ANES Scales Often Don't Measure What You Think They Measure – An ERPC2016 Analysis *

MATTHEW T. PIETRYKA
Florida State University
mpietryka@fsu.edu

RANDALL C. MACINTOSH
California State University, Sacramento
rmacintosh@csus.edu

December 14, 2017

Political surveys often include multi-item scales to measure individual predispositions such as authoritarianism, egalitarianism, or racial resentment. Scholars typically use these scales to examine how these predispositions vary across different subgroups, comparing women to men, rich to poor, or Republican to Democratic voters. Such research implicitly assumes that, say, Republican and Democratic voters' responses to the egalitarianism scale measure the same construct in the same metric. Unfortunately, this research rarely evaluates whether this assumption holds. We present a framework to test this assumption and correct scales when it fails to hold. We apply this framework to 13 commonly used scales on the 2012 and 2016 ANES. We find widespread violations of the equivalence assumption and demonstrate that these violations often lead to biased conclusions about the magnitude or direction of theoretically-important group differences. These results suggest that researchers should not rely on multi-item scales without first establishing measurement equivalence.

*For thoughtful suggestions, we thank Doug Ahler, Quintin Beazer, Rob Carroll, Kelley Doll, Brad Gomez, Chris Hare, Kelsey Houser, Bob Jackson, David Macdonald, and Jessica Parsons.

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; Jacoby 2006, 2014; 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 complex 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 and Tagar 2014; Federico, Fisher and Deason 2017; Hajnal and Rivera 2014; Hetherington and Suhay 2011; Hetherington and Husser 2012; Hutchings, Walton and Benjamin 2010; Kalkan, Layman and Uslaner 2009; Kam 2012; Miller, Saunders and Farhart 2016; O'Brien et al. 2013; Tesler 2012).

Multi-item scales provide great advantages over single-item measures (Ansolabehere, Rodden and Snyder Jr 2008), but their analysis also 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 (Gregorich 2006). Without measurement equivalence, the scale cannot provide meaningful group comparisons. 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 (2009); Pérez and Hetherington (2014); Stegmueller (2011). 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 about the distribution of predispositions in the electorate may be wrong.

We evaluate the extent of this problem by examining 13 of the most commonly used scales included in both the 2012 and 2016 ANES. Rather than strategically selecting only a few scales or groups to demonstrate our point, we evaluate as many as feasible. Further, we treat the 2012 analysis as exploratory, using the results to preregister the 2016 analysis. 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.

[1] We found 125 hits when conducting a Google Scholar search 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".

We contribute methodologically by helping researchers identify and correct inequivalence, thereby providing a means to address problems created by differential response patterns like social desirability or other unknown or overlooked factors. Our analysis suggests that all of the uncorrected ANES scales lack measurement equivalence for at least some theoretically important groups. Therefore, researchers must evaluate this assumption or risk biasing their conclusions. We thus provide on our website [URL removed] the corrected scale scores and instructions for merging them with the rest of the ANES data. Moreover, we describe a simple method scholars can use to evaluate equivalence and resolve its absence for datasets, scales, or groups we do not examine here. Though we focus on ANES scales, the threat of inequivalence and the method we describe to address it applies for comparisons based on any multi-item scale.

We contribute substantively by purging errors induced by inequivalent scales, thus providing more accurate estimates of how predispositions vary across 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 still 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. 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 be due to actual differences in racial animus or irrelevant differences in policy preferences.

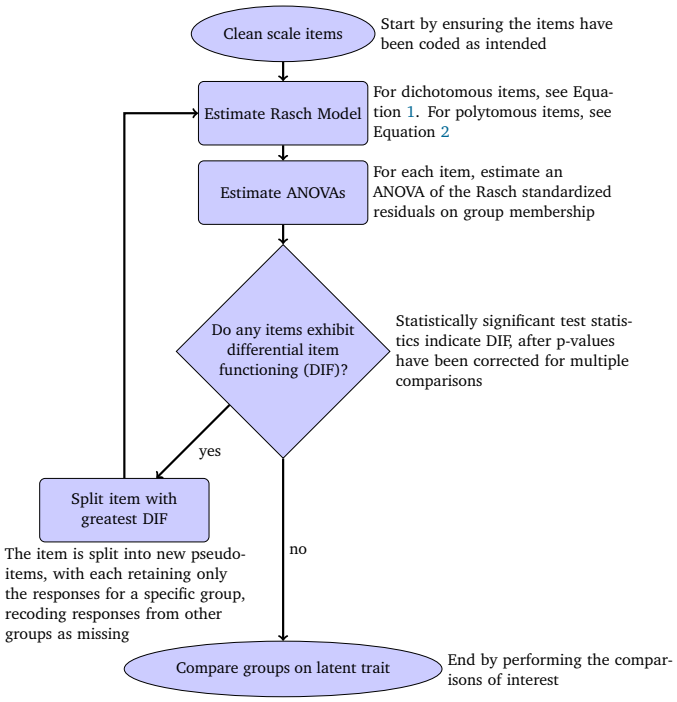
A well-developed framework exists to test for the presence of measurement equivalence (Andrich 2013*a*,*b*; Bond and Fox 2015; Gregorich 2006; Rasch 1980). The political knowledge literature provides a rare example where political scientists have applied this framework, finding that the apparent gender 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 2009, 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 political 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.

Meaningful comparisons require measurement equivalence

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

Figure 1: A workflow for establishing measurement equivalence



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). If true, 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. To assess this threat to a study's internal validity, researchers must test whether the measuring devices are sufficiently equivalent across groups to permit meaningful comparisons. This requirement applies in all cases, but inequivalence can be particularly misleading when expected differences are small, as is often the case in attitudinal research.

Differential item functioning indicates inequivalence

To evaluate this threat, we can test for differential item functioning (DIF), which indicates that items operate inconsistently across groups and therefore lack 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.[2] As no real-world data can be expected to fit a platonic model exactly, however, our interest lies primarily in the critical ways in which the data may fail to fit.

The Rasch model is surprisingly simple. 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

[2] Mokken (1971) provides an alternative, non-parametric scaling procedure, which has received several informative applications in the literature (e.g., Jacoby 1995, 2000; Mondak and Anderson 2004). The ordinal-level measures it yields, however, are insufficient for our purposes because we aim to improve the validity of these scales as they are typically used—as predictors or outcomes in regression analysis. Moreover, to correct scales when inequivalence arises, we conduct a form of test equating, which requires interval-level measurement (Meijer, Sijtsma and Smid 1990). Reassuringly, Mokken scaling tends to yield similar conclusions to the Rasch model (Molenaar 1997).

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—items with 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 the latent variable construction process requires additional corrective action (Andrich 2004, 2013*b*). 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.[3]

DIF occurs when an item's location varies systematically between groups, which produces statistically significant 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).[4]

[3] In contrast to other Item Response Theory models, the Rasch model exhibits an important property known as "specific objectivity." With specific objectivity, the scores across items are sufficient statistics to estimate the person location parameters. And the scores across persons are sufficient to estimate the item locations. When the data approximately fit the model, these properties mean that 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, the comparisons of items are independent of which persons participated in the survey. Specific objectivity is absent from other IRT models, such as the two- or three-parameter logistic models (2-pl and 3-pl), which are sample-dependent. Moreover, unlike the Rasch, these other IRT models may order the items inconsistently along the range of the latent continuum. This inconsistency occurs because each item response function takes on a different shape and the trace lines may cross at some point on the latent continuum, reversing the item order (Wilson 2005).

[4] 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

Correcting DIF by splitting items

To establish equivalence, all scale items must be evaluated for DIF, correcting it when it arises. 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 items are checked again for DIF. DIF is corrected by substituting the original biased item for new group-specific pseudo-items. 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. The process is repeated until no item shows DIF.[5]

Correction is not always possible because valid group comparisons require at least one DIF-free item since the corrected pseudo-items are group specific. This item acts 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. But this comparison requires that the same type of shoes is 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.[6]

Measurement inequivalence poses a serious internal validity threat, but political scientists have only examined the equivalence assumption for a few scales and grouping variables. At best, subsequent research has taken the results into account for those specific scales and

are typically too short to divide into more than four classes, presenting serious violations of the ANOVA assumptions. By examining only uniform DIF, we bias our conclusions *against* finding DIF, providing a more conservative test.

[5]Resolving DIF sequentially 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 appears negligible (Andrich and Hagquist 2012, 2015).

[6]The brevity of the scales may likewise limit their reliability or construct validity, which the DIF correction cannot address.

groups, but researchers working with other scales or groups have ignored the threat. We therefore seek to evaluate whether the lack of equivalence is limited to these instances or represents a more general threat to studies relying on ANES scales.

Data: 2012 and 2016 ANES

We draw data from the two most recent American National Election Studies. Rather than focus on just one scale and only a few grouping variables, we examine the most commonly used scales that are available in identical formats in both the 2012 and 2016 data. And we likewise examine the grouping variables we commonly see in analyses of these scales. By focusing on a broad search rather than a subset of scales and groups, readers can be confident that we have not cherry picked the scales and grouping variables to exaggerate the problems (Franco, Malhotra and Simonovits 2015). Instead, we provide a representative assessment of the problems researchers are likely to encounter. 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.[7] 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 13 scales that were included in both the 2012 and 2016 ANES: 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,

[7]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 by clicking the "View Registration Form" link at the following URL: https://osf.io/jc9nj/?view_only=bd5feb988241403496cd3d91a536a8dd

Extraversion, and Openness To Experiences).[8] We describe these scales below and explain their construction in online Supporting Information (SI) section A.1.

- The **Authoritarianism** scale relies on four dichotomous items, asking respondents about their child-rearing preferences. Scholars have relied on the child-rearing scale to measure authoritarianism because, unlike alternative measures, its items are conceptually distinct from political preferences (Feldman and Stenner 1997).

- The **Egalitarianism** scale was designed to measure the extent to which individuals value societal equality (Feldman 1988). The 2012 version includes six five-point Likert-type items, while the 2016 version includes only four of these. We therefore examine a scale constructed from the four common items.

- The **Limited Government** scale includes items intended to measure support for limiting government involvement in domestic affairs. Following previous work (Ansolabehere, Rodden and Snyder Jr 2008; Feldman and Huddy 2005), we rely on three dichotomous items.

- The **Moral Traditionalism** scale uses four five-point Likert-type items, intended to measure how opposed people are to changing moral standards.

- The **Negative Black Stereotypes** scale uses two items, intended to measure explicit acceptance of derogatory African American stereotypes. The first item asks respondents to place African Americans on a seven-point scale ranging from *hard-working* to *lazy*. In 2012, the second item asks an analogous question ranging from *intelligent* to *unintelligent*. In 2016, these anchors are replaced with *peaceful* and *violent*. Since

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

the items vary between years, we consider the 2016 analysis as exploratory.[9] This measure is typically applied only to whites and thus we restrict its analysis to white, non-Hispanic respondents.

- The **Non-Voting Participation** battery includes seven dichotomous items asking whether respondents have participated in a variety of political acts, including attending political meetings, displaying political signs, and donating money to campaigns. Though researchers sometimes analyze these items individually, they are often combined into a single multi-item scale (e.g., Dawkins 2017; Flavin and Griffin 2009; Valentino et al. 2011).

- The **Racial Resentment** scale includes four Likert-type items asking respondents whether they agree or disagree with statements about African Americans' place in society. As is typical (e.g., Tesler 2012; Feldman and Huddy 2005), we restrict the analysis of the racial resentment scale to white, non-Hispanic respondents.

- The **Ten Item Personality Inventory (TIPI)** relies on ten polytomous items to measure individual's personalities along five traits—two items per trait. Thus, the TIPI produces five two-item scales:

  - **Agreeableness** is intended to measure how sympathetic and warm individuals tend to be in their social interactions.
  - **Conscientiousness** is intended to measure how mindful, careful, and organized individuals tend to be.
  - **Emotional Stability** is intended to measure how even-tempered individuals tend to be. In some research, this trait is reverse coded and labeled instead as *Neuroticism*.

[9] These same items are also used in the Ethnocentrism scale (Kam and Kinder 2012). The Ethnocentrism scale is beyond the scope of our analysis because it relies on item transformations more complex than a simple average.

  - **Extraversion** is intended to measure how energetic and outgoing individuals tend to be.
  - **Openness To Experiences** is intended to measure individuals' open-mindedness and their propensity to seek new challenges and ideas.

- The **Wordsum** scale, originally developed for the General Social Survey, provides a vocabulary test intended to measure verbal skills, but is often used as a proxy for cognitive skills or general intelligence. Since 2012, the ANES has included the ten-item revised version developed by Cor et al. (2012).

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. 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. We describe how each grouping variable is constructed in SI-A.2.

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 proceed through each step in this process.

We use R version 3.3.3 (R Core Team 2017) to conduct our analysis, estimating the Rasch models with TAM 1.99993-0 (Kiefer, Robitzsch and Wu 2017), an IRT package that uses marginal maximum likelihood estimation. The models include as person weights the ANES Time Series Post-stratified full sample weight. We use the standard Rasch (Equation 1) for scales with dichotomous items and the partial credit model (Equation 2) for scales with polytomous items.[10]

Testing for DIF

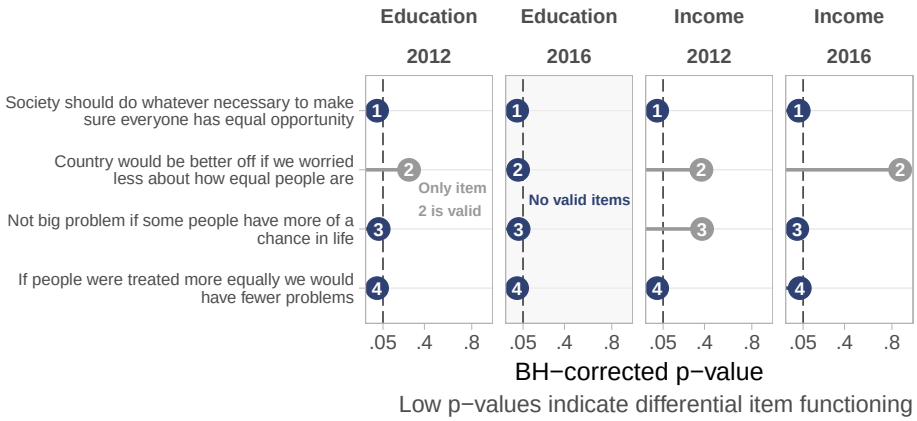
As a first step, we examine whether each scale exhibits DIF for each grouping variable using the one-way ANOVAs described above and in more detail in SI-B. As an example, we highlight DIF for the Egalitarianism scale grouped by education and income. Figure 2 shows the ANOVA p-values for each item. The 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 levels 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.[11] In the figure, dark boxes indicate that the item exhibited DIF for that grouping variable. When

[10] For polytomous items, an alternative to the partial credit model is the rating scale model, which constrains the threshold estimates (F_j in Equation 2) to be equal across items. For each of the polytomous scales in each year, however, a likelihood ratio test suggests the partial credit model provides significant improvement in fit over the rating scale model. Before examining DIF, we also assess threshold disorder for these scales, as described in SI-C.

[11] The ANOVAs used to detect DIF can be found in Tables B1–B13 of SI-B.

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 both years and item 2 in 2016. For income, the scale shows significant DIF for items 1 and 4 in both years and item 3 in 2016.

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 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 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.

The results seem consistent with 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.[12] 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.

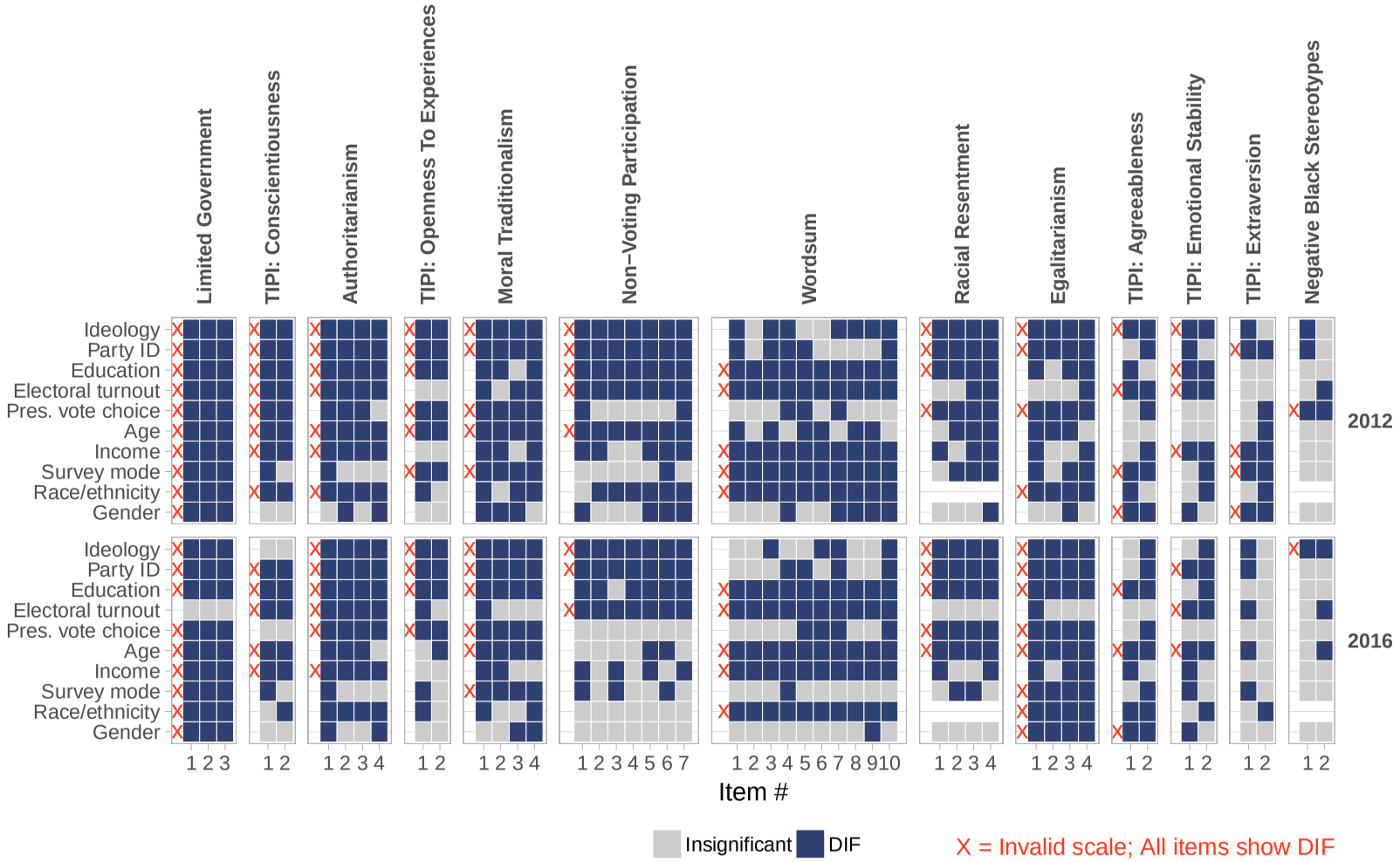
Though all scales exhibit DIF, some have more problems than others. On one end of the spectrum, all items from the Limited Government scale exhibit DIF for all grouping variables in 2012 and all but one in 2016. On the other end of the spectrum, the Negative Black Stereotypes items exhibit DIF in only a few cases. 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. Yet other grouping variables, such as gender, show DIF for relatively few items.

Despite widespread item-level validity problems, the scales can be corrected for many grouping variables.[13] This correction is possible as long as one or more items lacks DIF, acting as 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. For example, the data reveal that party identification groups differ qualitatively on moral traditionalism to the extent that they cannot be considered subgroups from the same population, at least as the construct is

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

[13] As we discuss in the conclusion, correcting DIF removes confounds created by unintended multidimensionality. Nonetheless, as Andrich and Hagquist (2015) notes, the source of DIF may be theoretically relevant and therefore worth examining in its own right.

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.

defined using this scale. 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.

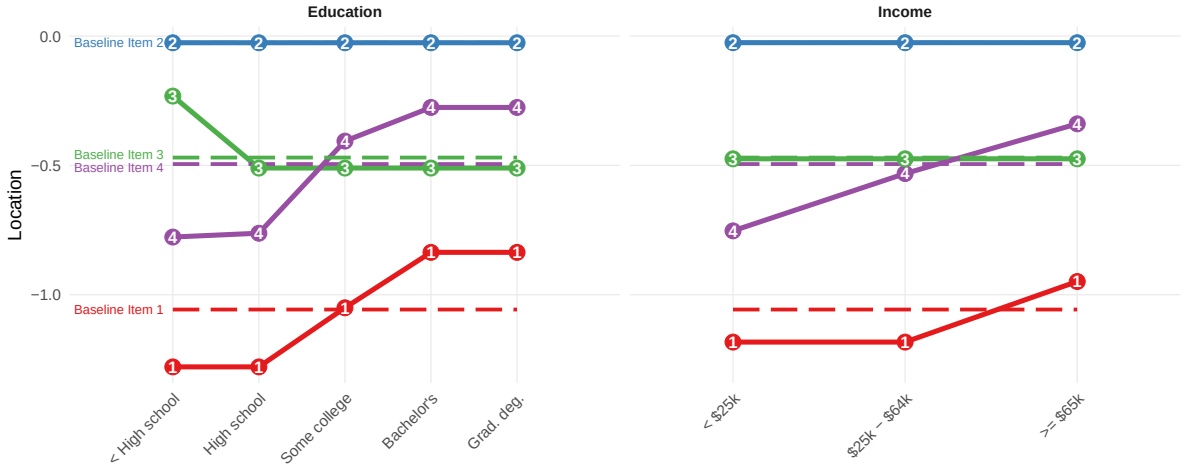
Correcting DIF

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 extract each person's location on the latent trait from the final Rasch model. Provided that no items show DIF in this final model, these corrected scales provide valid group comparisons on the latent trait.

As an example, Figure 4 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. The correction is possible 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.

In the left panel, 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

Figure 4: Egalitarianism item locations vary with education and income



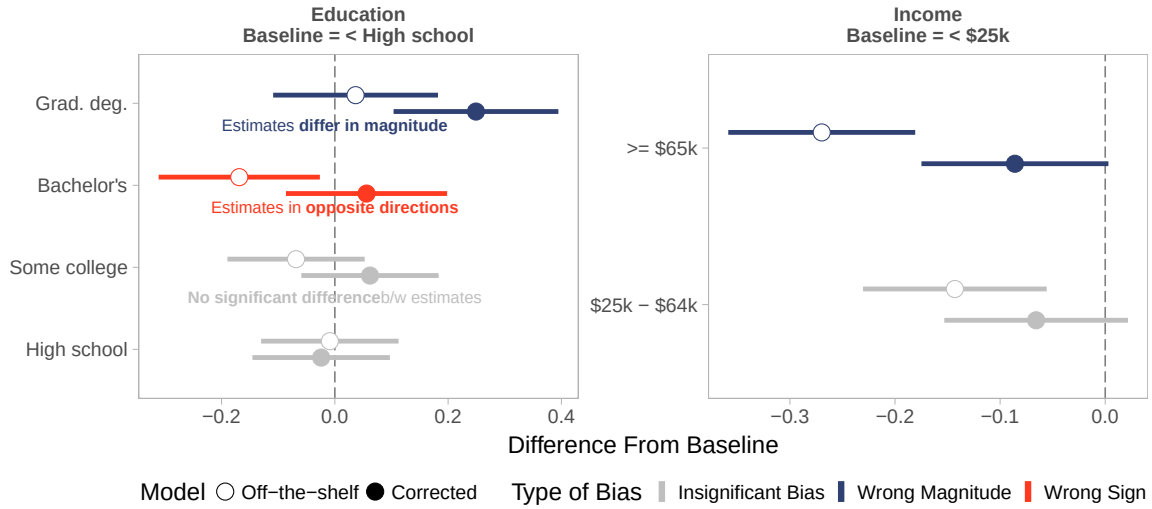
Note: The figure displays how the Egalitarianism item locations vary across education and income levels in the 2012 ANES. An item's location is inversely related to how likely someone is to choose the more egalitarianism response when choosing between two adjacent response categories. The dashed lines 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.

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 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 impact on the comparisons of interest. Thus, before abandoning the off-the-shelf scale, researchers should evaluate whether the DIF leads

Figure 5: 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).

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 5 plots the estimated relationships between the Egalitarianism scale and each of these grouping variables. Each panel in the figure provides coefficients from two linear regressions. In the first, the off-the-shelf, uncorrected scale is regressed on the focal grouping variable. In the second, the corrected, DIF-free scale is regressed on the grouping variable.[14] Since both grouping variables are ordinal, we include a dummy for each value, omitting the minimum value as the reference category.

[14]We standardize the off-the-shelf scores and corrected scores with mean = 0 and standard deviation = 1. We estimate the regressions in R using the `svyglm` function from the survey package (Lumley 2004, 2016). We estimate Taylor series standard errors for 2012 (DeBell 2010). The primary sampling unit variable on which these estimates rely is not available in the standard 2016 data, but is available in the restricted-access data. We have initiated a request for these data so we can update the 2016 analysis if given the chance to revise.

Figure 5 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 roughly a quarter of a standard deviation in egalitarianism.[15] 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.[16] 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.

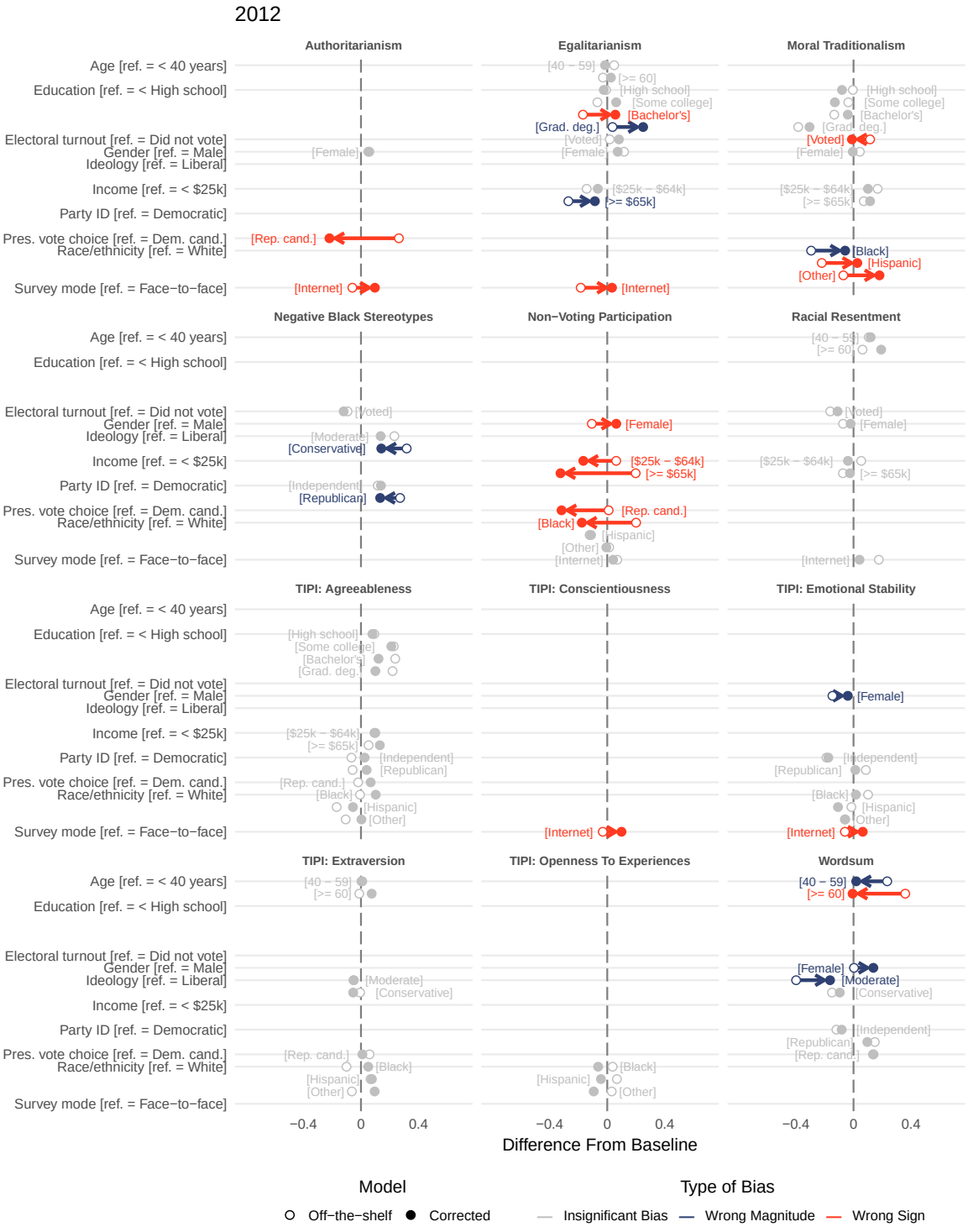
Given the large differences in the conclusions, readers may be surprised to learn that the corrected scales correlate strongly with the off-the-shelf scales. In 2012, the education-corrected Egalitarianism scale has a .98 correlation with the off-the-shelf scale. And the income-corrected Egalitarianism scale has a .97 correlation. These strong correlations occur among all of the off-the-shelf scales and their corrected counterparts, as shown in Figure D1 in SI-D. The correlations all exceed 0.8 in 2012 and 0.9 in 2016.

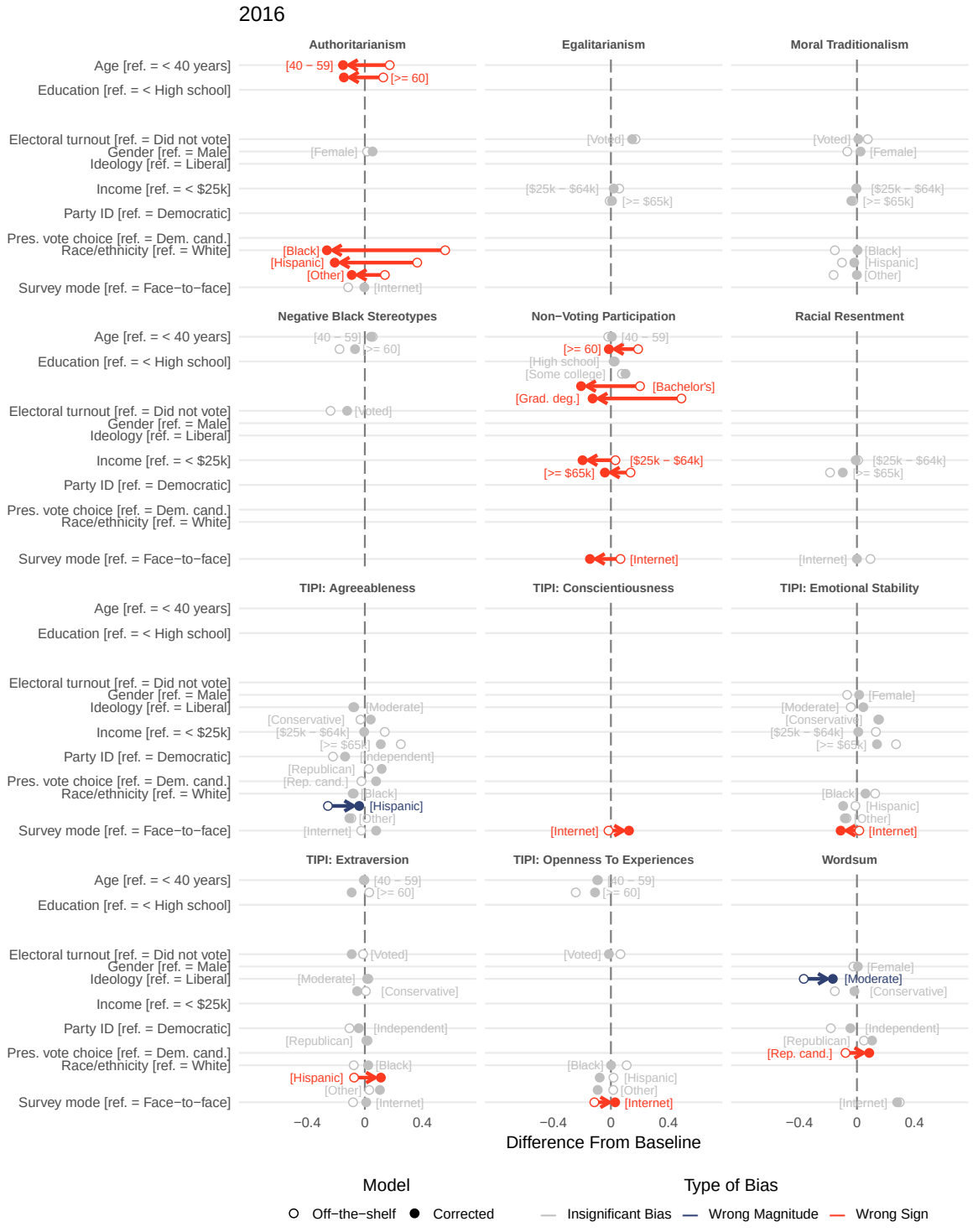
Despite these strong positive correlations, the off-the-shelf scales often suggest different substantive conclusions than do their corrected counterparts. Figure 6 displays the estimated relationship between each scale and each grouping variable.[17] The figure is restricted to scale-group combinations that could be corrected because they lacked significant DIF for at least one item. For exploratory analysis of the scales showing DIF for all items, see Figure E1 in SI-E.

The results in Figure 6 should give pause to researchers working with off-the-shelf scales. While many relationships remain unchanged, a large proportion of the corrected estimates

[15]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.
[16]The difference between the estimates is -0.18 (95%CI = [-0.31, -0.06])
[17]These estimates are derived from the same process described for Figure 5.

Figure 6: 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.

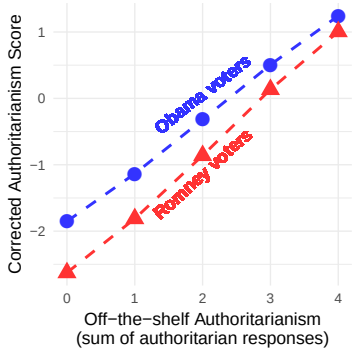
differ significantly from the uncorrected ones. In these cases, the off-the-shelf scale will lead to biased conclusions. For instance, the off-the-shelf 2012 Negative Black Stereotypes scale exaggerates the differences between Republican and Democratic respondents. In many cases, the direction of the 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. 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 calls for research that reexamines past findings after removing bias from measurement inequivalence.

Discussion

Readers who lack familiarity with these 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

Figure 7: 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.

a notable example, where Republican voters average lower scores on the latent variable than Democratic voters. This pattern emerges because three of the four child-rearing items are interpreted differently by the two groups. These data suggest 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. [18] 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 7. 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-F.

[18] Recall that the off-the-shelf score is a sufficient statistic to estimate the latent trait score.

We are not arguing that we have found the "true" relationship between vote choice in 2012 and authoritarianism. Rather, we argue that we have found the 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 child-rearing scale lacks construct validity. The challenge for this work will be to avoid returning to a tautological measure. As noted above, researchers first introduced the child-rearing scale because it measures authoritarianism with items that are substantively distinct from its theorized consequences. If scholars reject this scale based solely on its correlation with political preferences, they will negate this benefit.

Conclusion

Measurement equivalence is a fundamental element of high-quality scholarship. Its absence leaves any study vulnerable to the justified criticism that demonstrated effects (or absence of effects) 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. All of the 13 scales we examine lack measurement equivalence for theoretically-important grouping variables (Figure 3). For instance, each 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 (Jacoby 1991, Ch. 4). These results highlight the need for greater theoretical development, allowing future work to reveal what those differences represent. Such work would be of substantive interest and can also provide methodological guidance for creating improved scales.

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, off-the-shelf scales (Figure 6). Therefore, scholars should not assume results are valid unless measurement equivalence has been confirmed. 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 the off-the-shelf scale 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. 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 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 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*a*. "An Expanded Derivation of the Threshold Structure of the Polytomous Rasch Model That Dispels Any "Threshold Disorder Controversy"." *Educational and Psychological Measurement* 73(1):78–124.

Andrich, David. 2013*b*. "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.

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

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*. 3 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.

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

Cor, M. Ken, Edward Haertel, Jon A. Krosnick and Neil Malhotra. 2012. "Improving ability measurement in surveys by following the principles of IRT: The Wordsum vocabulary test in the General Social Survey." *Social Science Research* 41(5):1003–1016.

Dawkins, Ryan. 2017. "Political participation, personality, and the conditional effect of campaign mobilization." *Electoral Studies* 45:100–109.

DeBell, Matthew. 2010. How to analyze ANES survey data. Technical Report nes012492 ANES Technical Report. http://www.electionstudies.org/Library/papers/nes012492.pdf.

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.

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.

Federico, Christopher M. and Michal Reifen Tagar. 2014. "Zeroing in on the Right: Education and the Partisan Expression of Authoritarianism in the United States." *Political Behavior* 36(3):581–603.

Feldman, Stanley. 1988. "Structure and Consistency in Public Opinion: the Role of Core Beliefs and Values." *American Journal of Political Science* 32(2):416–440.

Feldman, Stanley and Karen Stenner. 1997. "Perceived Threat and Authoritarianism." *Political Psychology* 18(4):741–770.

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.

Flavin, Patrick and John D. Griffin. 2009. "Policy, Preferences, and Participation: Government's Impact on Democratic Citizenship." *The Journal of Politics* 71(2):544–559.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23(2):306–312.

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.

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.

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.

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.

Jacoby, William G. 1995. "The Structure of Ideological Thinking in the American Electorate." *American Journal of Political Science* 39(2):314–335.

Jacoby, William G. 2000. "Issue Framing and Public Opinion on Government Spending." *American Journal of Political Science* 44(4):750–767.

Jacoby, William G. 2006. "Value Choices and American Public Opinion." *American Journal of Political Science* 50(3):706–723.

Jacoby, William G. 2014. "Is There a Culture War? Conflicting Value Structures in American Public Opinion." *American Political Science Review* 108(04):754–771.

Kalkan, Kerem Ozan, Geoffrey C. Layman and Eric M. Uslaner. 2009. ""Bands of Others"? Attitudes toward Muslims in Contemporary American Society." *The Journal of Politics* 71(3):847–862.

Kam, Cindy D. 2012. "Risk Attitudes and Political Participation." *American Journal of Political Science* 56(4):817–836.

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.

Kiefer, Thomas, Alexander Robitzsch and Margaret Wu. 2017. "TAM: Test Analysis Modules.". https://CRAN.R-project.org/package=TAM.

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.

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

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9(1):1–19.

Lumley, Thomas. 2016. "survey: analysis of complex survey samples.". https://cran.r-project.org/web/packages/survey/index.html.

Meijer, Rob R., Klaas Sijtsma and Nico G. Smid. 1990. "Theoretical and Empirical Comparison of the Mokken and the Rasch Approach to IRT." *Applied Psychological Measurement* 14(3):283–298.

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.

Molenaar, Ivo W. 1997. Lenient or Strict Application of IRT with an Eye on Practical Consequences. In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, ed. Jürgen Rost and Rolf Langeheine. New York, NY: Waxmann Münster pp. 38–49.

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. 2009. "Lost in Translation? Item Validity in Bilingual Political Surveys." *The Journal of Politics* 71(4):1530–1548.

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.

R Core Team. 2017. "R: A Language and Environment for Statistical Computing.". https://www.R-project.org/.

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.

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

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. 2015. "The ANES 2012 Time Series Study [dataset].". Stanford University & the University of Michigan [producers]. http://www.electionstudies.org.

The American National Election Studies. 2017. "The ANES 2016 Time Series Study [dataset].". University of Michigan & Stanford University [producers]. http://www.electionstudies.org.

Valentino, Nicholas A., Ted Brader, Eric W. Groenendyk, Krysha Gregorowicz and Vincent L. Hutchings. 2011. "Election Night's Alright for Fighting: The Role of Emotions in Political Participation." *The Journal of Politics* 73(1):156–170.

Wilson, Mark. 2005. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, N.J: Routledge.

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.