Polling Across Borders:

The Promise and Pitfalls of Convenience Samples in a Cross-National Context

Abstract

This study evaluates the performance of two widely used survey platforms, Lucid and Morning Consult, across six diverse national contexts: Brazil, India, Japan, Nigeria, the Philippines, and the United States. We assess the impact of platform choice on sample composition, response quality, political judgments, and treatment effect estimation, focusing on a randomized corruption treatment embedded within the survey. Attention filter passage rates were similar and generally high across countries and platforms, while the percentage of high frequency survey-takers varied greatly across countries. Our findings reveal significant demographic skews, with both platforms consistently overrepresenting college-educated respondents. Despite these differences, political assessments and estimated average and heterogeneous treatment effects remain broadly consistent across platforms, and in Brazil our estimates largely tracked those from past research with a probability sample. We find some evidence of cross-national variation in the magnitude of treatment effects, but these differences were often platform-specific. These results suggest that convenience samples can provide reliable estimates of causal effects even in diverse contexts. Taken together, our research highlights the trade-offs between cost, speed, and representativeness in global public opinion research, offering insights into the challenges and opportunities of online survey platforms.

Introduction

Since George Gallup pioneered the scientific study of public opinion in 1935, surveys have become both more global and more accessible, fueled by the spread of internet and mobile technology, making it a growing, multibillion dollar, international industry (Business Research Company 2025). Within this landscape, survey experiments have become central to public opinion research because they allow scholars to identify causal effects behind political attitudes and behaviors. Yet the platforms that enable low-cost, rapid data collection—such as MTurk, Prolific, or Lucid—raise questions about generalizability. The sheer cost of probability-based samples can make them untenable for many cross-national or longitudinal studies. As a result, researchers often turn to opt-in convenience samples that can vary significantly in terms of sampling strategies, demographic composition, and data quality. Reliance on convenience samples, however, raises questions about the comparability of inferences and estimated experimental treatment effects across platforms and contexts.

Despite the growth in the global polling market, academic methodological critiques of polling techniques and convenience samples are still concentrated in the United States (e.g., Hillygus & Guay 2016; Krupnikov et al. 2021; Berrens et al. 2017; Kam et al. 2007; though for an assessment of the relative performance of convenience samples in Brazil and India, see Boas et al. 2020). In this study, we empirically assess the performance of two survey platforms—Morning Consult (MC) and Cint/Lucid—across six national contexts: Brazil, India, Japan, Nigeria, the Philippines, and the United States. These countries are spread across four continents and vary in political and economic systems, digital penetration, and survey norms, making them ideal for testing whether convenience samples produce consistent findings across contexts.

Morning Consult recruits quota-based samples from multiple large opt-in panels, conducts "digital fingerprinting" to protect against duplicate responses and participants who do not meet

recruitment criteria and applies post-stratification weights for representativeness (see Supporting Information Section B for more details). Lucid relies on lower-cost panels and does not employ quota-based sampling.[1] Kennedy et al. (2016) evaluate multiple convenience sample providers in the U.S. and show that vendor differences in recruitment and sampling design translate into measurable differences in data quality. Here, we assess the comparability of estimates of political quantities of interest and average and heterogeneous experimental treatment effects across these two platforms and highly diverse national contexts.

Our findings speak to ongoing methodological debates about representativeness and how estimates of experimental treatment effects and political quantities of interest vary across survey platforms in different national contexts. Echoing previous research (Berinsky, Huber & Lenz 2012; Hillygus & Guay 2016) our findings highlight persistent challenges with sample representativeness from non-probability-based samples, with and without quota-based sampling. Both platforms struggle to produce representative samples without significant weighting, consistently overrepresenting more educated respondents and exhibiting variable gender imbalances. The pattern is particularly evident in Lucid samples.[2] These biases raise questions about the generalizability of survey findings from online, opt-in panels and point to the importance of careful sample management when using convenience platforms for cross-national research.

In addition, we find that while platform differences emerged for trust in government in the United States and Brazil, with MC respondents reporting higher average trust than their Lucid counterparts, these discrepancies were not evident in the other four countries and were small for

[1] In the United States, Cint offers the Lucid Theorem service, which does provide quota-based survey samples matched to U.S. Census demographics. However, internationally (and including in our U.S. sample analyzed here), Lucid samples are not quota-based.

[2] It is important to note that in the United States and Japan, Morning Consult samples are advertised as representative of the general adult population. In Brazil, Nigeria, India, and the Philippines, they are advertised as representative of the online adult population.

alternative trust measures, such as social trust and trust in financial institutions. This pattern suggests that these isolated platform effects may reflect representativeness challenges in specific contexts, while the broader convergence across platforms supports the view that trust is shaped by deep-seated institutional and cultural factors, making it a relatively stable attitude despite differences in survey methods.

Finally, in our mayoral corruption experiment we find little evidence of significant within-country differences in treatment effects across platforms, particularly when using survey weights. While the estimated effect of the corruption treatment is negative across contexts, we do find some significant cross-national variation in the estimated treatment effect, but these differences are platform-specific. Our design can document this heterogeneity, but it cannot adjudicate among the many possible explanations for why treatment effects vary across countries. Finally, we explore treatment effect heterogeneity across gender and educational divides and find broadly consistent evidence across countries and platforms.

This paper makes three core contributions to the study of cross-national public opinion research and online survey platforms. First, on representativeness, it situates itself within a well-established line of evaluation studies showing that convenience samples differ descriptively from population benchmarks (Berinsky, Huber & Lenz 2012; Huff & Tingley 2015; Hillygus & Guay 2016; Coppock & McClellan 2019). We find that the discrepancies between population benchmarks and sample estimates are quite large in some cases, even with quota-based sampling and post-stratification weighting. Second, past work has repeatedly shown that estimates of experimental treatment effects are often comparable across probability-based and convenience based samples including student samples (Druckman & Kam 2011) and those recruited via MTurk (Berinsky, Huber & Lenz 2012; Mullinix et al. 2015; Coppock 2019; Coppock, Leeper & Mullinix 2018) and Lucid (Coppock & McClellan 2019) in the United States (for evidence in Brazil and India, see Boas

et al. 2020), even during COVID (Peyton, Huber & Coppock 2022). Other studies relying solely on convenience samples have explored the generalizability of international relations experiments to non-U.S. contexts (Bassan-Nygate et al. 2024) and found little evidence of differential treatment effects across "eager" and "reluctant" respondents (Moniz et al. 2024). Our study builds on prior work by providing the first systematic cross-national evaluation of how estimates of political quantities of interest and experimental treatment effects vary across two platforms, MC and Lucid, that employ different sampling strategies and data quality assurance procedures across six diverse country contexts. And in the case of Brazil, we are able to compare our estimates to those obtained from a probability sample benchmark from previous research (Weitz-Shapiro & Winters 2017).

Finally, the study contributes to debates about the portability of experimental findings by showing broadly similar estimates of average and heterogenous treatment effects across countries and platforms within country. The general pattern of findings suggests that, across many contexts, human subjects respond to experimental stimuli in broadly similar ways, supporting cross-national comparability.

Research Design

This study assesses whether the representativeness, estimates of political quantities of interest, and experimental treatment effects vary systematically across two widely used online platforms, Morning Consult and Lucid/Cint, by fielding parallel surveys across six countries. We treat Morning Consult as our baseline platform because of its greater investment in panel infrastructure, data quality assurance procedures such as digital fingerprinting to guard against duplicate and non-eligible responses, quota management, and post-stratification weighting, which are intended to produce samples that more closely approximate population benchmarks. At the same time, it is important to underscore that Morning Consult is not probability-based: like Lucid, it relies on opt-in recruitment, and quota-based sampling by itself does not guarantee

representativeness.[3] As Bethlehem (2010) and Cornesse et al. (2020) emphasize, weighting can reduce observable imbalances but cannot fully eliminate selection bias in nonprobability samples. Our use of Morning Consult as a baseline is therefore relative rather than absolute, reflecting its more structured and resource-intensive approach compared to Lucid, while recognizing that both remain subject to the inherent limitations of online opt-in sampling.

We fielded the surveys with the two platforms in six countries – Brazil, India, Japan, Nigeria, the Philippines, and the United States, a diverse set of countries with varying contextual factors across four continents – between January and March 2025 (see Supplemental Information for additional detail on case selection). The surveys were administered in Portuguese in Brazil, Japanese in Japan, and English in the United States, and respondents had the option of choosing between English and Hindi in India, Hausa in Nigeria, and Tagalog in the Philippines. MC generated weights to better approximate the population in each country based on a set of demographics that varied across countries (see Supporting Information Section B). We generated weights based on gender, age, race, and ethnicity for our Lucid samples.

In each country and across platforms, we fielded an identical survey instrument with several components. After several introductory questions the survey began with two screeners employed by MC (See Supporting Information Section B). Subjects that answered incorrectly were exited from the survey. The first substantive block assessed demographics and trust in government. Scholars have studied trust as a correlate of participation in the political process and compliance with the law because they believe that the government and its associated functions are legitimate (Marozzi 2014).

To assess treatment effects, we embedded a randomized survey experiment on attitudes toward corruption. Adapted from Weitz-Shapiro & Winters (2017), respondents were presented with a hypothetical mayoral candidate accused (or not accused) of corruption. Respondents were

[3] Quota-based sampling will not address the problem of MNAR bias and could conceivably introduce bias.

randomly assigned to two experimental conditions, a "clean" control, without mention of corruption and a "corrupt" condition in which a non-governmental organization points to corruption in the mayor's awarding of government contracts; all respondents were asked how likely they would be to vote for the candidate on a four-point scale.[4] Towards the middle of the survey, we included a standard attention check question. This design allows for direct comparisons of how platform choice affects the representativeness, cost-efficiency, and replicability of experimental research conducted in diverse global contexts.

Table 1 provides a comparison of sample sizes and per-interview costs across six countries for both MC and Lucid survey platforms. The United States has the largest sample sizes, with 2,000 respondents from each platform, while the remaining countries maintain sample sizes of roughly 1,000 per platform. Interview costs on Lucid vary significantly across countries, ranging from $1.50 per interview in the United States to $4.75 in Nigeria, reflecting both local market conditions and platform pricing strategies. Japan and the Philippines fall in the middle of this range, at $3 and $3.25 per interview, respectively. Morning Consult prices per respondent – which included a range of additional services including translation, survey programming, and team consultations, among others – were higher in each case ranging from only modestly higher in Nigeria ($5.80 vs. $4.75) to almost ten times higher in the United States ($11.41 vs. $1.50).

[4] Full wording and additional results dichotomizing the dependent variable are reported in SI.

Table 1. Survey Platform Comparison of Sample Size and Cost across Countries

Country	Morning Consult			Lucid		
	Sample Size	Cost/Interview	Dates	Sample Size	Cost/Interview	Dates
Brazil	1,027	$11.60	1/8-12/25	1,000	$2	3/6-12/25
India	1,011	$5.80	2/1/25	1,055	$2	3/4-6/25
Japan	1,007	$11.60	2/19-21/25	1,000	$3	3/11-20/25
Nigeria	1,012	$5.80	2/1-11/25	1,001	$4.75	3/6-11/25
Philippines	1,016	$11.60	1/27 -2/19/25	1,126	$3.25	3/6-11/25
United States	2,108	$11.41	12/16-17/24	2,000	$1.5	3/4-12/25

Results

We first present results on respondent attention checks across platforms and countries. The analysis distinguishes between weighted and unweighted estimates. As shown in the top panel of Figure 1, attention check pass rates were generally high, but they varied across countries, with MC generally outperforming Lucid. The Philippines was an exception, where Lucid performed better: about 70% of weighted Lucid respondents passed, compared to only about 50% of weighted MC respondents. Weighting had a modest but inconsistent impact, slightly increasing pass rates in some contexts (e.g., Japan) while reducing them in others (e.g., Philippines). Notably, the passage rates in India and the U.S. across both platforms were higher than those reported for Qualtrics panel and Facebook respondents in Boas et al. (2020, 245) and in the U.S. higher than the comparable figure in a nationally representative SSI panel (Berinsky, Margolis & Sances 2014).

Another concern with opt-in panels centers on "professional" survey-takers. To measure the frequency with which respondents took surveys across countries and platforms, we asked each respondent how many surveys they take in an average week. As shown in the bottom panel of Figure 1, the share of respondents who report taking more than 7 surveys per week varied significantly across countries. The greatest of self-reported high frequency survey takers was in the United States where roughly 60% of respondents reported averaging taking more than one survey per day across both platforms. Japan followed closely behind with approximately 50-60% of respondents identifying as high frequency survey takers. By contrast only 10-25% of respondents in India, Nigeria, and the Philippines reported taking seven or more surveys per week, and Brazil fell in the middle of these ranges with approximately 40% identifying as high frequency survey takers. We observe only modest differences across platforms within each country.

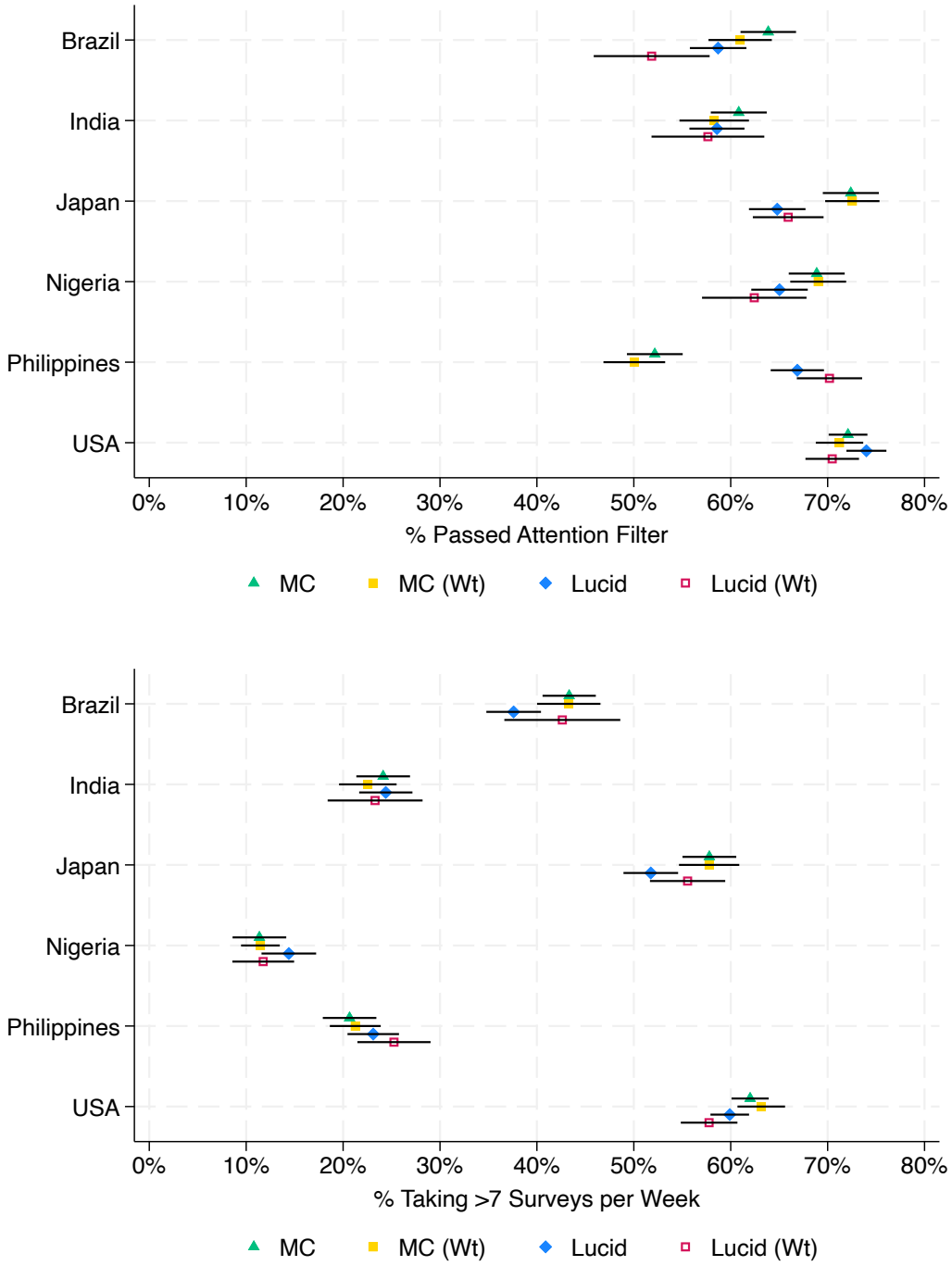

Figure 1. Top panel: Percentage of respondents passing attention filters across six countries (Brazil, India, Japan, Nigeria, Philippines, and the United States) for two survey platforms (Lucid and MC). Bottom Panel: The percentage of respondents saying they take more than seven surveys per week across countries and platforms. Results are reported both with and without weighting to reflect known population characteristics. Error bars represent 95% confidence intervals.

Demographic and Political Representativeness

To compare the demographic representativeness of higher vs. lower quality opt-in panel platforms across contexts, Figure 2 presents sample demographics for gender, age, and educational attainment across countries and platforms, with and without weights, alongside population benchmarks[5] For gender and age, unweighted Lucid sample demographics routinely deviated more from population values than did MC samples. This is to be expected given MC's use of quota-based sampling. The magnitude of these deviations differed across countries, and weighted Lucid estimates generally reflected population benchmarks. Somewhat surprisingly, even with weights the MC sample sometimes diverged significantly from the adult population baseline, particularly for gender and age in India, and for age in Nigeria and the Philippines.

The starkest differences from population parameters occurred with respect to education. In every country except the United States, college educated respondents were significantly over-represented across platforms, a common challenge in online survey platforms. In Brazil, Japan, and the Philippines, the skew towards higher education respondents was much more pronounced in the Lucid samples. The generally inflated estimates across countries (aside from the U.S.) and platforms may reflect the digital divide, where more educated individuals are more likely to have reliable internet access and be familiar with survey-taking platforms (though this is unlikely to explain over-representation in Japan). Neither sample was weighted with respect to education. The large discrepancies between weighted estimates and educational population parameters clearly shows that weighting on other demographic characteristics will not necessarily produce more representative outcomes on other dimensions.

[5] Population median ages are estimates from WHO population pyramids for adults 18 and older. See SI for more information.

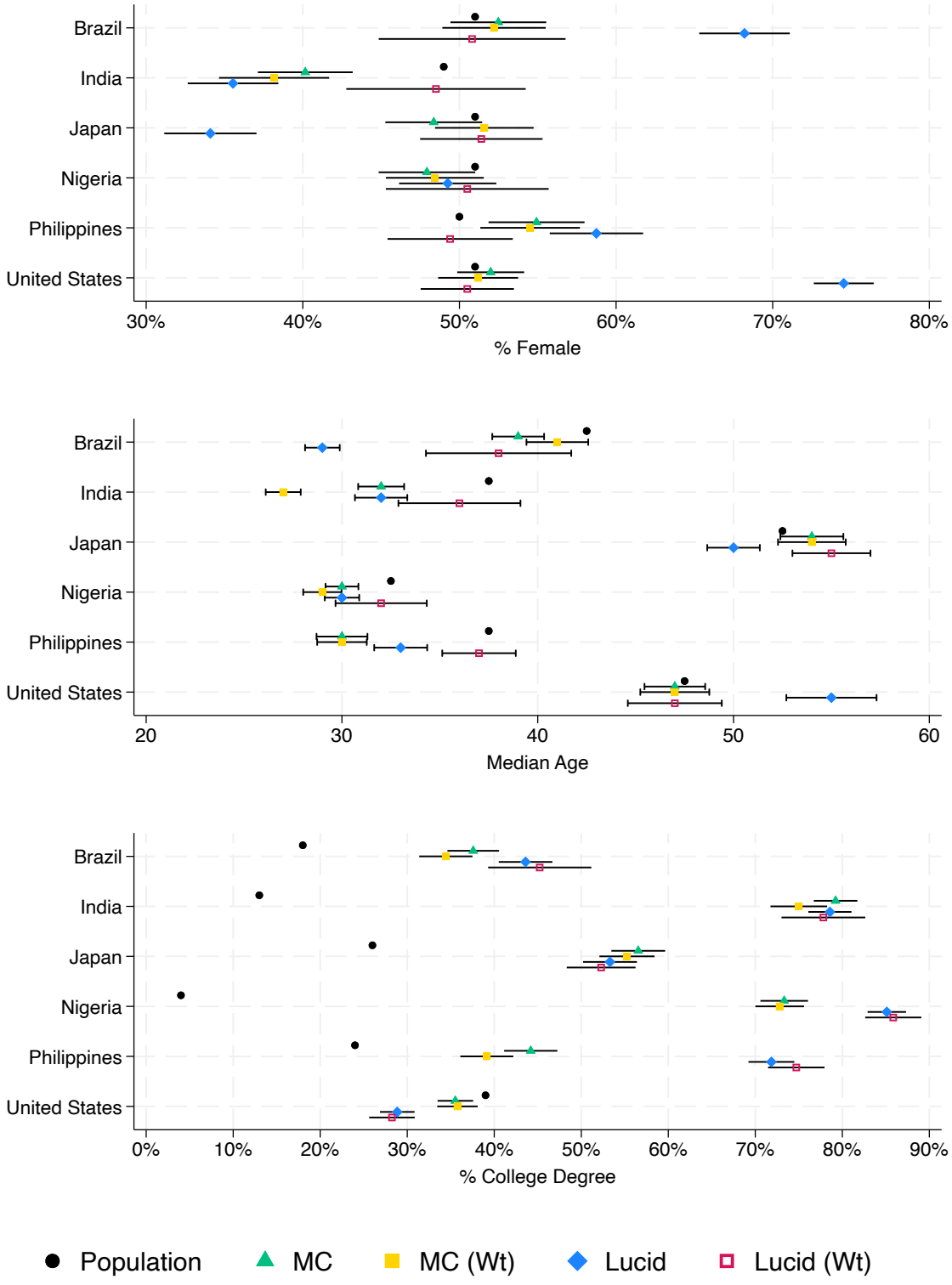

Figure 2. Proportion of female respondents, median age (for adults 18+), and percentage with a college degree across countries and platforms compared to population benchmarks. Results are shown with and without weighting. Horizontal lines indicate 95% confidence intervals.

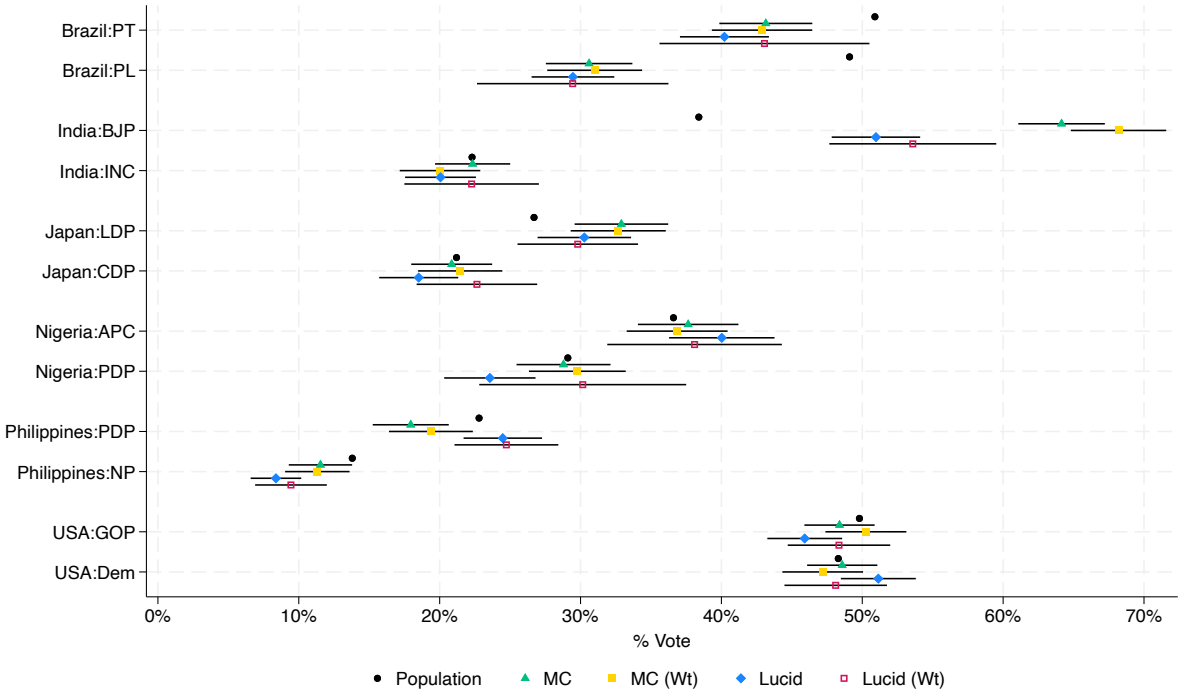

Figure 3. Support for the two largest parties in the last general election in each country across countries and platforms compared to actual election results.

Finally, our survey also allowed us to compare political characteristics of samples across surveys through self-reported partisan voting in the last general election to actual election outcomes. Figure 3 summarizes the results for the two largest parties across countries and platforms.[6] The disparities between self-reported voting behavior (albeit a far-from-perfect measure) and actual election results vary significantly across countries, as well as across platforms within the same country in some cases. In the U.S., Philippines, Nigeria, and Japan, the disparities are relatively modest. In India, both Lucid and particularly MC samples significantly over-represented BJP voters. And both platforms produced estimates that under-represented support for the two leading parties in Brazil. As we discuss in more detail in the Supporting Information Section D an over-representation of urban respondents may be contributing to the significant over-estimation of

[6] For comparisons of smaller parties, see Supporting Information, SI Table 2.

support for BJP in India. The Brazil case is complicated by the simultaneous elections for president and the Chamber of Deputies. The MC and Lucid shares of reported support for the PT and PL lag the vote shares received by Lula and Bolsinaro, but exceed the vote shares the two parties received in the Chamber of Deputies elections.

Experimental Treatment Effects

To assess the generalizability of estimated experimental treatment effects across platforms and countries, we replicated the mayoral corruption experiment in Weitz-Shapiro and Winters (2017). Across all six countries and in both platforms, the alleged corruption treatment significantly decreases electoral support for the hypothetical mayoral candidate (left panel of Figure 4). In the unweighted data we observe significant differences in estimated treatment effect sizes across platforms in four of six countries (right panel of Figure 4); however, the direction of this difference is not consistent. In Brazil and Japan, the estimated effect was greater in the MC sample than in the Lucid sample; in India and the U.S. the reverse was true. Interestingly, all of these differences in estimated treatment effects are no longer statistically significant when using survey weights, however, in the Philippines the estimated treatment effect in the weighted Lucid sample is significantly greater than in the weighted MC sample. The overall pattern suggests that both lower-cost and higher-quality convenience samples produce broadly consistent estimates of treatment effects, even in diverse political contexts.

Although not our prime focus, our data does show some evidence of treatment effect heterogeneity across countries.[7] The MC estimate (weighted and unweighted) for India is significantly lower than for any other country. Similarly, the MC estimate for Brazil (weighted and unweighted) is the largest among all six countries, and it is significantly greater than the corruption

[7] Our non-U.S. samples are sufficiently large to detect differences of approximately .25 points on the 4-point DV at alpha = .05 with 80% power. For greater discussion of statistical power, including for detecting heterogenous treatment effects, see Supporting Information Section E.

treatment effect estimates for the Philippines and the U.S. Here, however, we see important differences across platforms as the Lucid estimate for Brazil is in the middle of the range and lower than for the Philippines and U.S.; indeed, these latter differences are statistically significant (in the opposite direction of the MC results), but both differences become statistically insignificant when using weights. This suggests that estimates of cross-country heterogeneity in treatment effects may be more sensitive to differences across survey samples and platforms.

Taken together, these results indicate that while cross-national differences in treatment effects may be platform-specific, within countries MC and Lucid returned broadly comparable treatment effect estimates, particularly when using weights. This suggests, at least in this specific context, that lower-cost Lucid samples generally return treatment effect estimates comparable to those of higher quality quota-based samples.

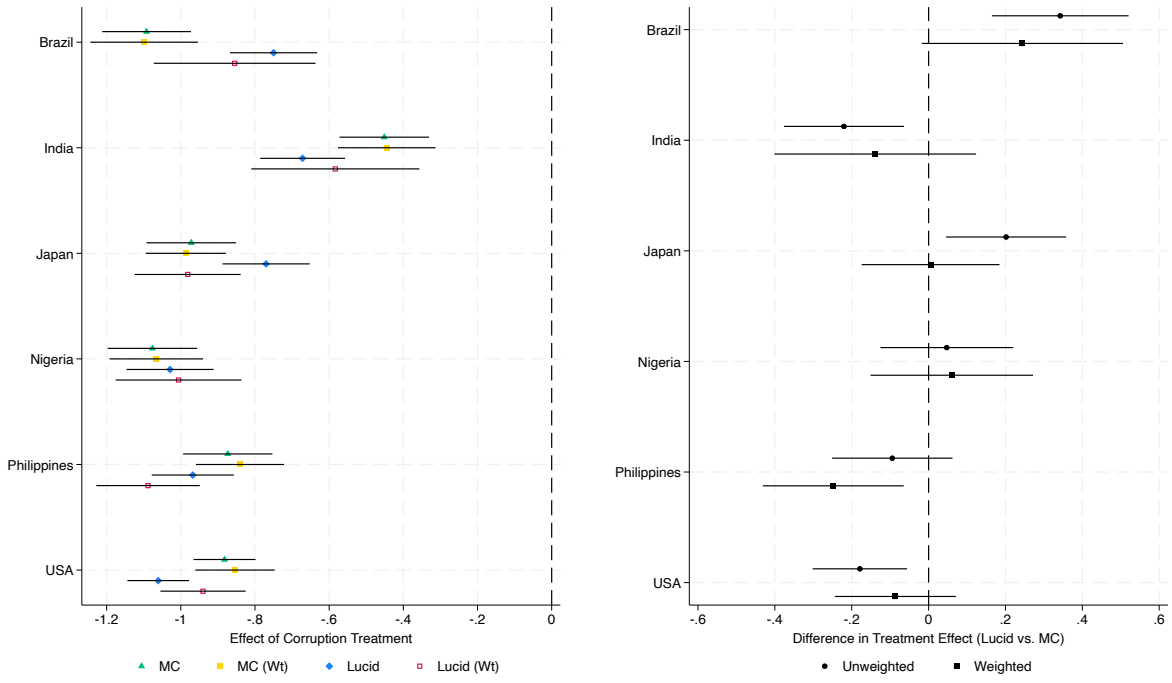

Figure 4. Left Panel: Estimated effects of the corruption treatment on survey responses across countries and panels. Right panel: Difference in treatment effects within countries by platform. Error bars represent 95% confidence intervals.

Heterogeneous Treatment Effects

We also evaluated differences across countries and platforms regarding heterogeneous treatment effects. Two factors in particular are prominent in the corruption literature: gender and education.[8] Recent research finds that whether women are less tolerant of corruption is at best highly contextual (Esarey & Chirillo 2013; Alatas et al. 2009), and possibly stronger in democratic countries. We assess whether the corruption penalty is higher among women, whether this gender penalty varies across contexts (if democracy were associated with the gender penalty, we would expect Japan and the US to show higher penalties among women and Nigeria and the Philippines to show the lowest penalty), and whether our within-country estimates of gender heterogeneity vary across platforms.

The left panel of Figure 5 presents the estimated gender differences in the corruption treatment effect across six countries. In no case do we find evidence of a significantly greater corruption penalty for women than for men. Most estimates are substantively small and statistically insignificant. The exceptions are India and the United States where we find some evidence that women punished corruption less than men; in the former the unweighted MC and Lucid estimates are statistically significant ($p = .08$ and $p = .06$, respectively), and in the latter the unweighted and weighted Lucid estimates are significant, ($p = .04$ and $p = .07$, respectively). The smaller CIs for the U.S. estimates reflect the larger sample size (see Supporting Information Section E for a discussion of statistical power).

The right panel of Figure 5 test for differences in estimated heterogeneous treatment effects across platforms. In no case are the estimated differences in treatment effects significantly different across the two platforms and in every country but Brazil the estimated differences are quite small in

[8] On our survey, both gender and educational attainment were measured before the experiment.

magnitude. Overall, we find strikingly little evidence of gender differences in corruption tolerance across contexts and platforms.

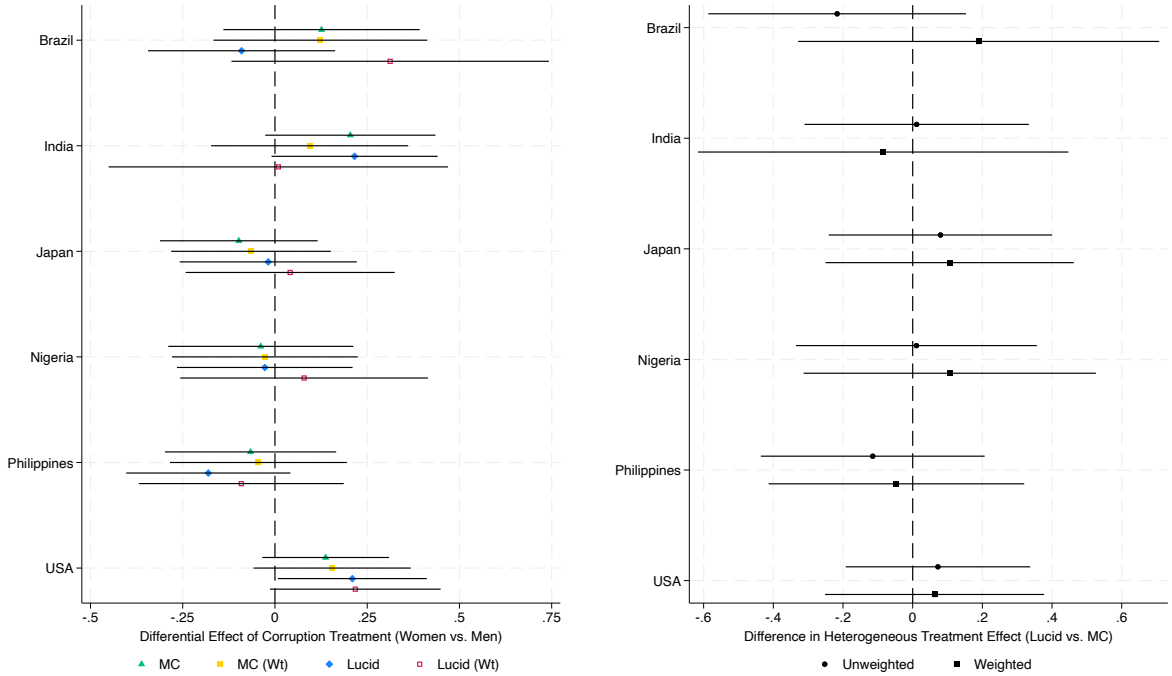

Figure 5. Estimated gender differences in the effect of the corruption treatment across countries and platforms. Positive values indicate a smaller negative treatment effect for female respondents compared to male respondents. Results are shown with and without weighting to reflect known population characteristics. Error bars represent 95% confidence intervals.

Past work also suggests that education is negatively associated with tolerance for corruption (Gouda & Park 2016; Truex 2011; Melgar, Rossi & Smith 2010, Lavena 2013). We therefore expect that the corruption treatment effect will be larger (i.e. more negative) among respondents with higher levels of education. We evaluate whether the corruption penalty is higher among college-educated respondents and whether there is variation across countries and survey platforms.

While our statistical power is somewhat limited (see Supporting Information Section E), the left panel of Figure 6 shows fairly consistent evidence across countries that education strengthens the corruption penalty. In the US, India, Philippines, and Brazil, multiple estimates of educational

heterogeneity in treatment effects are statistically significant. We caution that our sample sizes limit cross-country comparisons. For example, the unweighted MC estimate for India is negative and statistically significant, while the corresponding estimate for Brazil is not significantly different from zero; however, the two negative estimates are not significantly different from one another. However, the generally consistency in the direction of the estimated heterogeneity is striking. Finally, the right panel of Figure 6 presents the differences in estimated heterogeneous treatment effects within countries across platforms. We find no evidence of significantly different estimates across platforms, with or without survey weights.

Despite some significant differences, our study generally affirms that lower-cost, non-quota-based convenience samples like Lucid can consistently reproduce treatment effects, including heterogeneous treatment effects across groups, that largely reflect those observed in higher quality, but higher cost convenience samples such as those provided by Morning Consult.

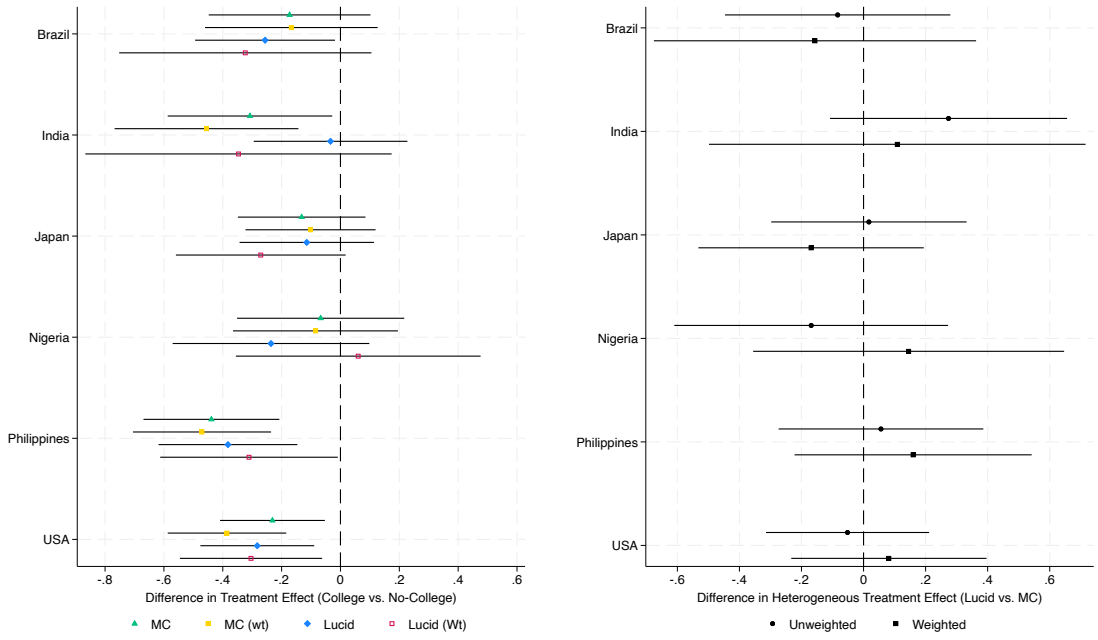

Figure 6. Estimated education differences in the effect of the corruption treatment across countries and platforms. Negative values indicate a larger treatment effect for college-educated respondents compared to those without a college degree. Results are shown with and without weighting to reflect known population characteristics. Error bars represent 95% confidence intervals.

Sensitivity Analysis

Despite the strong similarities in estimated average and conditional treatment effects across platforms, a major concern with generalizing from convenience samples to general populations is an omitted moderator that influences both selection into the sample and the magnitude of the treatment effect. In the specific context of the corruption experiment, the preceding analyses suggest that gender disparities between our convenience samples and the general population are unlikely to skew our average treatment effect estimates, but the greater share of college educated respondents in our samples may influence generalizability. We can never be certain that we have identified all possible treatment effect moderators. But we can conduct a sensitivity analysis to see how strong an omitted moderators would have to be to overwhelm our estimated average treatment effect.

Accordingly, we follow the method developed by Cinelli and Hazlett (2020) and Huang (2024) to estimate Robustness Values for our unweighted MC and Lucid estimates across each country. In most cases, these tests suggest that an omitted moderator would have to explain about 40% of the residual variance in both the treatment and the outcome to reduce our estimated average treatment effect to 0. Full results and bias contour plots that help visualize this sensitivity to an omitted moderator relative to observed benchmarks for gender and college educational attainment are reported in the Supporting Information, SI Table 3 and SI Figures 1-6. While it is likely there are MNAR biases in both the MC and Lucid samples, an omitted moderator would have to be much more powerful than educational attainment to seriously degrade our estimated average treatment effect.

Direct Comparisons of Brazil Results to Prior Research

A limitation of our research is that we can only compare estimated average and heterogeneous treatment effects across a higher-quality, quota-based sampling convenience sample provider and a more economical provider that does not use quota-based sampling. However, we do have one opportunity to compare our findings to those from a probability-based sample: Weitz-Shapiro and Winters' (2017) results in Brazil. The top panel of Figure 7 directly compares the magnitude of the estimated treatment effects from the MC and Lucid samples to those reported in Weitz-Shapiro and Winters. As noted above, our treatment wording was slightly adapted from the original Weitz-Shapiro and Winters study in Brazil. That study had two corruption treatments, one in which the allegation of corruption came from a federal audit – a credible source – and another in which the allegation came from the opposition party – a less credible source. Because the term "federal audit" would not travel across all the contexts in our study, we adapted language from a follow-on study by Winters and Weitz-Shapiro (2020) in Argentina and made the allegation from an "independent organization."[9] While we have little in the way of firm priors, we speculated that many respondents may deem this source less credible than a formal government audit, but more credible than an allegation by the opposition party. The MC treatment effect estimates (both weighted and unweighted) fall in between the more and less credible WSW treatment effect estimates. Thus, in the context of this specific experiment in Brazil, the MC sample produced treatment effect estimates that were very similar to those of a higher quality probability-based sample.[10] The MC-Lucid differences in Brazil (see Figure 4 in the text) were among the largest of the six countries in our study. And the estimated treatment effects in the Lucid sample were

[9] In the Argentina experiment, Winters and Weitz-Shapiro attribute the allegation to "an independent NGO." We used "an independent organization" to eliminate any possible confusion over the meaning of NGO across the varied countries in our sample.

[10] The Weitz-Shapiro and Winters (2017) survey was fielded by the Brazilian Institute of Public Opinion and Statistics to 2,002 individuals across 25 of Brazil's 27 states. The study employed a multistage sample with probability proportional to size sampling of cities across the states and then quota sampling at the level of the individual.

significantly smaller than those in the WSW study, but much closer than many of the convenience sample replicates from probability benchmarks reported in Boas et al. (2020).

The middle panel of Figure 7 examines heterogeneous treatment effect differences between men and women. Across the WSW, MC, and Lucid samples, the estimates are substantively small (with the weighted Lucid estimate being considerably larger) and not statistically significant. Thus, we see little evidence of significant differences in estimates across platforms and sampling strategies.

Finally, the bottom panel of Figure 7 compares estimates of heterogeneous treatment effects between high and lower-educated Brazilians across samples. The WSW sample is much closer to population benchmarks and includes an indicator for those with some tertiary education. We created the same variable for our MC and Lucid samples and interacted it with assignment to the corruption treatment. Each estimate (apart from the weighted Lucid estimate) is negative and of generally comparable magnitudes. The small sample sizes restrict our ability to detect small differences. However, the general comparability of results across sampling strategies and modes is notable.

Political Evaluations: Trust in Government

We conclude by evaluating differences in an important political assessment: trust in government. As the top panel of Figure 8 shows, trust in government is generally low across countries, and the level is surprisingly similar across countries. On a scale of 0-10, respondents' trust in their national government hovers around 4. This is true regardless of levels of economic development, perceptions of corruption, or democracy, as countries on the top and bottom of these distributions showed very similar levels of trust in government. The clear exception is India, where across platforms we observe estimates of approximately 6.5 on the 0-10 scale, the only country in which trust in government is above the midpoint of the scale. The popularity of the Modi

administration in India, despite ongoing corruption concerns, may also contribute to the comparatively higher trust scores observed in that context.

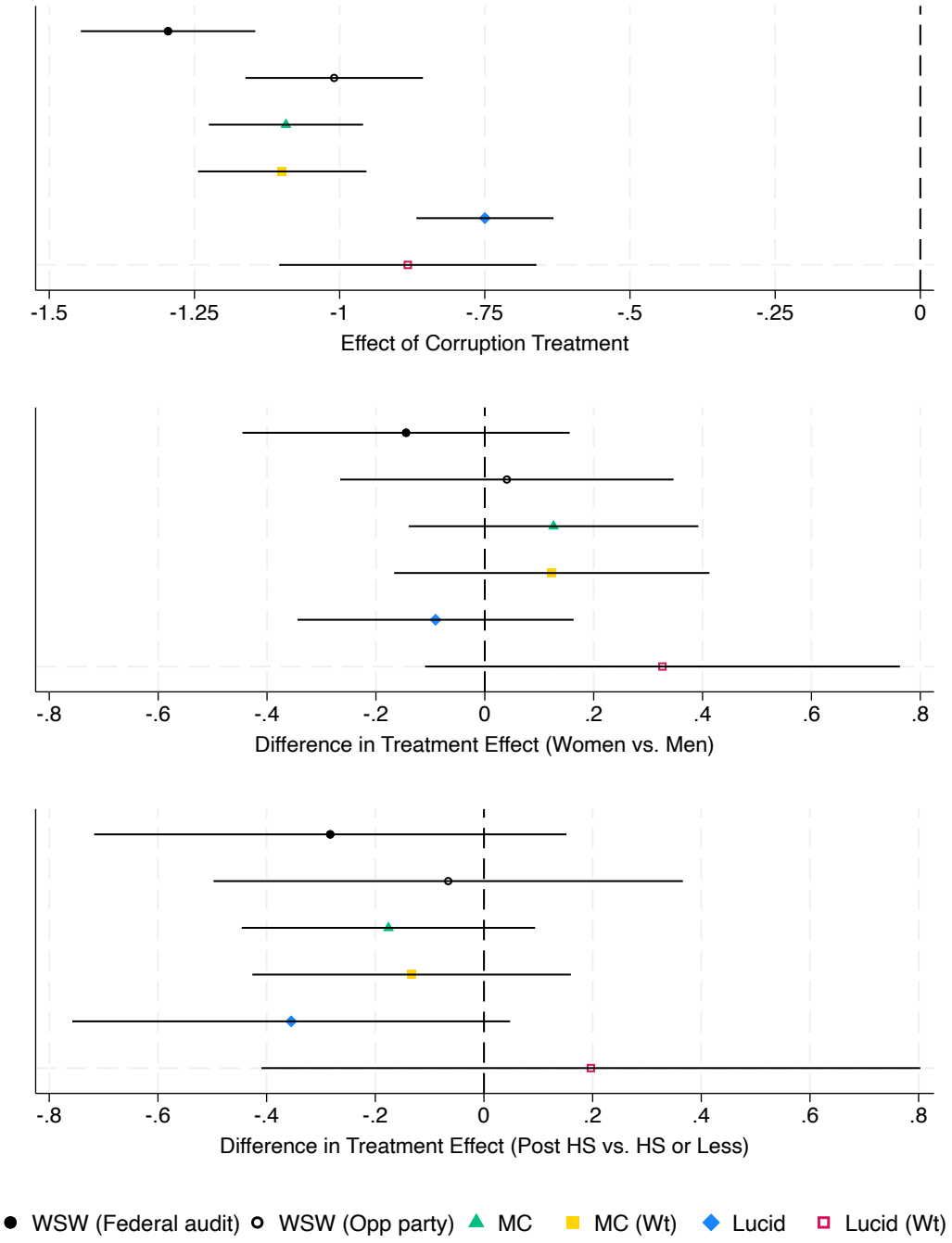

Figure 7. Brazil Experiment across Samples. The top panel presents average treatment effect estimates for the WSW study and our MC and Lucid studies, and for the latter two with and without weights. The middle and bottom panels present estimates of heterogeneous treatment effects by gender and college respectively. Horizontal bars present 95% confidence intervals.

However, the more surprising finding is the consistency within countries across platforms. The weighted and unweighted estimates also show minimal divergence, indicating that demographic adjustments do not dramatically shift the observed averages. The consistency of estimates, despite differences in sampling strategies, respondent pools, and weighting approaches, is striking. Other attitudinal measures may be more context dependent, however. As shown in the middle and bottom panels of Figure 8, we also see very similar estimates across platforms in estimates of social trust and confidence in financial institutions.

One outlier, in particular, stands out. In all three trust measures we observe statistically significant discrepancies across platforms in Brazil, where, the Lucid estimates are significantly greater than the MC estimates. This pattern may reflect differences in sample recruitment, respondent demographics, or platform effects that differentially capture political attitudes. The United States shows a similar, though less pronounced and directionally reversed pattern for trust in government and social trust, with MC respondents reporting slightly higher average trust than their Lucid counterparts; however, there estimates of confidence in financial institutions are similar across platforms. In the Philippines, average estimated confidence in financial institutions was modestly, but significantly higher in the Lucid samples than in the MC samples. Nevertheless, the stability of trust in government estimates across platforms and weighting strategies accords with work by scholars such as Mishler and Rose (2001), who have emphasized that trust in government is shaped by enduring institutional and cultural factors.

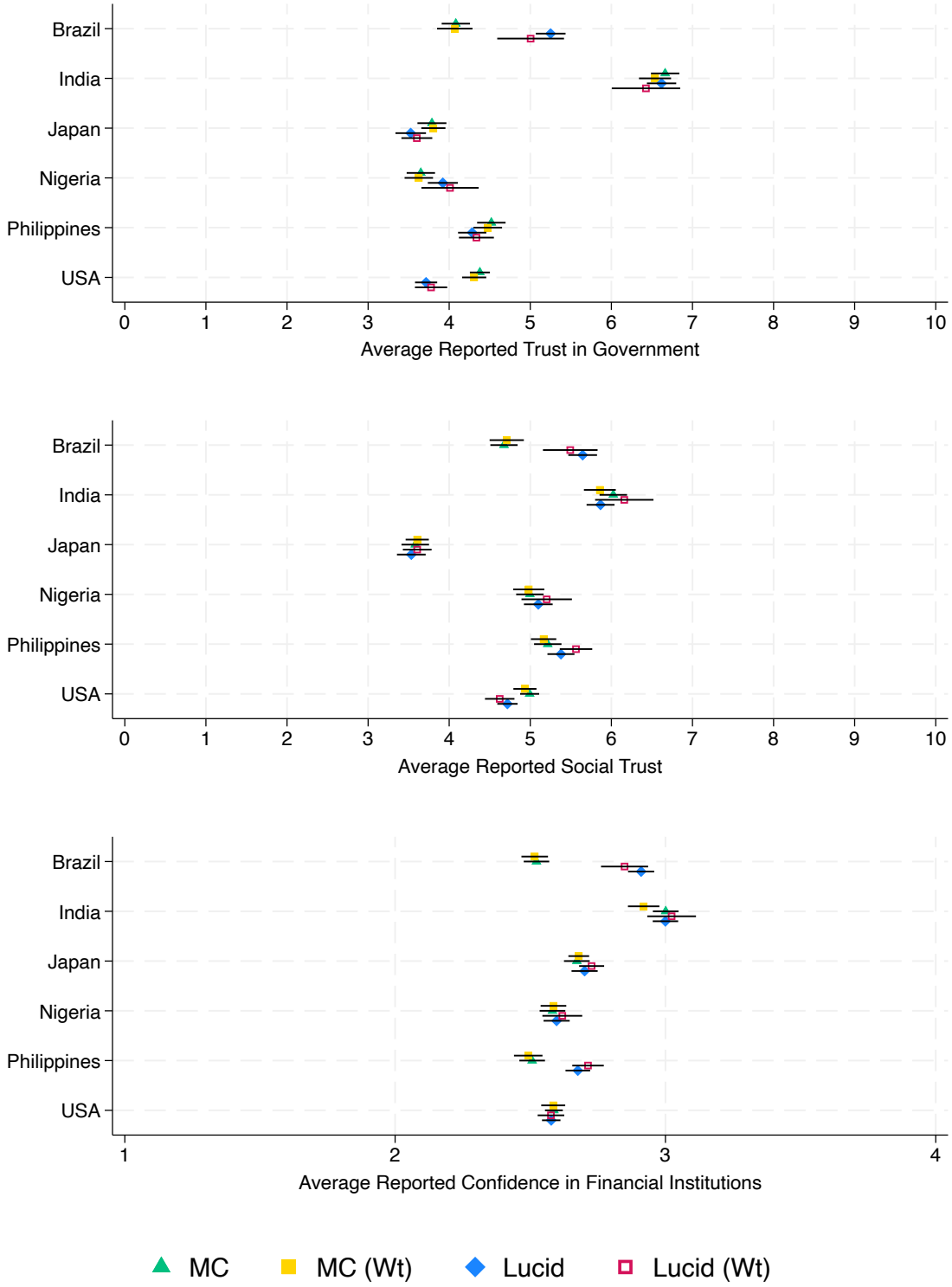

Figure 8. Trust Average reported trust in government, social trust, and confidence in financial institutions across countries and platforms. Results are shown with and without weighting to reflect known population characteristics. Error bars represent 95% confidence intervals.

Discussion

Several lessons emerge from the cross-national surveys. First, both platforms exhibit considerable demographic skews, particularly in terms of education and gender. For example, both platforms consistently overrepresent college-educated respondents, particularly in countries with lower levels of internet penetration. These skews may reflect the digital divide, as more educated individuals tend to have better internet access and familiarity with survey-taking platforms (although this explanation does not fit with the over-representation of college-educated respondents in Japan), a challenge noted in earlier work by Mercer et al. (2018) and Berinsky, Huber & Lenz. (2012), who note the risks of relying on opt-in panels without careful weighting. This bias highlights the persistent challenge of achieving representative samples in cross-national research, even when post-stratification weights are applied (Boas et al. 2020).

Second, and despite the demographic differences, treatment effects show general consistency across platforms within each country, suggesting that while baseline attitudes may vary, the relative impact of experimental interventions is less sensitive to platform choice. This finding offers a cross-national extension of past work (e.g. Berinsky, Huber & Lenz (2012)), finding that while demographic biases can distort descriptive statistics, the generalizability of experimental treatment effects is robust to sample composition. While estimates of the corruption treatment effect were consistently negative, we do find evidence of significant differences in its magnitude across countries. However, estimates of these cross-national differences varied across platforms. Finally, while our power is limited, we find broadly consistent evidence concerning heterogeneous treatment effects across countries and platforms. We find little evidence that women penalized corruption more than men, and considerable evidence of a stronger treatment effect among college-educated respondents in many contexts. The broad consistency of experimental findings across platforms may reflect the psychological consistency of treatment effects, which may depend more on the

salience of the experimental prompt than on the precise demographic composition of the sample (Coppock, Leeper, and Mullinix 2018). However, we caution that the considerable cross-platform similarities we observe may not hold in other experimental contexts. The effects of the specific treatment here, an allegation of corruption, may not vary significantly across groups that are under or over-represented in convenience samples. Other experimental treatments may be more strongly moderated by unobserved confounders. In such cases, we recommend researchers conduct sensitivity analyses for experimental treatment effect generalizability (Huang 2024).

Finally, baseline attitudes like trust in government were also generally similar across platforms within each country in our sample. While we did observe significant differences in Brazil and the U.S., they were small in magnitude. While caution must be used when using non-probability panels for population estimates of general political attitudes, the broad consistency of estimates across platforms was striking.

Conclusion

This study represents one of the most ambitious empirical efforts to date to compare convenience and quota-based sampling platforms across multiple national contexts. By fielding parallel surveys with an embedded experiment in six countries—Brazil, India, Nigeria, the Philippines, Japan, and the United States—our design allows for the first systematic comparison of platform effects on representatives, respondent attentiveness, and treatment effect consistency in such an expansive cross-national setting. While prior research has explored these dynamics within the United States or between one or two countries (Bassan-Nygate et al 2024; Boas et al 2020; Reynolds & Greenacre 2015; Moresche et al. 2018), no study to our knowledge has deployed this type of multi-platform, multi-country design with harmonized instruments and randomized treatments.

The structure of the study points to several extensions. Beyond providing the first systematic cross-national comparison of two widely used online survey platforms, our design offers a blueprint for evaluating the portability of causal inferences in political behavior experiments across diverse contexts. The approach can be adapted to other domains—such as attitudes toward international cooperation, misinformation, or climate change—where researchers are interested in whether experimental findings travel across institutional and cultural settings.

The findings also highlight a set of practical considerations for researchers. Platform choice is not just a question of cost, but of how different recruitment and weighting strategies shape the descriptive and experimental properties of samples. Future work might extend this logic by testing additional providers, benchmarking against probability samples where feasible, or tracking whether platform effects change over time as online participation and digital infrastructure evolve.

Finally, our results raise broader methodological and ethical questions about the role of convenience samples in comparative research. If experimental effects replicate across platforms and countries, as our study and others suggest, then low-cost online surveys offer a compelling tool for comparative research. When descriptive precision or subgroup balance matters most, researchers may need to turn to more resource-intensive sampling methods. This tension is unlikely to disappear and instead highlights the need for transparent reporting, thoughtful engagement with the limits of online samples, and continued comparative evaluation.

References

Alatas, V., Cameron, L., Chaudhuri, A., Erkal, N, and Gangadharan, L. 2009. "Gender Culture and Corruption: Insights from an Experimental Analysis," *Southern Economic Journal* 75(3): 663-680.

Bassan-Nygate, L., Renshon, J., Weeks, J.L., and Weiss, C.M. 2004. "The Generalizability of IR Experiments beyond the United States." *American Political Science Review*, 1–16. doi:10.1017/S0003055424001199.

Berinsky, A.J., Huber, G. & Lenz, G. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351-368.

Berinsky, A.J., Maroglis, M.F. & Sances, M.W. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58(3): 739-753.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., & Weimer, D. L. 2003. "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples." *Political Analysis*, *11*(1): 1-22.

Bethlehem, J. 2010. "Selection Bias in Web Surveys." *International Statistical Review* 78(2): 161–188.

Boas, T. C., Christenson, D. P., & Glick, D. M. 2020. "Recruiting Large Online Samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics." *Political Science Research and Methods 8*(2): 232-250.

The Business Research Company. 2025. "Public Opinion and Election Polling Market Report." Available at https://www.thebusinessresearchcompany.com/report/public-opinion-and-election-polling-global-market-report

Cinelli, C. & Hazlett, C. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82: 39-67.

Coppock, A. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7(3): 631-628.

Coppock, A., Leeper, T.J., and Mullinix, K. J. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115(49): 12441-12446.

Coppock, A. & McClellan, O. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research and Politics* 6(1): 2053168018822174.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., and Wenz, A. 2020. "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research." *Journal ofSurvey STatisticvs and Methodology* 8(1): 4-36.

Druckman, J. N., and C. D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base'." In *Cambridge Handbook of Experimental Political Science,* edited by J. N. Druckman, New York: Cambridge University Press.

Esarey, J., & Chirillo, G. 2013. "Fairer Sex" or Purity Myth? Corruption, Gender, and Institutional Context. *Politics & Gender 9*(4): 361-389.

Gouda, M and Park, SM. 2016. "Religious Loyalty and Acceptance of Corruption." *Jahrbücher für Nationalökonomie und Statistik* 235(2): 184-206.

Huang, M. 2024. "Sensitivity Analysis for the Generalization of Experimental Results." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187: 900-918.

Huff, C., & Tingley, D. 2015. "'Who are these people?' Evaluating the demographic characteristics and political preferences of MTurk survey respondents." *Research & Politics*, 2(3). https://doi.org/10.1177/2053168015604648

Hillygus, D. S., & Guay, B. 2016. "Polling in the United States." *Seminar Magazine* 684: 51-55.

Kam, C. D., Wilking, J. R., & Zechmeister, E. J. 2007. "Beyond the 'narrow data base': Another Convenience Sample for Experimental Research." *Political Behavior* 29: 415-440.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K. & Gimenez, A. 2016. "Evaluating Online Nonprobability Samples." *Pew Research Center* 61: 1-60.

Krupnikov, Y.H., Nam, H., & Style, H. 2021. "Convenience Samples in Political Science Experiments." In *Advances in Experimental Political Science*, Chapter 9. Cambridge: Cambridge University Press.

Lavena, C. F. 2013. "What Determines Permissiveness Toward Corruption? A Study of Attitudes in Latin America." *Public Integrity* 15(4): 345-366.

Marozzi, M. (2015). "Measuring Trust in European Public Institutions." *Social Indicators Research 123*: 879-895.

Melgar, N., Rossi, M., and Smith, T. W. 2010. "The Perception of Corruption." *International Journal of Public Opinion Research* 22(1): 120-131.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Pew Research Center. https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/

Mishler, W., & Rose, R. 2001. "What are the Origins of PoliticalTtrust? Testing Institutional and Cultural Theories in Post-Communist Societies." *Comparative Political Studies 34*(1): 30-62.

Moniz, P., Ramirez-Perez, R., Hartman, E., and Kessee, S. 2024. "Generalizing toward Nonrespondents: Effect Estimates in Survey Experiments are Broadly Similar for Eager and Reluctant Participants." *Political Analysis* 32(4): 507-520.

Moresche, T., Juliano D., and Melo, M.A. 2018. "Corruption Perceptions and Political Behavior: Evidence from Brazil." *Brazilian Political Science Review* 12 (1): 35-58.

Mullinix, K., Leeper, T.J., Druckman, J.N., and Freese, J. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(2): 109-138.

Peyton, K., Huber, G.A., and Coppock, A. 2022. "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic." *Journal of Experimental Political Science* 9(3): 379-394.

Reynolds, G, and Greenacre, J. 2015. "The Role of Social Media in Public Opinion Formation." *Social Science Computer Review* 33(1): 56-70.

Truex, R. 2011. "Corruption, Attitudes, and Education: Survey Evidence from Nepal." *World Development* 39(7): 1133-1142.

Weitz-Shapiro, R., & Winters, M. S. 2017. "Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil." *Journal of Politics 79*(1): 60-74.

Winters, M.S. & Weitz-Shapiro, R. 2020. "Information Credibility and Responses to Corruption: A Replication and Extension in Argentina." *Political Science Research and Methods* 8(1): 169-177.

Appendix A

First data source: All data analyzed in the paper come from original surveys fielded via Morning Consult and Cint/Lucid.

Data Collection Strategy: The first set of data was collected by Morning Consult. The second set of data was collected via Qualtrics from survey respondents recruited by Cint/Lucid.

Research Sponsor and Conductor: This research was supported by gifts from the Reynolds Foundation and Human Rights Foundation.

Measurement Tools/Instruments: The complete wording of all questions analyzed in the manuscript are provided in the Supporting Information.

Population Under Study: All survey samples were general adult populations.

Methods Used to Generate and Recruit the Sample: Morning Consult recruits respondents from multiple online panels via quota-based sampling; after data collecting, Morning Consult applies post-stratification weights to match national census parameters. Lucid/Cint recruits survey respondents from a global marketplace of online panels. Morning Consult also employs a range of data quality assurance strategies, including digital fingerprinting, to guard against non-eligible respondents, double respondents, and inattentive respondents. Lucid/Cint also employs fraud detection strategies, but it does not use quota-based sampling. Lucid/Cint does not provide survey weights.

Method and Mode of Data Collection: All survey data was collected via the internet.

Dates of Data Collection: The date for each survey across countries and platforms is provided in Table 1.

Sample Sizes: The sample size for each survey across countries and platforms is provided in Table 1.

Whether and How the Data were Weighted: Morning Consult constructed survey weights based on a range of demographic characteristics that varied across countries (for a complete list across

countries, see SI). For our Lucid samples, we created survey weights on age, gender, race, and ethnicity using the *ipfweight* package in STATA 18.

How the Data Were Processed and Procedures to Ensure Data Quality: Morning Consult employs two screener questions. The first (with varying numbers) was of the form: "A boy had two marbles and lost one. How many marbles does the boy have now?" The second asked "How often do you…" followed by a series of choices that varied somewhat across countries. Common options included "Use Instagram" and "Use YouTube". Respondents could then choose between "several times a day", "about once a day", "a few times a week", "about once a week", "at least once a month or less often" and "I do not have an account or do not use." One or two of the options, depending on country, referenced a media source that does not exist (e.g. "Use Appleton Post-Dispatch"). Respondents who did not choose the last response option ("I do not have an account or do not use") failed the screener. Survey-takers who failed either or both screeners were then exited from the survey. We included the same screeners and the same exit procedures with our Lucid sample.

Measurement and Model Specifications: All analyses presented in the text are simple differences in means. For the analyses of heterogenous treatment effects, the analyses report the difference in treatment effect (i.e. the difference in mean mayoral support across the treatment and control group) across the target groups (e.g. men vs. women, college-educated vs. non-college educated).

Limitations of the Design and Data Collection: The purpose of this study is to compare estimates from high quality, quota-based survey samples to convenience samples across a diverse range of countries and contexts. With the exception of Brazil, where we compare our average and heterogeneous treatment effect estimates with those reported in Weitz-Shapiro and Winters' 2017 study, we are unable to compare our estimates to comparable estimates from benchmark probability-based surveys.