Comparing In-Person, Sona, and Mechanical Turk Measurements of Three Prejudice-Relevant Constructs

Bradlee W. Gamblin, Matthew P. Winslow, Benjamin Lindsay, Andrew W. Newsom & Andre Kehn

Current Psychology
A Journal for Diverse Perspectives on Diverse Psychological Issues

ISSN 1046-1310
Volume 36
Number 2

Curr Psychol (2017) 36:217-224
DOI 10.1007/s12144-015-9403-1

Springer

Curr Psychol (2017) 36:217–224
DOI 10.1007/s12144-015-9403-1

Comparing In-Person, Sona, and Mechanical Turk Measurements of Three Prejudice-Relevant Constructs

Bradlee W. Gamblin[1] · **Matthew P. Winslow**[2] · **Benjamin Lindsay**[2] · **Andrew W. Newsom**[3] · **Andre Kehn**[1]

Published online: 7 January 2016

Abstract Electronic data collection and participant pool management tools give researchers new ways to conduct research. The current study investigated the equivalency of in-person and online administrations of the Right-Wing Authoritarianism, Social Dominance Orientation, and Modern Racism scales across three modalities (administration in person, online through Sona Systems, and online through Mechanical Turk). Results indicate that in-person administration was largely equivalent to the randomly assigned online sample (Sona Systems) in terms of their intercorrelations, mean scores, variability, and reliability. However, the Sona sample consistently responded with strong attitudes for all measures, suggesting that social desirability may be decreased when completing these measures online. In addition, the Mechanical Turk sample differed in many ways from both in-person and Sona administration; although this nonequivalence is at least partially explained by sample demographic differences, other considerations may have exacerbated nonequivalence, including prior exposure to the measures and a desire to respond correctly.

Keywords Sona · Mechanical Turk · MRS · SDO · RWA

✉ Bradlee W. Gamblin
bradlee.gamblin@ndus.edu

[1] Psychology Department, University of North Dakota, Grand Forks, ND, USA

[2] Psychology Department, Eastern Kentucky University, Richmond, KY, USA

[3] Marquette University, Milwaukee, WI, USA

Computer technology has changed the way psychological researchers recruit and track participants. Companies like Sona Systems Limited (Sona; www.sona-systems.com) and Amazon's Mechanical Turk (MTurk; www.MTurk.com) provide tools that allows researchers to post their studies online and allows participants to search for and complete these studies on their own time. Web-based participant pools provide many advantages over traditional paper-and-pencil bulletin board systems, including faster and cheaper data collection and a larger and more diverse pool of participants (Paolacci and Chandler 2014). These advantages have quickly been incorporated into the repertoire of psychological researchers. Indeed, at the time of this writing, approximately 19 % of all studies published in *Current Psychology* since March 2014 included some form of online research methodology. Overall, the replacement of in-person methods with online methods is supported by research assessing the equivalency of the two methodologies (Buhrmester et al. 2011). However, past research has not established the equivalency of studying *prejudice* online compared to its in-person counterpart. Thus, the purpose of the current study was to extend the literature on the equivalency of online and in-person data collection methodologies by establishing the online equivalency of three constructs relevant to the prejudice domain.

There is a considerable amount of research on the validity of electronic data collection and online participant pools. For example, Gosling et al. (2004) found that online participants had similar individual difference and motivational characteristics when compared to data collected using traditional methods. Overall, these studies have concluded that online participant pools (particularly MTurk) is a valid alternative to in-person data collection. Buhrmester et al. (2011), for instance, found that data collected through MTurk was equal to in-person data in terms of quality, interrater reliability, and test-retest reliability; Paolacci et al. (2010) also concluded that

MTurk and in-person samples were equivalent in their representativeness of the overall population. Past research also indicates that MTurk participants are attentive to experimenter instructions and are unlikely to complete the same study more than once (Berinsky et al. 2012).

Although the majority of the literature suggests that online research methodologies are as valid and reliable as in-person methods, there is reason to caution these claims of equivalency. For example, past research suggests that online participants are capable of participating multiple times in the same experiment (Berinsky et al. 2012), particularly when participation involves incentives (Bowen et al. 2008). Online participants are also more likely to participate in related experiments (Chandler et al. 2014). Both of these issues have established negative effects on internal validity (Birnbaum 2004; Brock and Becker 1966). Furthermore, although some researchers have concluded that online and in-person participant pools are equivalent (e.g., Paolacci et al. 2010), others have found evidence that online respondents differ from community samples (Berinsky et al. 2012) and college samples (Berinsky et al. 2012; Goodman et al. 2013); there is also evidence that online samples from different sources (e.g., MTurk vs. another website; Buhrmester et al. 2011). Goodman et al. (2013) also found that online participating receiving compensation were much more likely to look up the correct answers to objective questions in an experiment compared to those who were not compensated. On the other hand, some research suggests that online methodologies may also have benefits to internal validity. For example, Richman et al. (1999) found that online participation reduced social desirability biases compared to in-person participation.

The literature comparing specific scales and studies across modality also suggests that in-person and online data collection are equivalent. For example, online and in-person administration has been found equivalent for measures of depression, personality (Srivastava et al. 2003), political views (Berinsky et al. 2012), framing effects (Paolacci et al. 2010), self-esteem (Robins et al. 2002), and cognitive reflection (Goodman et al. 2013). Experiments have also been replicated using online methods; Berinsky et al. (2012), for instance, replicated Tversky and Kahneman's (1981) Asian disease problem through MTurk and found nearly identical results as those of the original study. Thus, across many different psychological domains, evidence suggests that the modality used to administer the study (in-person vs. online) does not affect the obtained data.

However, past research has not investigated the equivalency of online and in-person measurements of explicit prejudice measures. This is an important area of investigation for two reasons. First, online equivalency for prejudice measures cannot be assumed based on the fact that other constructs have been found equivalent when administered online (Buchanan 2002). Most of the constructs found equivalent between online and in-person administrations have not been investigated in relation to prejudice measures; of those that have been, none have established strong correlations (Ekehammar et al. 2004; Pratto et al. 1994; Schlachter and Duckitt 2002; Sibley and Duckitt 2008). Therefore, it is premature to assume that equivalency found for these previous constructs will transfer to measures of prejudice. Second, there is evidence to suggest that online administrations of prejudice measures may lead to different outcomes than in-person administrations because of their potential to reduce or eliminate experimenter effects, such as social desirability bias (Roberts 2007). Past research indicates that social desirability distortion is common and can lead to a variety of issues in interpreting responses (Tourangeau and Yan 2007). Furthermore, social desirability research asserts that this distortion is less severe in computer-based administrations (Evans et al. 2003), especially when respondents are alone and can backtrack (Richman et al. 1999) as they are when completing online research. Because prejudice measures are known to be influenced by social desirability, particularly when the respondent holds high levels of prejudice (Batson et al. 1978; Krysan 1998; Roese and Jamieson 1993), there may be important differences between prejudice measures and other constructs in how they translate from in-person to online administration.

Questions remain regarding the validity of data collected using online participant pools, both in general (Buhrmester et al. 2011; Goodman et al. 2013) and in terms of prejudice measures in particular (specifically regarding the impact of social desirability; Tourangeau and Yan 2007). Toward resolving these questions, the current study compared the equivalency of data collected for three prejudice measures (modern racism, right-wing authoritarianism, and social dominance orientation) across three different modalities: using traditional in-person methods, using online participant pools geared toward college students (i.e., Sona), and using online participant pools geared toward a general audience (i.e., MTurk). Although some past research suggests that prejudice measures may operate differently than other constructs (e.g., Batson et al. 1978; Ekehammar et al. 2004) and there is some evidence for non-equivalency in general (e.g., Chandler et al. 2014; Goodman et al. 2013), the majority of the literature points to equivalency across modality for psychological constructs (e.g., Buhrmester et al. 2011; Gosling et al. 2004). Therefore, we hypothesized that the three prejudice measures under investigation would be equivalent across modality.

Method

Participants

The In-Person and Sona samples consisted of 301 White undergraduate students (66.10 % female; M_{age} = 20.75,

SD = 4.79). Participants were randomly assigned to condition ($N_{\text{In-Person}}$ = 195; N_{Sona} = 136). Thirty participants in the Sona condition did not complete the optional online portion of the study and were thus excluded from our analyses. For the MTurk sample, participants were 240 White individuals from around the United States (52.5 % male; M_{age} = 36.06, *SD* = 12.86) collected online through the MTurk website and paid $0.40 for their participation (average completion time = 15.94 minutes), a pay rate consistent with past studies conducted through MTurk (e.g., Goodman et al. 2013; Paolacci et al. 2010). The study was advertised on MTurk as restricted to only White participants, and responses to a demographics questionnaire included at the beginning of the survey (hosted on Qualtrics) were used to exclude non-White participants who attempted to complete the survey. The MTurk sample was restricted to White participants because it is common in the prejudice literature to restrict participation to only White participants (Oliver and Wong 2003), and past research using these constructs has frequently used only White participants (e.g., Chambers et al. 2013; Ho et al. 2012; Trawalter et al. 2012).

Materials

Right-Wing Authoritarianism The Right-Wing Authoritarianism (RWA; Altemeyer 1998) scale consists of 20 items measuring three dimensions: submissiveness to authority figures, conventionalism, and a propensity to engage in aggression sanctioned by authority figures. Example items from the scale include, "There are many radical, immoral people in our country today who are trying to ruin it for their own godless purposes whom the authority should put out of action, " and, "The 'old-fashioned ways' and 'old-fashioned values' still show the best way to live." The RWA scale ranges from −4 (*very strongly disagree*) to +4 (*very strongly agree*), with higher scores indicating stronger endorsement of authoritarian ideas (see Table 1 for descriptive statistics and αs for all scales).

Social Dominance Orientation The Social Dominance Orientation (SDO; Pratto et al. 1994) scale consists of 16 items measuring preference for social inequality. Example items from the scale include, "Inferior groups should stay in their place," and, "To get ahead in life, it is sometimes necessary to step on other groups." The SDO scale ranges from 1 (*very negative*) to 7 (*very positive*), with higher scores indicating stronger endorsement of social dominance.

Modern Racism The Modern Racism Scale (MRS; McConahay 1986) consists of seven items measuring beliefs about race-relations in the United States. Example items from the scale include, "Blacks should not push themselves where they are not wanted," and, "Discrimination against Blacks is no longer a problem in the United States." The MRS ranges from −2 (*disagree strongly*) to +2 (*agree strongly*), with higher scores indicating stronger endorsement of prejudice toward African Americans.

Table 1 Scale properties for the in-person, Sona, and MTurk samples

Scale	Modality	Mean	*SD*	*N*	α
MRS	In-person	−.82	.71	195	.77
	Sona	−.54	.68	106	.68
	MTurk	−.91	.97	240	.89
SDO	In-person	2.42	.83	195	.86
	Sona	2.67	1.03	106	.92
	MTurk	2.27	1.12	240	.93
RWA	In-person	−.78	1.50	195	.93
	Sona	−.53	1.56	106	.93
	MTurk	−1.90	1.78	240	.96

Procedure

Participants in the In-Person and Sona conditions attended the initial session of data collection in person to establish random assignment to condition. Participants randomly assigned to the In-Person condition completed the RWA, SDO, and MRS scales in person in randomized order. Participants in the Sona condition were given instructions on how to complete the study through the university's Sona website. Participants in the MTurk sample signed up for the study through MTurk and were then given a link to complete the study through Qualtrics. Participants in both online conditions completed the three scales in randomized order.

Results

Sample Differences

Prior to conducting our main analyses, demographic differences across modality were examined in order to investigate how the college-based In-Person and Sona samples compared to the nationwide MTurk sample. Based on past research of MTurk sample demographic makeup (e.g., Berinsky et al. 2012), we expected to find that the samples would differ in both age and gender. Because participants in the In-Person and Sona conditions were randomly sampled from the same participant pool, analyses were conducted by collapsing across these two conditions and comparing these participants to the MTurk sample. Results indicated that there was a significant gender difference across modality, $\chi^2 = 19.41$, $p < .01$; there was a higher percentage of women in the In-Person/Sona

sample (66.3 %) than in the MTurk sample (52.5 %). Significant age differences were also found across modality, with participants in the MTurk sample being older (M_{age} = 36.06) and more varied in age (SD_{age} = 12.86) compared to participants in the In-Person/Sona sample (M_{age} = 20.75, SD_{age} = 4.79; Levene's F = 234.52, p < .01; t[291.96] = 17.50, p < .01). Because we were primarily interested in comparing convenience samples from the different modalities, the significant demographic differences were left unaccounted for in our analyses. However, differences in sample characteristics are important to note as both gender (Whitley 1999) and age differences (von Hippel et al. 2000) may affect interpretations of data obtained for these measures.

In order to investigate the potential role of gender and age in our data, preliminary analyses were also conducted using age and gender as predictors of the three prejudice measures. As expected (Whitley 1999), overall gender differences were found on two measures (t_{MRS}[464.04] = 2.40, t_{SDO}[411.41] = 4.55; both ps < .02; t_{RWA}[538] = −1.40, *ns*), and age was a significant overall predictor of two measures (β_{SDO} = −.12, β_{RWA} = −.17; both ps < .01; β_{MRS} = −.05, *ns*). Thus, sample characteristics are important for these measures not only because convenience samples will vary across modality but also because the demographics significantly impact responses to the measures.

Primary Analyses

To investigate equivalency across the three modalities, we initially calculated correlations amongst the three scales and compared the correlations across modality (see Table 2 for a summary of equivalency findings for all analyses). In general, the highest correlations were found in the MTurk sample ($r_{MRS\text{-}SDO}$ = .60, $r_{MRS\text{-}RWA}$ = .64, $r_{SDO\text{-}RWA}$ = .46), replicating past research on other constructs collected through MTurk (Buhrmester et al. 2011). The Sona sample ($r_{MRS\text{-}SDO}$ = .50, $r_{MRS\text{-}RWA}$ = .45, $r_{SDO\text{-}RWA}$ = .44) and the In-Person sample ($r_{MRS\text{-}SDO}$ = .60, $r_{MRS\text{-}RWA}$ = .41, $r_{SDO\text{-}RWA}$ = .45) displayed slightly weaker correlations in the same direction as the MTurk sample. We also compared these obtained correlations to those published in past research using these measures. For example, Pratto et al. (1994) reported a correlation of .53 between SDO and MRS; Bäckström and Björklund (2007) reported a correlation of .50 between RWA and MRS; and, in their meta-analysis of SDO and RWA, Wilson and Sibley (2013) report correlations between .12 and .48. All of these published correlations are similar to those reported here. Finally, we examined the intercorrelations across modalities using z-tests. Of the three correlations between MRS, RWA, and SDO, none were significantly different between the In-Person and Sona samples (all z-tests <1.2). However, the correlation between MRS and RWA was significantly higher in the MTurk sample than in both the Sona (z = 2.32) and the In-Person samples (z = 3.32).

Table 2 Summary of equivalency findings

Comparison	Analysis	MRS	SDO	RWA
In-person – Sona	Intercorrelations	E	E	E
	Mean differences	N	E	E
	Variability	E	N	E
	Reliability	E	E	E
In-person – MTurk	Intercorrelations	N[a]	E	N[a]
	Mean differences	E	E	N
	Variability	N	N	N
	Reliability	E	E	E
Sona – MTurk	Intercorrelations	N[a]	E	N[a]
	Mean differences	N	N	N
	Variability	N	N	N
	Reliability	E	E	E

E equivalent, *N* nonequivalent, p < .05

[a] Nonequivalence for the MRS-RWA correlation only

Mean scores and variances on the three scales were also compared across modality to investigate equivalency (see Table 1 for *M*s and *SD*s). Levene's test suggested heterogeneity of variance for all three scales (F_{MRS} = 17.73, F_{SDO} = 9.45, F_{RWA} = 9.13; all ps < .01); the MTurk sample consistently displayed the largest amount of variation across all scales. A second Levene's test was also conducted, excluding the MTurk sample, in order to test for homogeneity of variance across the two conditions using random assignment. This test revealed homogeneity of variance for MRS (F = .30, *ns*) and RWA (F = .02, *ns*), but heterogeneity for SDO (F = 6.51, p < .02). Next, one-way ANOVAs were conducted on the mean scores. Analyses revealed significant effects of modality for MRS, F(2, 538) = 7.40; for SDO, F(2, 538) = 5.75; and for RWA, F(2, 538) = 36.88 (all ps < .01). Importantly, the effect size of modality on both the MRS (η_p^2 = .03) and the SDO (η_p^2 = .02) was small, suggesting that any differences across modality were not very powerful. However, modality had a much stronger effect on RWA (η_p^2 = .12).

Post-hoc analyses were conducted on each scale to determine which modalities were driving our significant mean differences. Bonferroni-corrected analyses revealed that MRS scores were significantly higher in the Sona condition (M = −.54) compared to the MTurk (M = −.91) and In-Person conditions (M = −.82; both ps < .02), which were not significantly different from one another; that SDO scores were significantly higher in the Sona condition (M = 2.67) compared to the MTurk condition (M = 2.27; p < .01); and that RWA scores were significantly lower in the MTurk condition (M = −1.90) compared to the Sona (M = −.53) and

In-Person conditions ($M = -.78$; both $ps < .01$), which were not significantly different from one another.

Finally, scale reliabilities were compared across modality; α values for all scales were at acceptable levels (>.65) across modality (see Table 1). Thus, acceptable interitem reliability was achieved for all scales regardless of condition, although reliability tended to be higher in the MTurk condition.

Discussion

Online data collection presents researchers with attractive opportunities. Large data sets can be collected in a relatively short amount of time at a reduced cost. Online data collection can also reduce concerns with social desirability and experimenter effects. However, before seizing these opportunities, researchers must establish the equivalence of online measures and traditional in-person measures. The current research provides evidence both for and against equivalency of online and in-person administration of three commonly used measures of prejudice: the Modern Racism Scale (McConahay 1986), the Social Dominance Orientation Scale (Pratto et al. 1994), and the Right-Wing Authoritarianism Scale (Altemeyer 1998).

Using a variety of methods of comparison, and across two online modalities, online responses were in many ways comparable to those collected in person. In particular, the two randomly assigned conditions (in-person and Sona) yielded very few differences across modality. Two measures (SDO and RWA) had equivalent means, and two measures (MRS and RWA) had equivalent variances across the two modalities. In addition, all intercorrelations were equivalent between the in-person and Sona administrations, and reliability measures indicated similar levels of reliability across modality. The comparisons between these two modalities are perhaps the most important in establishing the equivalency of online and in-person administrations of prejudice measures in general. Because the In-Person and Sona conditions involved random assignment, these comparisons provide a true test of the effects of modality in isolation from the confounds discussed below that are present in the MTurk sample. Largely, the current study suggests that prejudice measures are unaffected by the method used to administer them. Thus, assuming that the characteristics of online samples are equivalent to those that would have been obtained in person, online responses obtained for these measures should also be equivalent to those they would have obtained in person.

However, the current study did identify some important differences between in-person and Sona administration. Most importantly, participants who completed the MRS through Sona reported significantly higher levels of modern racism than participants who completed the measure in person, and participants who completed the SDO through Sona exhibited more variability of responding than participants who completed the measure in person. Furthermore, although only MRS was significant, mean scores on all three scales were higher when administered through Sona than when they were administered in person. Although the effects of modality on MRS were small, our findings still suggest that administration method has some effect on participant responding. The literature on social desirability provides a potential explanation for these modality differences. Past research on the social desirability bias indicates that completing sensitive measures on a computer (Richman et al. 1999), particularly when completed over the internet away from the research lab (Evans et al. 2003), can reduce social desirability distortion and lead to more honest responses. Thus, participants completing the MRS through Sona (at home, over the internet) may feel less compelled to respond in a socially desirable manner, leading them to respond with higher (and more honest) levels of modern racism. Because the measures used in the current study are known to be influenced by social desirability bias (Batson et al. 1978; Duriez and Van Hiel 2002; Krysan 1998; Roese and Jamieson 1993), it seems plausible that differences on the MRS were due to reduced feelings of social desirability in the Sona participants. If this is the case, then researchers using this measure may actually prefer to gather data online to avoid the issues that can be caused by social desirability distortion (Tourangeau and Yan 2007). However, the current study did not include a measure of social desirability, so there may be an alternative explanation for the difference between the Sona and In-Person conditions. Future research investigating equivalency of prejudice measures (or other measures influenced by social desirability) should more thoroughly investigate the role of social desirability in influencing in-person and online responses.

In addition, many differences were identified between MTurk and the other two modalities. The MTurk sample displayed a stronger correlation between two of the scales (MRS-RWA), replicating past research using both in-person and MTurk samples (Crawford 2012). The MTurk sample also exhibited the highest variability on all three scales. Furthermore, the MTurk sample reported lower average scores on all scales, particularly in comparison to the Sona sample. The clearest explanation for these differences is that the convenience sample gathered through MTurk differed in important ways from the sample used for the In-Person and Sona conditions. Specifically, the MTurk sample was significantly older and included a higher proportion of men in comparison to the In-Person/Sona sample; in addition, gender and age effects were found on all three scales, replicating past research (Sidanius et al. 2000; Wilson and Sibley 2013). Therefore, the current study's findings of nonequivalency between MTurk and the other two conditions may be due largely to demographic differences between the samples.

By not randomly assigning participants to the MTurk condition, the current study introduces demographic confounds to

any interpretation of differences between MTurk and the other samples in our study. The effects of demographic differences between MTurk and other participant pools has been established previously (Paolacci et al. 2010); however, the current study extends these findings by providing evidence of demographic effects when collecting data using prejudice measures. Specifically, the current study suggests that participants on MTurk may exhibit different average attitudes on these three prejudice measures and, therefore, researchers should be cautious when interpreting, combining, or comparing results collected across MTurk and other modalities. For instance, when a researcher plans to combine prejudice data collected through MTurk with data collected using another modality, the researcher should consider all possible demographic confounds that may cause MTurk participants to respond uniquely and account for these potential confounds in their analyses. As an example, Crawford and Pilanski (2014) used a multi-sample design (in-person college students and an MTurk convenience sample) in which they obtained data for both RWA and SDO. The authors also chose to combine the two samples. However, the authors acknowledge potential issues with combining the samples and controlled for modality in their analyses. The current study supports this approach and suggests that it is a critical step when combining or comparing prejudice data collected through MTurk data to that of another modality.

Beyond the effects of demographics, past research provides alternative explanations as to why the current study found differences between MTurk and the other two modalities. For instance, past research indicates that MTurk participants are more likely to have prior exposure to methods and measures used in psychological studies, and that this prior exposure can influence participants' responses (Chandler et al. 2014). Further, Goodman et al. (2013) found that, compared to other modalities, MTurk participants were more likely to make efforts to appear factually correct (by looking up answers on the internet), even when the item asked participants to guess or give an opinion. Thus, our MTurk sample may have been exposed to the prejudice measures in past experiments they had completed and, having surmised that the experiment was investigating prejudice, attempted to answer "correctly" by reporting low levels of prejudice.

If MTurk responses are influenced by perceptions of correctness or experimenter intent (Goodman et al. 2013), then research conducted using MTurk may present a unique case by which social desirability is actually stronger online compared to in-person administration. Both data from the current study and previous research suggests that this may be the case. First, responses to our three measures were consistently lowest in the MTurk condition. This is particularly surprising when demographic differences are accounted for: although past research has found gender (Whitley 1999) and age differences (von Hippel et al. 2000) for these measures assessed in this study, these studies suggest that the MTurk sample (being the older, more male sample) should have expressed the highest levels of prejudice, not the lowest. Second, past research has found that MTurk samples score higher on measures of social desirability compared to college-based samples (Behrend et al. 2011). Behrend et al. (2011) suggest that the higher levels of social desirability found in MTurk samples may be due to a perception that giving an incorrect response will lead to decreased compensation for the participant, and Goodman et al. (2013) demonstrated that compensating MTurk participants led to higher rates of factually accurate responses. On the other hand, there is some evidence that older and more male populations are more likely to respond in a socially desirable manner (Watson et al. 1984). Thus, social desirability differences found between MTurk and other samples in the current study, and in past research, may instead be due to demographic characteristics. Future research should further investigate these relationships to better understand the role of social desirability bias within MTurk samples and account for it when comparing or combining datasets across modality.

Although the considerations discussed above are essential when designing studies and interpreting results obtained from online data collection, the results of the current study should not be interpreted as invalidating prejudice research conducted using online methodologies or comparisons made across modalities. Instead, the current study is intended to provide guidance and suggest some cautionary procedures future researchers may take when administering prejudice measures online (particularly through MTurk). It is integral that researchers investigate and control for any effects of administration method when combining datasets across modality. These effects may include social desirability distortion (Tourangeau and Yan 2007), demographic characteristics (Sidanius et al. 2000), prior exposure to the prejudice measures (Chandler et al. 2014), and participants' desire to give the correct answer (Goodman et al. 2013). Each of these effects may be accounted for when conducting online or in-person research. To address issues of social desirability bias, researchers may account for the effects of social desirability using items or scales created to measure an individual's level of social desirability (Nederhof 1985). Researchers may also try to reduce the effects of social desirability through techniques such as the bogus pipeline (Roese and Jamieson 1993; Tourangeau and Yan 2007), in which the participant is led to believe that any attempt at hiding their true attitudes will be detected by the researcher. Researchers may control for demographic characteristics when comparing datasets obtained online and in person, or they may set strict requirements for their online participant pool such that the online dataset matches the in-person dataset on any relevant demographic variables. Similarly, researchers can account for prior exposure to the measures by asking participants if they have seen any of the scales prior to their participation. Finally, Goodman et al. (2013) suggest that

including instructions asking participants to respond truthfully, not correctly, may help alleviate issues with participants giving socially desirable, but not necessarily honest, responses.

Finally, the current study suggests that online administration provides some important benefits to prejudice researchers. Participants who completed the MRS through Sona reported higher levels of modern racism than participants who completed the scale in person; since these conditions were randomly assigned, the differences across modality are likely due to experimenter effects such as social desirability (Evans et al. 2003). Furthermore, although the demographic differences between the MTurk sample and the college sample presented difficulties for the current study, they also indicate that MTurk provides a different (and in many ways superior; Buhrmester et al. 2011) participant pool that may be useful for investigating particular research questions about prejudice. Therefore, the evidence of nonequivalence presented in the current study should not be construed as evidence that in-person administrations of prejudice measures are necessarily *better* than online administrations.

Conclusion

It is integral to establish the validity of using online participant pools like Sona and MTurk because of their great utility for the researcher and the prevalence of their usage in today's field (Mason and Suri 2012). In many ways, the current supports the validity of data collected online for prejudice measures, suggesting that researchers can take advantage of this opportunity. However, the current study also highlights many critical considerations that future researchers should account for when using these measures of prejudice, particularly when combining or comparing data collected from MTurk with that collected from a typical, in-person, college-based sample. In summary, our results join others that support the view that data collected electronically are mostly equivalent to data collected in person. More specifically, we have demonstrated the equivalence of three commonly used measures of social attitudes (MRS, SDO, and RWA) when comparing in-person and electronic data collection using two participant pool management system (Sona Systems and Mechanical Turk). While not uniformly equivalent, these results provide researchers with evidence as to when and how the measures are equivalent across modality. In addition, our results provide insight as to how to design future studies to ensure equivalency, and the implications of our findings point out the potential utility of instances where the modalities were not equivalent.

Acknowledgments Thanks to Meredith Wells for reviewing a draft of the manuscript.

Compliance with Ethical Standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of Interest Bradlee W. Gamblin declares that he has no conflict of interest. Matthew P. Winslow declares that he has no conflict of interest. Benjamin Lindsay declares that he has no conflict of interest. Andrew W. Newsom declares that he has no conflict of interest. Andre Kehn declares that he has no conflict of interest.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

Altemeyer, B. (1998). The other "Authoritarian" personality. *Advances in Experimental Social Psychology*, *30*, 47–91.

Bäckström, M., & Björklund, F. (2007). Structural modeling of generalized prejudice: The role of social dominance, authoritarianism, and empathy. *Journal of Individual Differences*, *28*, 10–17.

Batson, C. D., Naifeh, S. J., & Pate, S. (1978). Social desirability, religious orientation, and prejudice. *Journal for the Scientific Study of Religion*, *17*, 31–41.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*, 800–813.

Berinsky, A. J., Huber, G. A., & Lenz, G. A. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368.

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, *55*, 803–832.

Bowen, A. M., Daniel, C. M., Williams, M. L., & Baird, G. L. (2008). Identifying multiple submissions in internet research: preserving data integrity. *AIDS Behavior*, *12*, 964–973.

Brock, T. C., & Becker, L. A. (1966). "Debriefing" and susceptibility to subsequent experimental manipulations. *Journal of Experimental Social Psychology*, *2*, 314–323.

Buchanan, T. (2002). Online assessment: desirable or dangerous? *Professional Psychology: Research and Practice*, *33*, 148–154.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Chambers, J. R., Schlenker, B. R., & Collisson, B. (2013). Ideology and prejudice: the role of value conflicts. *Psychological Science*, *24*, 140–149.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaivete among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112–130.

Crawford, J. T. (2012). The ideologically objectionable premise model: predicting biased political judgments on the left and right. *Journal of Experimental Social Psychology*, *48*, 138–151.

Crawford, J. T., & Pilanski, J. M. (2014). The differential effects of right-wing authoritarianism and social dominance orientation on political intolerance. *Political Psychology*, *35*, 557–576.

Duriez, B., & Van Hiel, A. (2002). The March of modern fascism: A comparison of social dominance orientation and authoritarianism. *Personality and Individual Differences*, *32*, 1199–1213.

Ekehammar, B., Akrami, N., Gylje, M., & Zakrisson, I. (2004). What matters most to prejudice: Big Five personality, social dominance orientation, or right-wing authoritarianism? *European Journal of Personality*, *18*, 463–482.

Evans, D., Garcia, D., Garcia, D., & Baron, R. (2003). In the privacy of their own homes: using the Internet to assess racial bias. *Personality and Social Psychology Bulletin*, *29*, 273–284.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224.

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104.

Ho, A. K., Sidanius, J., Pratto, F., Levin, S., Thomsen, L., Kteily, N., & Sheehy-Skeffington, J. (2012). Social dominance orientation: revisiting the structure and function of a variable predicting social and political attitudes. *Personality and Social Psychology Bulletin*, *38*, 583–606.

Krysan, M. (1998). Privacy and the expression of White racial attitudes. *Public Opinion Quarterly*, *62*, 506–544.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1–23.

McConahay, J. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In *Prejudice*, *discrimination*, *and racism* (pp. 91-125). San Diego: Academic Press.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: a review. *European Journal of Social Psychology*, *15*, 263–280.

Oliver, J. E., & Wong, J. (2003). Intergroup prejudice in multiethnic settings. *American Journal of Political Science*, *47*, 567–582.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184–188.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Pratto, F., Sidanius, J., Stallworth, L., & Malle, B. (1994). Social dominance orientation: a personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*, 741–763.

Richman, W., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*, 754–775.

Roberts, L. D. (2007). Equivalence of electronic and off-line measures. In R. A. Reynolds, R. Woods, & J. D. Baker (Eds.), *Handbook of research on electronic surveys and measurements* (pp. 97–103). Hershey: Idea Group Reference/IGI Global.

Robins, R. W., Trzesniewski, K. H., Tracy, J. L., Gosling, S. D., & Potter, J. (2002). Global self-esteem across the life span. *Psychology and Aging*, *17*, 423–434.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: a critical review and meta-analysis. *Psychological Bulletin*, *114*, 363–375.

Schlachter, A., & Duckitt, J. (2002). Psychopathology, authoritarian attitudes, and prejudice. *South African Journal of Psychology*, *32*, 1–8.

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: a meta-analysis and theoretical review. *Personality and Social Psychology Review*, *12*, 248–279.

Sidanius, J., Levin, S., Liu, J., & Pratto, F. (2000). Social dominance orientation, anti-egalitarianism and the political psychology of gender: an extension and cross-cultural replication. *European Journal of Social Psychology*, *30*, 41–67.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84,* 1041–1053.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883.

Trawalter, S., Hoffman, K. M., & Waytz, A. (2012). Racial bias in perceptions of others' pain. *PloS One*, *7*(11), 1–8.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

von Hippel, W., Silver, L. A., & Lynch, M. E. (2000). Stereotyping against your will: the role of inhibitory ability in stereotyping and prejudice among the elderly. *Personality and Social Psychology Bulletin*, *26*, 523–532.

Watson, P. J., Grisham, S. O., Trotter, M. V., & Biderman, M. D. (1984). Narcissism and empathy: validity evidence for the narcissistic personality inventory. *Journal of Personality Assessment*, *48*, 301–305.

Whitley, B. E. (1999). Right-wing authoritarianism, social dominance orientation, and prejudice. *Journal of Personality and Social Psychology*, *77*, 126–134.

Wilson, M. S., & Sibley, C. G. (2013). Social dominance orientation and right-wing authoritarianism: additive and interactive effects on political conservatism. *Political Psychology*, *34*, 277–284.