Improving the Measurement of Hostile Sexism

Allison M. N. Archer*
Assistant Professor
Department of Political Science &
Valenti School of Communication
University of Houston
amarcher3@uh.edu

Scott Clifford
Associate Professor
Department of Political Science
University of Houston
sclifford@uh.edu

Abstract. In recent years, sexism has played an increasingly pivotal role in American politics, and scholarship examining the importance of gender attitudes for political behavior has surged. Researchers have largely relied on the hostile sexism scale to measure prejudice against women, and this scale seems particularly relevant to political science research. However, this scale measures attitudes with an agree-disagree response format, which has long been recognized as a source of substantial measurement error. In this paper, we introduce a revised version of the hostile sexism scale that instead relies on an item-specific question format. Across three studies, we show that the item-specific scale is strongly related to the agree-disagree scale, but that the item-specific version reduces problems with truncation and tends to improve discriminant and predictive validity. Given these advantages, we conclude by recommending that researchers adopt the item-specific hostile sexism scale.

The authors would like to thank Erin Cassese, Cindy Kam, Beth Simas and Steve Utych for helpful comments and suggestions. Data collection was funded by the University of Houston.

*Corresponding author.

In recent years, gender has been at the forefront of national politics and sexism has played an increasingly salient role in public opinion and political behavior. Hillary Clinton's 2016 candidacy marked the first time a woman was chosen to run at the top of the ticket. Clinton's opponent, Donald Trump, frequently used sexist rhetoric during the campaign and faced multiple allegations of past sexual misconduct. In the following presidential election, a record number of women ran in the Democratic primary, and Kamala Harris made history as the first woman vice president among many other "firsts." Throughout this same period, the rise of the #MeToo movement focused attention on allegations of sexual harassment and assault against individuals throughout society, including political figures like Roy Moore, Al Franken, and Brett Kavanaugh. Political activism related to the #MeToo movement has forced legislators, businesses, and the media to pay greater attention to systemic gender imbalances and consider policies to address sexual misconduct.

Given these events, scholars have increasingly focused on sexism as an explanation for political behavior (Schaffner 2021). Research suggests sexism significantly affected political outcomes in the 2012 (Simas and Bumgardner 2017), 2016 (Banda and Cassese 2021; Cassese and Holman 2019; Godbole, Malvar, and Valian 2019; Knuckey 2019; Schaffner, Macwilliams, and Nteta 2018; Setzler and Yanus 2018; Sides, Tesler, and Vavreck 2018), and 2020 presidential elections (Utych 2020). Notably, the impact of gender attitudes has increased in recent years (Cassese and Barnes 2019; Kam and Archer 2021; McThomas and Tesler 2016; Valentino, Wayne, and Oceno 2018), and sexism significantly affects views on sexual misconduct and the #MeToo movement (Archer and Kam 2021). Thus, sexism has been activated in public opinion and political behavior, particularly since the 2016 election and the rise of the #MeToo movement.

Given increased scholarly attention to the topic, researchers have sought a more unified approach to measuring sexist attitudes. Recent work evaluated several alternative measures and concluded by encouraging scholars to adopt the hostile sexism scale for studying political behavior (Schaffner 2021). However, the hostile sexism measure (and many similar alternatives) relies on an agree-disagree response format, which is known to introduce measurement error (Pasek and Krosnick 2010). Specifically, agree-disagree formats introduce a strong response bias that conflates agreement with item content, inflates correlations with other constructs measured on the same scale, and encourages satisficing. Thus, while the hostile sexism scale is conceptually relevant to politics, the scale itself needs improvement.

In this paper, we introduce an item-specific version of the hostile sexism scale, which incorporates the relevant concept into the response options. Through three studies, we show that the item-specific version improves upon the performance of the agree-disagree version by reducing truncation, improving discriminant validity, and increasing predictive validity. We conclude by recommending researchers use the item-specific version of the hostile sexism scale.

Measuring Sexism

Scholars have measured prejudice against women using several prominent scales. Initially, researchers asked questions assessing traditional views of women through negative stereotypes about competence (e.g., intelligence) and beliefs about differential rights and roles ascribed to women and men (Spence, Helmreich, and Stapp 1973; Swim et al. 1995). One popular question asks whether men and women should have equal roles in society or if a woman's place is "in the home." Over time, attitudes have liberalized, leading to widespread consensus on these questions (Glick and Fiske 1996) due to socialization, work experiences, and

family context (Banaszak and Plutzer 1993; Plutzer 1991; Powell and Steelman 1982; Rhodebeck 1996).[1] Thus, measures of traditional views about women and gender roles have proven to be less useful over time in applied research.

To move beyond traditional gender attitudes, psychologists created the Ambivalent Sexism Inventory (ASI) (Glick and Fiske 1996) and Modern Sexism (Swim et al. 1995) scales.[2] The ASI consists of two theoretically distinct elements of prejudice: hostile and benevolent sexism. Benevolent sexism reflects attitudes about women that are subjectively positive in tone yet ultimately reinforce a viewpoint that places men above women in the social hierarchy. Its items encompass three broad categories: protective paternalism, complementary gender differentiation, and heterosexual intimacy (Glick and Fiske 1996). Benevolent sexists believe women should be safeguarded, cherished, and chaste. Hostile sexism reflects the antipathy toward women that is more traditionally associated with prejudice against women (Glick and Fiske 1996). Its items tap three broad categories: dominative paternalism, competitive gender differentiation, and heterosexual hostility (Glick and Fiske 1996).[3] Hostile sexism is similar to modern sexism, which consists of the denial of gender discrimination and resentment toward people or policies seeking to address gender inequality (Swim et al. 1995; Swim and Cohen

[1] Indeed, surveys like the American National Election Studies no longer carry this question (see Archer and Kam 2020 and Schaffner 2021 for discussion).

[2] The Neosexism scale (Tougas et al. 1995) is similar, but much less prominent in political science.

[3] We discuss question wording for the hostile sexism battery in detail later in the paper. The Appendix includes the full scale, plus benevolent and modern sexism scales.

1997). Indeed, the two are strongly related (Schaffner 2021). Taken together, the hostile, modern, and benevolent sexism scales represent measures of gender prejudice that are subtler than old-fashioned sexism and that provide greater predictive validity than the traditional scales (Glick and Fiske 1996; Swim et al. 1995).

We focus on hostile sexism in this paper for several reasons. First, recent work suggests hostile sexism is more relevant to politics than modern and benevolent sexism. Evaluations of the three scales' convergent and predictive validity, plus their proximity to politics suggest a subset of hostile sexism is most useful for political science (Schaffner 2021).[4] Second, hostile and modern sexism are theoretically and empirically similar. Both scales measure antagonistic views toward women, and they are strongly related (Glick and Fiske 1996; Schaffner 2021). Further, hostile and modern sexism are sometimes discussed and/or presented interchangeably given their similarity (Cassese and Barnes 2019) and due to limitations in item availability (Valentino et al. 2018). Finally, we acknowledge that benevolent sexism represents a distinct and important dimension of gender attitudes (Glick and Fiske 1996; Schaffner 2021). For example, benevolent sexism helps explain reactions to politicians' involvement in sex scandals (Barnes, Beaulieu, and Saxton 2020). However, the popularity of hostile sexism and its greater influence on highly studied political outcomes like candidate evaluations (Schaffner 2021) motivate the present focus on this scale. Notably, however, modern and benevolent sexism also use agree-disagree scales, so our results hold implications for these batteries as well.

[4] Schaffner (2021) examines predictive validity using candidate choice and views on policies about gender, plus a nonpolitical outcome of boss choice.

The Problems with Agree-Disagree Scales

The hostile sexism scale relies on agree-disagree (AD) response choices to measure prejudice against women. AD scales are extremely popular in survey research, likely because they are easy to construct and allow for the placement of many attitudes on a single scale. The AD scale also facilitates the use of grids, which can save survey time (Couper et al. 2013). Despite their popularity, influential texts on questionnaire design have long discouraged the use of this response format in favor of item-specific (IS) response formats, which incorporate the relevant concept into the response options (Krosnick and Presser 2010; Pasek and Krosnick 2010; Saris et al. 2010).

A large body of research suggests AD scales suffer from several shortcomings (e.g., Höhne and Krebs 2018; Höhne and Lenzner 2018; Höhne, Schlosser, and Krebs 2017; Pasek and Krosnick 2010; Saris et al. 2010). First, they are more cognitively difficult to answer and thus encourage satisficing. This is because respondents are presumed to think in terms of the specific construct at hand, then have to translate it to the AD scale (Fowler and Cosenza 2008; Pasek and Krosnick 2010). For example, consider asking how much someone agrees or disagrees with the statement that "the issue of abortion is important to me." The respondent might first think that the issue is only somewhat important, then must figure out how to translate "somewhat important" into agreement with the statement (e.g., "somewhat agree" that "abortion is important"). In contrast, if an IS scale is used, the respondent can skip the last step and simply select "somewhat important." As a result, respondents actually devote more effort to answering IS scales, as indicated by response times (Höhne, Schlosser, and Krebs 2017), eye-tracking (Höhne and Lenzner 2018), and self-reports (Höhne and Krebs 2018). This literature suggests that by varying response options across questions, the IS format encourages more conscientious

and effortful responding, while the repetitive and less conceptually relevant AD format lulls respondents into low-effort responding.

A second shortcoming is that AD scales are particularly prone to method or response bias. That is, a significant proportion of the variance in responses is not due to the latent construct (e.g., sexism), but is instead a response to the scale format (e.g., agreement). Respondents are more inclined to agree with a statement than disagree, regardless of the content, and this effect is estimated at about 10 percentage points (Krosnick and Presser 2010). This acquiescence bias is likely driven by some combination of conversational norms, politeness, and satisficing (Krosnick 1991). However, some people are more susceptible to acquiescence bias, particularly those who are unmotivated or unable to put in the required effort. For example, older and less-educated respondents are more prone to acquiescence bias (Roberts et al. 2019). There is also some evidence for differences by ethnicity and cultural norms (Baron-Epel et al. 2010; Javeline 1999; Ross and Mirowsky 1984) and by gender (Weijters, Geuens, and Schillewaert 2010). Thus, the error introduced by acquiescence bias is not simply a uniform shift.

Some researchers have tried to ameliorate this problem by introducing reversed items. For example, for eight of the hostile sexism questions, agreement indicates high levels of sexism (e.g., "Women are too easily offended"), while for three questions agreement indicates low levels (e.g., "Feminists are making entirely reasonable demands of men"). As the authors of this scale explain, "items were reworded to yield the reverse meanings to control for acquiescence bias" (Glick and Fiske 1996, 496). However, this strategy simply replaces one problem with another (Zhang, Noor, and Savalei 2016). Rather than the reversed items canceling out bias, respondents who are prone to acquiescence bias are pooled to the middle of the scale, regardless of where they belong on the latent variable. Supporting this concern, factor analysis of the hostile

sexism scale reveals that reversed items loaded "much less strongly" on the latent dimension even after attempting to explicitly model out acquiescence bias (Schaffner 2021).

The problems of acquiescence bias and reversed items are exacerbated by the common practice of placing these items in one or more grids or matrices. Grids are appealing because they typically reduce survey time. However, grids also increase correlations between individual items. Respondents are less likely to differentiate between items placed in a grid, increasing measurement error (Couper et al. 2013), likely due to satisficing. Respondents move through the questions quickly and may not notice that an item is reversed.

Taken together, the evidence suggests AD scales introduce substantial bias. In one application of AD scales (measuring Facebook usage), this bias was estimated to account for roughly 20% of the variance in responses (Kuru and Pasek 2016). This can have several deleterious effects, including biasing correlations between the target construct and variables that predict acquiescence bias (e.g., education and age). For example, a study of conspiracy beliefs found that AD scales, relative to an alternative format, substantially increased conspiracy endorsement, particularly among those low in political knowledge and cognitive reflection (Clifford, Kim, and Sullivan 2020). It can also inflate correlations between items within a given scale, leading to overestimates of internal coherence, as well as with other constructs also measured on an AD scale (Kuru and Pasek 2016).

Item-specific (IS) scales are typically less susceptible to the problems faced by AD scales. As noted above, an eye-tracking study found respondents give more attention to response options for IS scales than AD scales (Höhne and Lenzner 2018). IS scales also seem to be less influenced by scale direction (Höhne and Krebs 2018) and correlations between measures are less affected by shared response formats (Kuru and Pasek 2016). However, one application

found little difference between AD and IS scales' predictive validity (Lelkes and Weiss 2015). Clearly, IS items are not a panacea, but the literature suggests they tend to reduce substantial biases observed in AD scales.

Development of an Item-Specific Hostile Sexism Scale

The available evidence suggests the AD format may also introduce bias into the hostile sexism scale. Two sets of authors have independently created short versions of the scale, and both removed all reversed items after finding they loaded poorly on the primary factor (Glick and Whitehead 2010; Rollero, Glick, and Tartaglia 2014; Schaffner 2021). Hostile sexism scales have also suffered from truncation. Roughly 15% of respondents in a recent study scored at the minimum value of hostile sexism, suggesting non-trivial floor effects (Schaffner, Macwilliams, and Nteta 2018). Others have used tobit models to address similar truncation issues with a closely related measure (Archer and Kam 2021). The censorship of low scores, which seem to represent "an actively anti-sexist" perspective (Schaffner 2021), may hamper the functioning of the scale. Taken together, existing work is suggestive of problems due to the AD format.

In an effort to improve the hostile sexism scale, we revised it to accommodate an IS format. Table 1 shows the original AD statements next to our IS questions and response options for the 5-item subscale suggested by Schaffner (2021) (see Appendix for the full scale). We attempted to keep the question stem as similar as possible to the original wording, with most items requiring only modest changes. For example, rather than ask respondents about their agreement with statements like "women seek to gain power by getting control over men," we asked respondents how *often* or how *many* women engage in a particular kind of behavior. For these questions, we used five-point unipolar response scales.

In some cases, a bipolar scale seemed more appropriate. Consider the statement that "women are too easily offended." A respondent might disagree either because they think that women take offense precisely when they should, or because they think women should be offended more often. So, we used a bipolar scale ranging from women are offended "much too often" to "not nearly often enough," with a midpoint of "about the right amount." By incorporating a bipolar scale, we can measure a wider range of attitudes than the AD scale, which should reduce truncation.[5]

Table 1. AD and IS Wording of the Hostile Sexism Scale

AD	IS	IS Response Options
Women seek to gain power by getting control over men	How often do women seek to gain power by getting control over men?	Never, rarely, sometimes, often, always
Women exaggerate problems they have at work	How often do women exaggerate problems they have at work?	Never, rarely, sometimes, often, always
Once a woman gets a man to commit, she puts him on a tight leash	When a woman gets a man to commit, how often does she put him on a tight leash?	Never, rarely, sometimes, often, always
Women are too easily offended	Would you say that women are offended too often, or not offended often enough?	Offended [much too often, a bit too often, about the right amount, not quite often enough, not nearly often enough]
	Do you think women give men too much credit or too	

[5] Another potential solution is adding more scale points to the typical 5-point scales. However, research suggests 5-point AD scales are preferable to longer ones, as more categories reduce data quality by increasing variance in respondents' interpretations of the response choices (Revilla, Saris, and Krosnick 2014).

Most women fail to appreciate fully all that men do for them	little credit for what men do for them?	Women give men [far too much, a bit too much, about the right amount of, a bit too little, far too little] credit

Study 1

As a first test of our IS scale, we recruited 300 respondents from Amazon's Mechanical Turk to complete a survey. Respondents were required to have completed at least 500 studies, have an approval rate of at least 98%, and be in the US. Additionally, we blocked respondents whose IP address indicated they were using a VPN or taking the survey from outside the US (Kennedy et al. 2020). The study was fielded on November 10, 2020, and 309 respondents completed the study. We retain only the 301 who completed all of the sexism questions. Respondents tended to be young (median age = 36) and tended to identify with the Democratic party (52%); 79% were white, and 54% were college-educated. See Appendix for full details. As with all three of our studies, surveys were fielded using Qualtrics software, and respondents were allowed to answer using their preferred device.

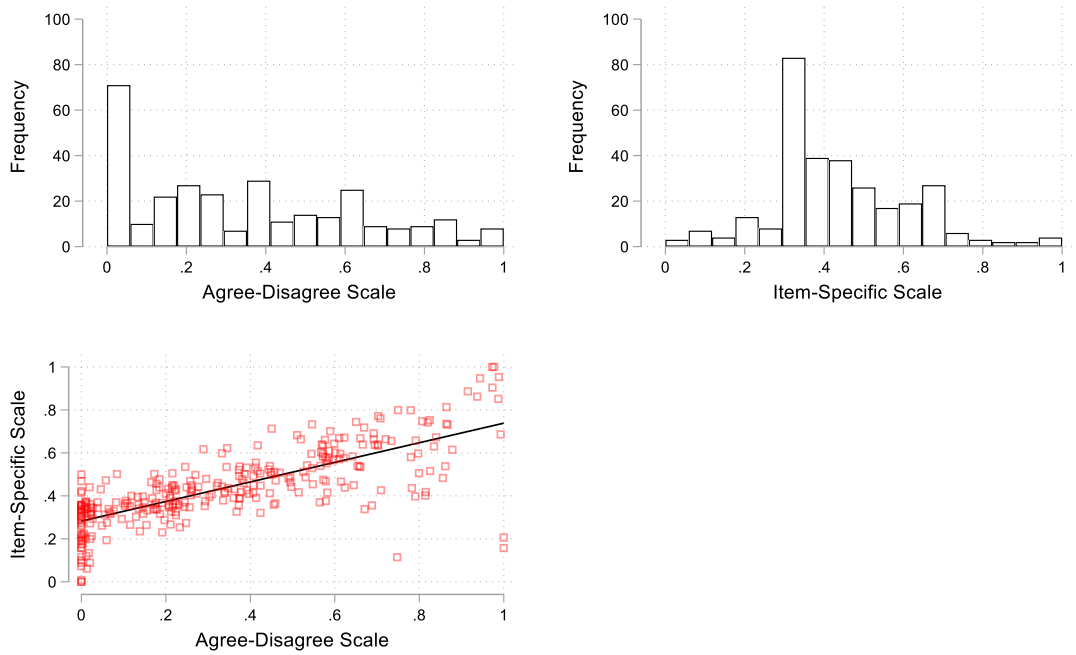
After answering questions regarding demographics and partisanship, respondents were asked both the AD version and the IS version of the 5-item scale in randomized order. As is common in applied research, the AD items were displayed horizontally in grids; the IS items were displayed vertically and individually. The AD items were also presented on a 6-point scale with no midpoint. Finally, respondents were asked one question regarding their feelings about the scales. The question randomly selected one of the hostile sexism items and displayed images of the AD and IS versions side-by-side, in random order. Respondents were asked which of the two questions "best measured your own opinion on the topic." Because Turkers are expert survey takers, we expect them to have well-formed opinions about survey design.

Results

Both the IS and AD formats yielded reliable scales ($\alpha = 0.84$, $\alpha = 0.94$, respectively), though the latter had slightly higher reliability. Of course, this could be due to the AD format and grid design inflating internal reliability (Couper et al. 2013). The median respondent spent 27 seconds on the IS scale, but only 16 seconds on the AD, likely due to the AD scale's grid format. Thus, there is a small cost in reaction time to using the IS scale.

Figure 1 displays histograms of both scales, which were rescaled to range from 0-1. The AD scale clearly suffers from floor effects, with 22% of respondents at the minimum value. In contrast, only 1% of respondents receive the minimum score on the IS scale. To further illustrate, among respondents at the minimum value of the AD scale ($n = 65$), their scores on the IS scale range from 0 to 0.5 (mean = 0.27). The two scales are strongly correlated ($r = .77$; see bottom panel of Figure 1), but the IS scale clearly captures variation missed by the AD scale.

Figure 1. Distributions of the Agree-Disagree and Item-Specific Scales



Finally, we look at respondent reactions. Even though it took more time to complete the IS questions, 64% of respondents (95% CI: 58%, 69%) preferred the IS question. Thus, despite

being incentivized to complete tasks as quickly as possible, Turkers prefer the IS scale by about 1.8 to 1.

Discussion

Overall, Study 1 demonstrates that the IS and AD scales are strongly related. However, the IS scale picks up variation missed by the AD scale due to floor effects, and a clear majority of respondents preferred the IS scale. While Study 1 provides some evidence that the IS scale improves upon the AD version, it is limited by a reliance on a relatively small convenience sample that leaned liberal, perhaps inflating floor effects.

Study 2

To expand on Study 1, we conducted a second online study on a more diverse sample from January 4-5, 2021. Respondents were recruited through Lucid Theorem, a platform that allows researchers to target samples balanced on age, gender, ethnicity and region.[6] Due to concerns about data quality, we excluded respondents who failed an attention check placed among the first few questions of the survey. After screening out inattentive respondents, 2,258 completed the study.[7] Compared to Study 1, respondents were somewhat older (median age =

[6] Lucid partners with companies to provide convenience samples of online survey participants. Individuals are invited to participate in research, and companies provide incentives like cash or reward points. For more on Lucid, see Coppock and McClellan (2019).

[7] We included a more challenging attention check later in the study, and 91% of those who completed the survey passed it, suggesting acceptable data quality (Thomas and Clifford 2017).

45), less college-educated (44%), and less Democratic (39%), but with similar racial demographics (74% white). See Appendix for details.

In contrast to Study 1, respondents were randomized to either the AD or the IS version of the hostile sexism scale. Respondents first answered the five questions used in Study 1, then on a separate page answered the remaining six items from the full hostile sexism scale. Again, in keeping with common practices, the AD items were displayed horizontally in grids, while the IS items were displayed vertically and individually. Additionally, the AD items were presented on a 6-point scale with no midpoint. Because political scientists rarely use the full 11-item scale, we focus our attention on the five-item scale (see the Appendix for analyses of the 11-item scale) and examine the discriminant and predictive validity of each format.

To examine discriminant validity, we included a scale measuring feelings of control over one's life (Mirowsky and Ross 1990), which consists of statements like "I am responsible for my own success" (see Appendix for the full battery). We selected this scale not for substantive content, but because it consists of eight items, all measured on an AD scale, with half of the items reversed. This allows us to test whether the sexism scale is differentially related to the two halves of the control scale. The four reversed items were placed in a separate grid from the other four items, and the grids were placed on separate pages. This design should reduce response bias because respondents will be more likely to notice the change in direction when opposing items are placed in separate grids rather than intermixed. This scale allows a test of discriminant validity in that hostile sexism should be equally related to the two halves of the control scale. But, if the AD format introduces a response bias, then hostile sexism should be more strongly related to the half of the control scale oriented in the same direction as the hostile sexism scale as compared to the half that is reversed. Moreover, this difference should be larger when the hostile

sexism scale is measured with an AD format than an IS format. If so, then it suggests the response bias in the AD hostile sexism battery could artificially inflate or deflate relationships with other constructs also measured on an AD scale.

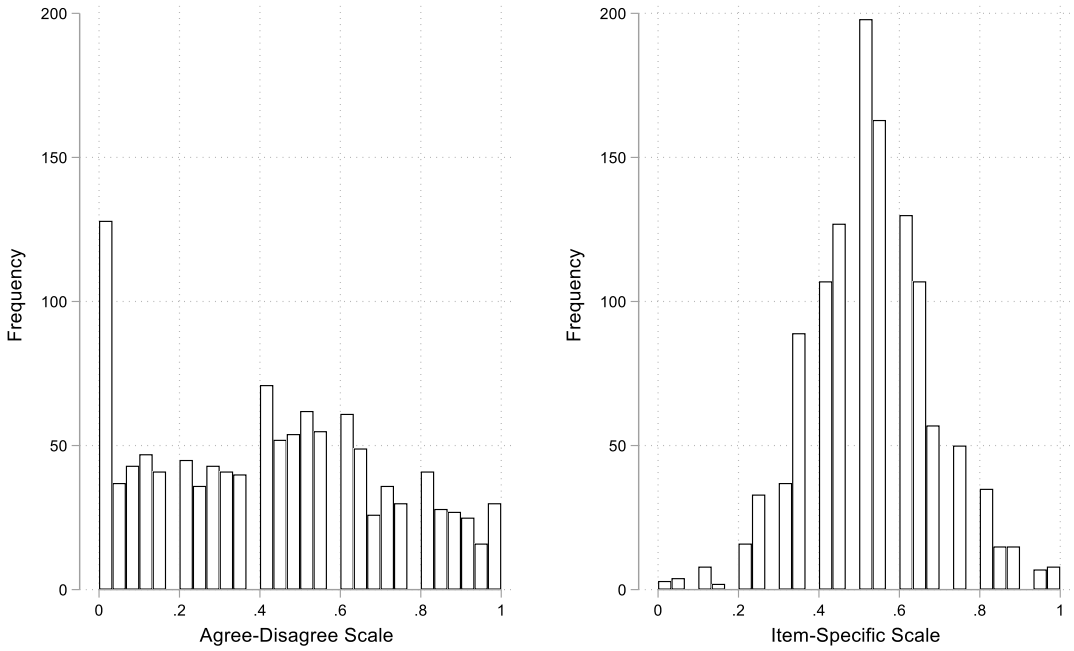
To assess predictive validity, we followed past work and measured attitudes toward a variety of men and women politicians (Utych 2020), plus attitudes toward the #MeToo movement (Archer and Kam 2021). Attitudes toward politicians were measured on a standard 101-point feeling thermometer in a single grid. Politicians included three Democratic women (Hillary Clinton, Kamala Harris, Elizabeth Warren), three Democratic men (Joe Biden, Barack Obama, Bernie Sanders), two Republican women (Amy Coney Barrett, Sarah Palin), and two Republican men (Mike Pence, Donald Trump). We measured attitudes toward #MeToo with five questions regarding whether the movement has gone too far, whether workplace sexual harassment should be addressed by employers or individuals, support for mandatory workplace sexual harassment training, perceptions of backlash to #MeToo, and whether recent allegations of sexual harassment were isolated incidents or part of a larger societal problem. Finally, respondents rated the credibility of sexual misconduct allegations against several celebrities. However, due to high no-opinion rates for these credibility questions, we relegate these analyses to the Appendix.[8]

Results

[8] We find no significant difference in the predictive validity of the scales for the credibility items.

Figure 2 shows the distributions of the AD and IS versions of the hostile sexism scale. Consistent with Study 1, truncation is a problem for the AD scale (14% at the minimum or maximum value), but not the IS format (1%).

Figure 2. Distributions of the Item-Specific and Agree-Disagree Scales



Both versions formed reliable scales, though the AD format had higher reliability (α = .92) than the IS version (α = .71). While this may seem like a virtue of the AD scale, it is also consistent with the possibility that this format inflates the relationship between items scored in the same direction.[9] Respondents also spent less time on the five AD items (median = 32 seconds) than the same IS items (44 seconds), though the longer time may indicate deeper processing of the questions (Höhne and Lenzner 2018).

Discriminant Validity

[9] Indeed, the 11-item AD scale is *less* reliable than the 5-item AD scale, likely because some of the additional items are reversed. See Appendix for details.

To test whether the AD format inflates correlations with other constructs measured using the same format, we examine the feelings of control scale, described above. We create two control subscales composed of the forward and reverse-worded items ($\alpha = .79$, $\alpha = .77$, respectively). Both subscales are coded so that higher values indicate higher feelings of control. Given that agreement with the hostile sexism statements indicates higher levels of sexism, we expect a stronger positive correlation between sexism and feelings of control for the forward-worded control items than for the reversed control items. Moreover, the differences in correlations should be larger for the AD sexism scale than the IS.

The AD hostile sexism scale is strongly and positively related to the forward-worded control scale ($r = .40$). However, it is weakly and negatively related to the reversed control scale ($r = -.04$), even though the two control scales are designed to measure the same construct. In contrast, the correlations between the IS sexism scale and the two halves of the control scale are both weak (forward: $r = .16$; reversed: $r = -.08$). To test whether these patterns differed between the two scales, we first calculated the difference between the correlations with the forward-worded and reversed scales for each measure of hostile sexism (AD: 0.44, IS: 0.24). Then we differenced the two differences (0.20). To test whether this difference is statistically significant, we used bootstrapping to draw 1,000 random samples from our data and recorded the difference-in-differences in each. Of the 1,000 samples, 99.9% were larger for the AD scale than the IS scale, suggesting the AD format is reliably inflating correlations with other constructs measured with the same response scale.

Predictive validity

To directly compare the two scales' predictive validity, we rescale each to range from 0-1 (for a similar approach, see Bakker and Lelkes 2018). We predict each dependent variable as a

function of hostile sexism, an indicator of the format (AD or IS), and an interaction between the two variables, which tests whether the two scales have the same effects. Additionally, we control for partisanship, ideology, political interest, age, education, gender, and race. Model results are shown in Table 2, and Figure 3 plots the predicted values for each outcome as a function of each version of hostile sexism.

We begin with attitudes toward the #MeToo movement. To simplify and reduce measurement error, we combine these items into a single latent variable using an item response model (see Appendix for details). Higher values indicate greater opposition to the movement and greater resistance to workplace-based solutions to sexual harassment.[10] We follow the modeling approach described above.

As expected, the AD scale significantly predicts #MeToo attitudes ($b = 0.77$, $p < .001$). However, the interaction term ($b = 0.25$, $p = .062$) suggests the IS version has a marginally stronger effect. Notably, the gains in predictive validity occur primarily at the low end of the scale, where the IS scale predicts significantly lower levels of opposition to #MeToo than the AD scale (difference = -0.21, $p = .004$), while the predictions do not significantly differ at the highest values of the scales (difference = 0.04, $p = .576$). Gains at the low end of the scale are useful in helping researchers better understand public opinion in *support* of the movement and policies aimed to address sexual misconduct. While measuring opposition to #MeToo with nuance at the high end of sexism is undoubtedly important, adding nuance to our measurement of the low end is equally critical to understanding ongoing policy debates.

[10] We excluded the backlash items, as they were virtually unrelated to the rest of the scale.

We next analyze feeling thermometer ratings of politicians. To simplify analyses and focus on the role of candidate gender, we subtract the average rating of women candidates from the average rating of men. Because we selected an equal number of Democratic women and men and an equal number of Republican women and men, we assume that this approach isolates the effect of candidate gender from partisanship.

Surprisingly, the AD scale does not significantly predict relative candidate ratings (b = 0.35, p = .806). However, the interaction term indicates that the IS scale has a larger effect than the AD scale (b = 10.12, p < .001), such that higher values on the IS scale predict a relative preference for men. As shown in Figure 3, the IS scale improves prediction at both the low and high ends of the scale. The poor performance of the AD scale is surprising. Auxiliary analyses in the Appendix suggest this is, in part, because the AD scale responds primarily to a candidate's partisanship, rather than their gender.

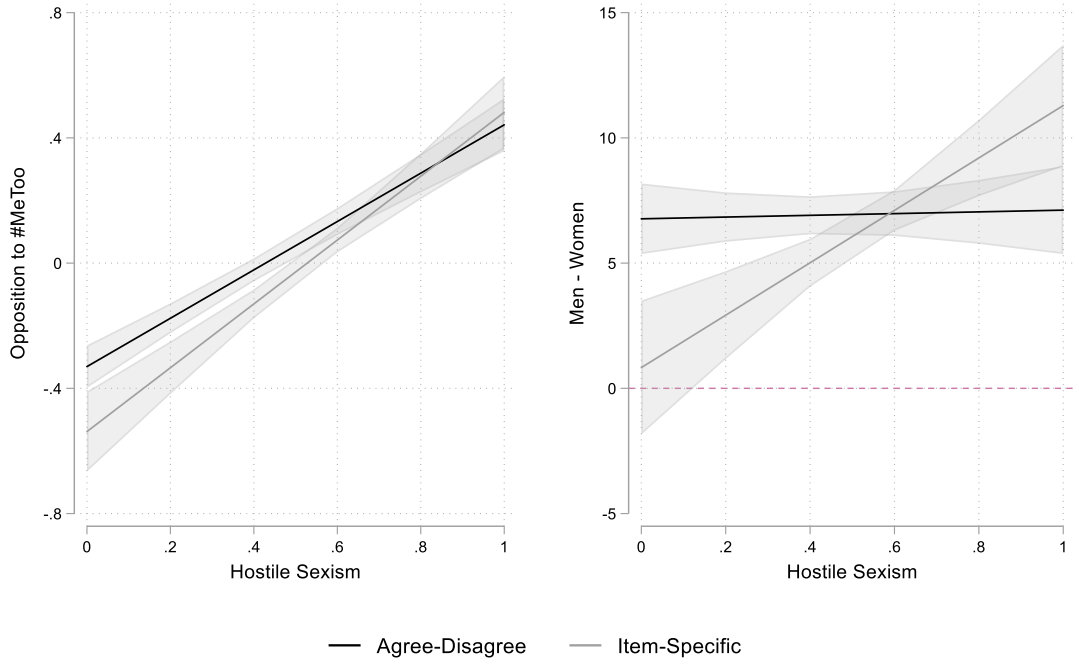
Table 2. Comparing the Predictive Validity of the AD and IS Scales (Study 2)

	Opposition to #MeToo Movement		Feeling Thermometers (Men - Women)	
	Beta (SE)	*p*-value	Beta (SE)	*p*-value
Hostile Sexism	0.772***	0.000	0.347	0.806
	(0.068)		(1.416)	
IS Format	-0.208**	0.004	-5.940***	0.000
	(0.072)		(1.503)	
IS Format × Hostile Sexism	0.247	0.062	10.115***	0.000
	(0.132)		(2.772)	
Partisan Identity	0.050***	0.000	1.397***	0.000
	(0.008)		(0.160)	
Ideology	0.076***	0.000	0.189	0.356
	(0.010)		(0.205)	
Political Interest	-0.057***	0.000	-0.688**	0.006
	(0.012)		(0.248)	
Education	0.001	0.900	-0.532**	0.006
	(0.010)		(0.194)	
Age	0.005***	0.000	-0.133***	0.000
	(0.001)		(0.018)	
Male	0.185***	0.000	-1.686**	0.004

	(0.028)		(0.580)	
Asian	0.043	0.534	-0.281	0.843
	(0.069)		(1.419)	
Hispanic	-0.068	0.307	-0.023	0.987
	(0.067)		(1.412)	
Native American	-0.196	0.143	1.794	0.516
	(0.134)		(2.763)	
White	0.055	0.243	1.091	0.264
	(0.047)		(0.976)	
Other Race	-0.058	0.555	3.783	0.075
	(0.099)		(2.122)	
Constant	-0.997***	0.000	12.281***	0.000
	(0.080)		(1.695)	
Effect of AD Scale	0.772	0.000	0.347	0.806
	(0.068)		(1.416)	
Effect of IS Scale	1.120	0.000	10.463	0.000
	(0.119)		(2.492)	
R^2	0.27		0.13	
N	2343		1993	

Note: bottom panel shows the marginal effect of each scale. For the IS scale, this is the linear combination of the hostile sexism coefficient and the coefficient on the interaction term. All p-values are two-tailed.

Figure 3. Predictive Validity of the Item-Specific and Agree-Disagree Scales



Discussion

Taken together, Study 2 further illustrates how the IS scale improves upon the AD scale. The AD format still faces truncation problems that may weaken the scale's validity. Additionally, the AD format can inflate relationships with other constructs measured with the same response format, potentially creating false positives. Finally, we found some evidence that the IS version increases the hostile sexism scale's predictive validity, particularly at the low end of the scale.

Despite this, there are limitations to the results. While we designed the study as it would likely be carried out by applied researchers, this led to several differences in the visual presentation of the two scales as well as differences in the number of response options, which may have contributed to our findings (Höhne and Krebs 2018).

Study 3

To address the limitations of Study 2, we fielded a third study that holds visual presentation constant across formats. We view Study 3 as a better controlled test of how the two scale formats differ, though less representative of how the scales would typically be used in the field. Respondents were recruited from Mechanical Turk (N=2,003) on September 1, 2021. Respondents were required to have completed at least 100 HITs, earned an approval rate of at least 95%, and passed CloudResearch's approved worker test. Similar to Study 1, the sample tended to be young (median age = 38) and Democratic (46%); 74% of respondents were white

and 60% were college-educated.[11] Approximately 103 respondents completed the study using a smartphone.[12]

For our independent variables, we follow the same design as Study 2, with a few changes to the AD scale's presentation. In contrast to the original hostile sexism scale, we placed the AD questions on a five-point scale with a midpoint. Additionally, AD questions were not placed in a grid, and response options were displayed vertically rather than horizontally. As with Study 2, we measured the primary five items on an initial page and the remaining six items on the following page, but we focus our attention here on the short scale. Both formats are rescaled to range from 0 to 1.

Our criterion variables consist of the same feeling thermometers, questions on the #MeToo movement,[13] and feelings of control over one's life. Consistent with our presentation of the AD hostile sexism scale, the control scale items were presented individually and vertically with 5-point AD response choices. Respondents were first asked the criterion variables, then randomized to one version of the hostile sexism questions, followed by the control scale. Finally, respondents answered demographics questions.

Results

Once again, the IS and AD formats yield reliable scales (IS: $\alpha = .82$; AD: $\alpha = .92$). Truncation was again a problem for the AD scale (25% at the minimum or maximum value), but

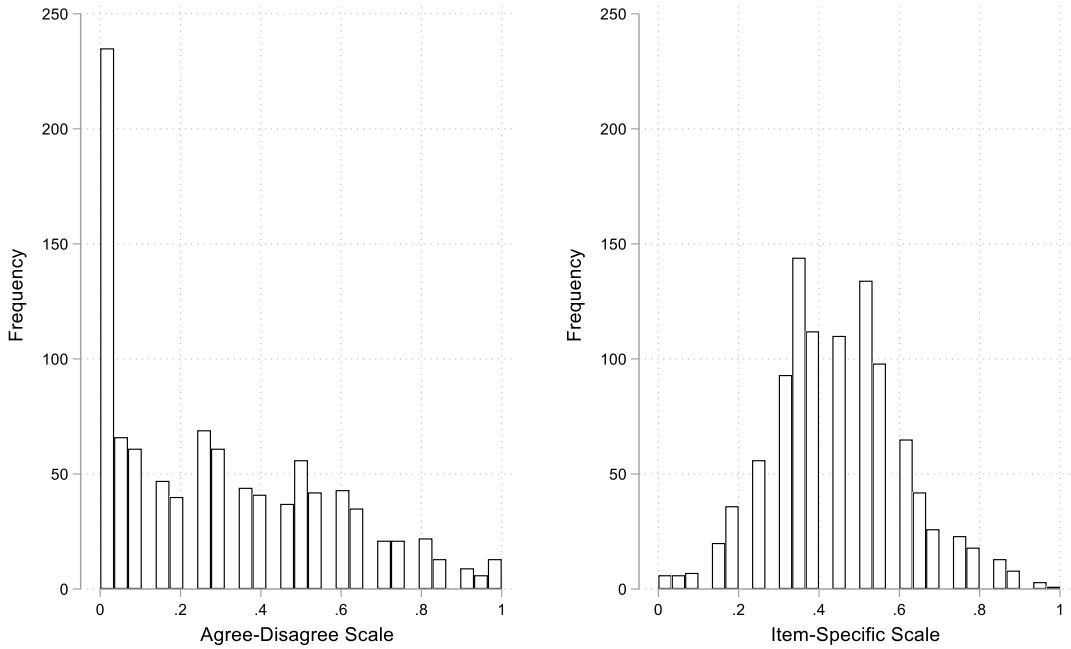
[11] We used the same attention check from Study 2. Attention was higher in Study 3 (99%) than Study 2 (91%), suggesting higher data quality.

[12] We infer smartphone usage from the respondent's operating system.

[13] We omitted the fifth question that did not load on the latent factor in Study 2.

not the IS scale (1%; see Figure 4). The median respondent spent 32 seconds on the IS scale, but only 20 seconds on the AD scale, suggesting modest differences in reaction time.

Figure 4. Distributions of the Item-Specific and Agree-Disagree Scales



Discriminant Validity

For discriminant validity, we compare the correlations between each hostile sexism scale and each of the two halves of the control scale – the four reversed items and four remaining items. The AD scale is positively related to the control scale when the items are worded in the same direction (r = .08), but negatively related when the items are reversed (r = -.11). The IS scale is positively related to the first half of the scale (r = .14), but unrelated to the reversed half of the scale (r = .00). We again tested the difference-in-differences using bootstrapping. In 88%

of cases, the difference in correlations was larger for the AD scale, providing suggestive evidence that the IS scale has better discriminant validity.[14]

Predictive Validity

We first examine attitudes toward the #MeToo movement, which we again scale together using an item response model (see Appendix for details). We use the same modeling approach and control variables described in Study 2 in addition to a control for usage of a smartphone. Model results are shown in Table 3 and predicted values are plotted in Figure 5. As expected, the AD scale predicts greater opposition to #MeToo (b = .95, p < .001). However, the interaction term indicates the IS scale has a significantly larger effect than the AD scale (b = .88, p < .001).

Turning to the feeling thermometers, we again subtract ratings of women from ratings of men and use the same modeling approach as described above. This time, the AD scale predicts more favorable views of men (b = 2.90, p = .027), but the interaction term shows no difference between the two scales (b = -1.16, p = .616).

Table 3. Comparing the Predictive Validity of the AD and IS Scales (Study 3)

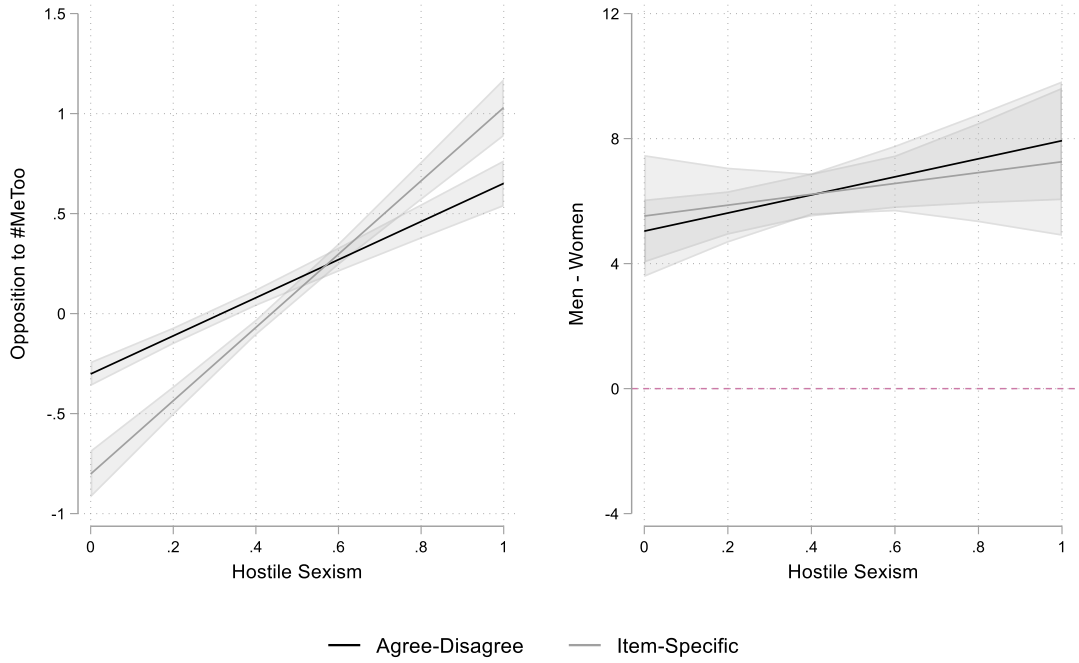
	Opposition to #MeToo Movement		Feeling Thermometers (Men - Women)	
	Beta (SE)	*p*-value	Beta (SE)	*p*-value
Hostile Sexism	0.953***	0.000	2.893*	0.027
	(0.079)		(1.308)	
IS Format	-0.500***	0.000	0.483	0.651
	(0.065)		(1.068)	
IS Format × Hostile Sexism	0.879***	0.000	-1.157	0.616
	(0.140)		(2.306)	
Partisan Identity	0.048***	0.000	1.949***	0.000
	(0.013)		(0.209)	

[14] Unlike Study 2, the two halves of the control scale were moderately related to each other (r = .47), suggesting weaker acquiescence bias. This could be due to the change in format or higher effort from respondents.

Ideology	0.117***	0.000	-1.314***	0.000
	(0.015)		(0.254)	
Political Interest	-0.024	0.124	-0.493	0.053
	(0.015)		(0.254)	
Education	-0.001	0.964	-0.711***	0.000
	(0.011)		(0.187)	
Age	0.001	0.233	-0.072***	0.000
	(0.001)		(0.019)	
Male	0.324***	0.000	0.412	0.400
	(0.030)		(0.489)	
Asian	-0.036	0.596	-1.144	0.306
	(0.068)		(1.118)	
Hispanic	0.020	0.785	-0.840	0.490
	(0.074)		(1.217)	
Native American	0.516*	0.016	-2.781	0.453
	(0.214)		(3.702)	
White	0.112*	0.028	-0.359	0.669
	(0.051)		(0.840)	
Other Race	0.111	0.322	-1.994	0.282
	(0.112)		(1.852)	
Smartphone User	0.019	0.762	-0.668	0.526
	(0.063)		(1.053)	
Constant	-1.082***	0.000	11.800***	0.000
	(0.099)		(1.635)	
Effect of AD Scale	0.953	0.000	2.893	0.027
	(0.079)		(1.308)	
Effect of IS Scale	1.832	0.000	1.736	0.410
	(0.128)		(2.107)	
R^2	0.44		0.09	
N	2000		1958	

Note: bottom panel shows the marginal effect of each scale. For the IS scale, this is the linear combination of the hostile sexism coefficient and the coefficient on the interaction term. All p-values are two-tailed.

Figure 5. Predictive Validity of the Item-Specific and Agree-Disagree Scales



Discussion

Overall, Study 3's results largely reinforce the findings of Study 2, however, the differences between the two scales tended to be smaller. This could be due to differences across studies in scale presentation or differences between samples. The AD scale again showed substantial levels of truncation, while the IS scale did not. The IS scale showed better predictive validity on one of two outcomes, while the evidence was only suggestive for improved discriminant validity. Thus, there are still benefits to the IS scale, even when visual display is held constant.

Conclusion

In recent years, sexism has been increasingly pivotal for national politics, particularly since the 2016 election (Cassese and Barnes 2019; Kam and Archer 2021; Valentino et al. 2018). The subsequent rise in scholarly attention to sexism has highlighted the need for a unified

approach to the measurement of this concept, and hostile sexism seems to be the most promising candidate (Schaffner 2021). However, hostile sexism, like other popular measures of sexism, relies on AD scales that are prone to response bias. As shown across three studies, the AD format exhibits substantial truncation at the lower end of the scale and introduces bias in relationships with other concepts also measured with an AD format. Specifically, when compared with a scale measuring feelings of control, the hostile sexism scale was more strongly related to the feelings of control items worded in the same direction than those worded in the opposite direction. That is, a substantial proportion of the variance in responses to AD scales seemingly captures a general tendency to agree with statements in surveys rather than the intended concept, which biases the relationships between different concepts using AD scales.

To address these issues, we adapted the hostile sexism questions to an IS response format. Across all studies, the evidence suggests the IS version performs at least as well as, but typically better than, the AD version. The IS scale was strongly related to the AD scale, suggesting that we captured the same concept, but the IS scale significantly reduced truncation effects. The IS format also reduced the tendency to bias relationships with other scales. And finally, we provided some evidence that the IS format improves the predictive validity of the scale, particularly at its low end.

One potential drawback of the IS format is that it takes a bit longer for respondents to complete. However, in our second and more diverse study that fielded questions as they are typically used in applied research, the survey time only increased from 32 to 44 seconds. Thus, the costs to survey length seem small relative to the gains in measurement. Additionally, in our first study with more experienced survey-takers, a sizable majority explicitly preferred the IS version. Thus, the increased time seemed worthwhile to respondents.

Taken together, our results suggest that researchers should adopt the IS scale over the AD version. The advantages of the IS format are especially useful considering its reduced left-censoring and its gains in predictive validity at the low end of the hostile sexism scale. These gains may be particularly useful for examining which Democrats are most likely to vote (Kam and Archer 2021) and who is most supportive of the #MeToo movement and policies aimed at reducing sexual misconduct (Archer and Kam 2021). Improving our ability to measure the views of those low in sexism is crucial to continued work in these domains, as their opinions and behavior are consequential for modern elections and policy debates.

There are several limitations to the current study that point to future directions for research. For example, it may be possible to improve on the wording of the new IS items. Questions asking "how often" women do something may be better suited with a "how many women" frame, and vice versa. Modifications that increase differentiation at the middle of the scale might be particularly valuable. Additionally, the optimal set of items for a shortened IS scale may differ from the optimal AD items. Further, although our research relies on samples commonly used by applied researchers, they are not nationally representative. Thus, although experiments conducted on MTurk and on nationally representative samples tend to produce the same treatment effects (e.g., Coppock 2018; Mullinix et al. 2016), there is no guarantee that our findings generalize to the population, particularly the distributions of hostile sexism. Future work evaluating the IS version of hostile sexism with a probability-based sample of the general population in the US and in other countries would provide useful additional validation. Finally, the insights from our three studies suggest other measures of sexism using AD scales (e.g., benevolent and modern sexism) would be improved by adopting the IS format. Future research

might directly compare the AD and IS versions of these scales to provide definitive evidence of this.

The recent surge in research on sexism often features calls for future work to continue examining the effects of sexism on key political outcomes (e.g., candidate evaluations or views on policies aimed to reduce sexual harassment) over time. To do so, a unified approach to the measurement of gender attitudes is needed. Based on prior work (Schaffner 2021) and evidence from our three studies, we encourage researchers to use the IS version of the hostile sexism scale to minimize response bias and measurement error when assessing prejudice against women.

References

Archer, Allison M. N., and Cindy D. Kam. 2021. "Modern Sexism in Modern Times: Public Opinion in the #MeToo Era." *Public Opinion Quarterly*.

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): 1311–25.

Banaszak, Lee Ann, and Eric Plutzer. 1993. "Contextual Determinants of Feminist Attitudes: National and Subnational Influences in Western Europe." *American Political Science Review* 87: 145–57.

Banda, Kevin K., and Erin C. Cassese. 2021. "Hostile Sexism, Racial Resentment, and Political Mobilization." *Political Behavior*.

Barnes, Tiffany D., Emily Beaulieu, and Gregory W. Saxton. 2020. "Sex and Corruption: How Sexism Shapes Voters' Responses to Scandal." *Politics, Groups, and Identities* 8(1): 103–21.

Baron-Epel, Orna, Giora Kaplan, Ruth Weinstein, and Manfred S. Green. 2010. "Extreme and Acquiescence Bias in a Bi-Ethnic Population." *European Journal of Public Health* 20(5): 543–48.

Cassese, Erin C., and Tiffany D. Barnes. 2019. "Reconciling Sexism and Women's Support for Republican Candidates: A Look at Gender, Class, and Whiteness in the 2012 and 2016 Presidential Races." *Political Behavior* 41(3): 677–700.

Cassese, Erin C., and Mirya R. Holman. 2019. "Playing the Woman Card: Ambivalent Sexism in the 2016 U.S. Presidential Race." *Political Psychology* 40(1): 55–74.

Clifford, Scott, Yongkwang Kim, and Brian Sullivan. 2020. "An Improved Question Format for

Measuring Conspiracy Beliefs." *Public Opinion Quarterly*.

Coppock, Alexander. 2018. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods*: 1–16.

Couper, Mick P., Roger Tourangeau, Frederick G. Conrad, and Chan Zhang. 2013. "The Design of Grids in Web Survey." *Social Science Computer Review* 31(3): 322–45.

Fowler, Floyd J., and Carol Cosenza. 2008. "Writing Effective Questions." In *International Handbook of Survey Methodology*, eds. Edith D. de Leeuw, Joop Hox, and Don A. Dillman. New York, NY: Routledge, 136–60.

Glick, Peter, and Susan T. Fiske. 1996. "The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism." *Journal of Personality and Social Psychology* 70(3): 491–512.

Glick, Peter, and Jessica Whitehead. 2010. "Hostility Toward Men and the Perceived Stability of Male Dominance." *Social Psychology* 41(3): 177–85.

Godbole, Maya A., Noelle A. Malvar, and Virginia V. Valian. 2019. "Gender, Modern Sexism, and the 2016 Election." *Politics, Groups, and Identities* 7(3): 700–712.

Höhne, Jan Karem, and Dagmar Krebs. 2018. "Scale Direction Effects in Agree/Disagree and Item-Specific Questions: A Comparison of Question Formats." *International Journal of Social Research Methodology* 21(1): 91–103.

Höhne, Jan Karem, and Timo Lenzner. 2018. "New Insights on the Cognitive Processing of Agree/Disagree and Item-Specific Questions." *Journal of Survey Statistics and Methodology* 6(3): 401–17.

Höhne, Jan Karem, Stephan Schlosser, and Dagmar Krebs. 2017. "Investigating Cognitive Effort and Response Quality of Question Formats in Web Surveys Using Paradata." *Field Methods*

29(4): 365–82.

Javeline, Debra. 1999. “Response Effects in Polite Cultures: A Test of Acquiescence in Kazakhstan.” *The Public Opinion Quarterly* 63(1): 1–28.

Kam, Cindy D., and Allison M. N. Archer. 2021. “Mobilizing and Demobilizing: Modern Sexism and Turnout in the #MeToo Era.” *Public Opinion Quarterly*.

Kennedy, Ryan et al. 2020. “The Shape of and Solutions to the MTurk Quality Crisis.” *Political Science Research and Methods* 8(4): 614–29.

Knuckey, Jonathan. 2019. “‘I Just Don’t Think She Has a Presidential Look’: Sexism and Vote Choice in the 2016 Election.” *Social Science Quarterly* 100(1): 342–58.

Krosnick, Jon A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5(3): 213–36.

Krosnick, Jon, and Stanley Presser. 2010. “Question and Questionnaire Design.” In *Handbook of Survey Research*, eds. P.V. Marsden and J.D. Wright. Emerald, 263–314.

Kuru, Ozan, and Josh Pasek. 2016. “Improving Social Media Measurement in Surveys: Avoiding Acquiescence Bias in Facebook Research.” *Computers in Human Behavior* 57: 82–92.

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): 205316801560417.

McThomas, Mary, and Michael Tesler. 2016. “The Growing Influence of Gender Attitudes on Public Support for Hillary Clinton, 2008-2012.” *Politics & Gender* 12(1): 28–49.

Mirowsky, John, and Catherine E. Ross. 1990. “The Consolation-Prize Theory of Alienation.” *American Journal of Sociology* 95(6): 1505–35.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2016. “The

Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(02): 109–38.

Pasek, Josh, and Jon A. Krosnick. 2010. "Optimizing Survey Questionnaire Design in Political Science: Insights from Psychology." In *The Oxford Handbook of American Elections and Political Behavior*, ed. Jan E. Leighley. Oxford University Press, 27–49.

Plutzer, Eric. 1991. "Preferences in Family Politics: Women's Consciousness or Family Context?" *Political Geography Quarterly* 10(2): 162–73.

Powell, Brian, and Lala Carr Steelman. 1982. "Testing an Undertested Comparison: Maternal Effects on Sons' and Daughters' Attitudes toward Women in the Labor Force." *Journal of Marriage and Family* 44: 349–55.

Revilla, Melanie, Willem E Saris, and Jon A. Krosnick. 2014. "Choosing the Number of Categories in Agree-Disagree Scales." *Sociological Methods & Research* 43(1): 73–97.

Rhodebeck, Laurie A. 1996. "The Structure of Men's and Women's Feminist Orientations: Feminist Identity and Feminist Opinion." *Gender & Society* 10: 386–403.

Roberts, Caroline, Emily Gilbert, Nick Allum, and Leila Eisner. 2019. "Satisficing in Surveys: A Systematic Review of the Literature." *Public Opinion Quarterly* 83(3): 598–626.

Rollero, Chiara, Peter Glick, and Stefano Tartaglia. 2014. "Psychometric Properties of Short Versions of the Ambivalent Sexism Inventory and Ambivalence Toward Men Inventory." *Testing, Psychometrics, Methodology in Applied Psychology* 21(2): 1–11.

Ross, Catherine E, and John Mirowsky. 1984. "Socially-Desirable Response and Acquiescence in a Cross-Cultural Survey of Mental Health." *Journal of Health and Social Behavior* 25(2): 189–97.

Saris, Willem E, Melanie Revilla, Jon A Krosnick, and Eric M Shaeffer. 2010. "Comparing

Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options." *Survey Research Methods* 4(1): 61–79.

Schaffner, Brian F. 2021. "Optimizing the Measurement of Sexism in Political Surveys." *Political Analysis*.

Schaffner, Brian F., Matthew Macwilliams, and Tatishe Nteta. 2018. "Understanding White Polarization in the 2016 Vote for President: The Sobering Role of Racism and Sexism." *Political Science Quarterly* 133(1): 9–34.

Setzler, Mark, and Alixandra B. Yanus. 2018. "Why Did Women Vote for Donald Trump?" *PS: Political Science & Politics* 51(3): 523–27.

Sides, John, Michael Tesler, and Lynn Vavreck. 2018. *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America*. Princeton: Princeton University Press.

Simas, Elizabeth N., and Marcia Bumgardner. 2017. "Modern Sexism and the 2012 US Presidential Election: Reassessing the Casualties of the 'War on Women.'" *Politics & Gender* 13(3): 359–78.

Spence, Janet T., Robert Helmreich, and Joy Stapp. 1973. "A Short Version of the Attitudes toward Women Scale (AWS)." *Human Memory, Learning, & Thinking* 2: 219–20.

Swim, Janet K., Kathryn J. Aikin, Wayne S. Hall, and Barbara A. Hunter. 1995. "Sexism and Racism: Old-Fashioned and Modern Prejudices." *Journal of Personality and Social Psychology* 68: 199–214.

Thomas, Kyle A., and Scott Clifford. 2017. "Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments." *Computers in Human Behavior* 77: 184–97.

Tougas, Francine, Rupert Brown, Ann M. Beaton, and Stephane Joly. 1995. “Neosexism: Plus Ça Change, Plus c’est Pareil.” *Personality and Social Psychological Bulletin* 21(8): 842–49.

Utych, Stephen M. 2020. “Sexism Predicts Favorability of Women in the 2020 Democratic Primary... and Men?” *Electoral Studies*.

Valentino, Nicholas A., Carly Wayne, and Marzia Oceno. 2018. “Mobilizing Sexism: The Interaction of Emotion and Gender Attitudes in the 2016 US Presidential Election.” *Public Opinion Quarterly* 82(S1): 799–821.

Weijters, Bert, Maggie Geuens, and Niels Schillewaert. 2010. “The Stability of Individual Response Styles.” *Psychological Methods* 15(1): 96–110.

Zhang, Xijuan, Ramsha Noor, and Victoria Savalei. 2016. “Examining the Effect of Reverse Worded Items on the Factor Structure of the Need for Cognition Scale.” *PLOS ONE*.