Survey Research Methods (2024)
Vol. 18, No. 1, pp. 59-78
https://doi.org/10.18148/srm/2024.v18i1.7948
European Survey Research Association

We have come a long way and we have a long way to go. A cross-survey comparison of data quality in 16 Arab countries in the Arab Barometer vs the World Values Survey

Saskia Glas[1] · Veronica Kostenko[2,3]
[1]Radboud University
[2]European University
[3]Tel-Aviv University

With the launch of the Arab Barometer (AB) project and the incorporation of Arab countries in the World Values Survey (WVS) in the 2000s, public opinion scholars have increasingly turned their attention to the Arab region. However, remarkably little is however known about the quality of these data. To our knowledge, Arab surveys have never been scrutinized in a systematic empirical cross-survey study. Therefore, this study compares sixteen surveys from the AB with sixteen from the WVS concerning four attitudes widely studied by substantive scholars: generalized and institutional trust and gender equality in education and in politics. We assess the comparability of their univariate distributions and their predictors in multivariate models. Our results show considerable diversity across and even within surveys in quality, indicating that blanket statements on Arab surveys' (lack of) quality are inappropriate. In a minority of tested cases (17%), the conclusions of scholars on what predicts trust or gender equality depend completely on the chosen data source. We also test whether often-heard reasons for Arab surveys' supposed lack of quality explain the diversity in survey quality. Our results show that neither sample differences nor enumerator fraud drives discrepancies, but there might be some influence of socially desirable answers

Keywords: data quality; survey comparison; comparative data analysis; gender equality; trust; Arab Barometer; World Values Survey

1 Introduction

Since the 2000s, the Arab region has been included into the scope of comparative social surveys, broadening the landscape for comparative social scholars (e.g. Inglehart and Norris, 2003; Jamal and Tessler, 2008). Roused by public debates on the region, comparative scholars have delved into questions on Arab people's democratic attitudes, trust, religiosity, and gender norms in particular (e.g., Rizzo, Helen, Abdel-Latif, Abdel-Hamid, and Meyer, Katherine, 2007; Kostenko, Kuzmuchev, and Ponarin, 2016; Tessler and Tout, 2017; Glas and Spierings, 2021). This fascinating endeavour shed light on many processes in the region and helped explain seemingly unexpected events, such as the advent of Arab Spring given the high support for democracy in many Arab countries.

With the inclusion of Arab countries, the diversity of countries covered rose dramatically, which meant scholars also took a gigantic leap towards better explaining attitudes on the global level. For instance, after Jamal's (2007) groundbreaking conclusion that generalized trust reinforces rather than reduces support for existing authoritarian regimes in the Arab states, a whole new subfield on the dark side of trust was born (Diop, Tessler, Wittrock, and Jardina, 2017; Spierings, 2019; Sika, 2020). Arab people's gender attitudes also sparked substantial interest. Inglehart and Norris' thesis that "an Islamic heritage is one of the most powerful barriers to the rising tide of gender equality" (Inglehart & Norris, 2003, pg. 49) was itself met with a wealth of nuances by specialists on the region (Ross, 2008; Alexander and Welzel, 2011; Glas, Spierings, and Scheepers, 2018). Fish (2011) showed that, in general, the values of Muslim publics showed little particularities, but

Corresponding author: Saskia Glas, Radboud University, Nijmegen, Netherlands (Email: s.glas@ru.nl)

he did note specificities in support for gender equality, tolerance of homosexuality, and religiosity.

However, remarkably little attention has been paid to the data themselves. As Benstead argues: "very few systematic efforts have been made to assess the surveys' comparability" (Benstead, 2018a, pg. 224). This might be because some seem to assume that Arab survey quality is stagnant (as noted by Gengler, Tessler, Lucas, and Forney, 2021; Benstead, 2018b), but, at the very least, whether there is indeed no diversity in the quality of Arab surveys should be empirically assessed rather than assumed. This paper takes up that challenge and takes the first steps in systematically studying the quality of Arab survey data.

Our aim is to shed light on the Arab survey data in such a way that our conclusions do not remain hidden in the survey methodology literature but actually help substantive Arab public opinion scholars further expand a field that is already filled with fascinating insights. We take four steps to ease the connection with the substantive literature.

First, we focus on two publicly available surveys that are often used by substantive scholars: the Arab Barometer (Jamal et al., 2014, 2019) and the World Values Survey (Inglehart et al., 2014; Haerpfer et al., 2020). Second, we focus on four attitudes that have been widely studied in this region particularly, and hotly debated: trust (both generalized and institutional), and gender equality (both in education and in politics) (e.g., Diop et al., 2017; Rizzo, Helen et al., 2007). Third, we do not only assess descriptive levels of trust and gender equality, but we also focus on what substantive scholars concentrate on, namely what predicts them. Fourth, this paper does not nitpick each and every technical abnormality in the AB and the WVS surveys that in the end is inconsequential to the conclusions of substantive scholars. Rather, we emphasize phenomena that truly alter main substantive conclusions and provide some tips and tricks for how substantive scholars can tackle these.

Our assessment entails a comparison between Arab WVS and AB surveys. Although such cross-survey comparisons are becoming more and more common in educational and demographic studies on Muslim-minority countries (e.g., Hayford and Morgan, 2008; Ortmanns and Schneider, 2016; Manning, Joyner, Hemez, and Cupka, 2019), survey comparison has not been applied yet to public opinion in Arab countries. It is regrettable that methodologists have not risen to this task, because substantive works themselves have pointed out discrepancies between the AB and the WVS (Glas, Spierings, Lubbers, & Scheepers, 2019, pg. 305). Moreover, survey comparison is an especially useful tool for Arab surveys, because quality tests that compare surveys and general populations are tricky, not only because censuses obviously lack attitudinal data, but also because we lack reliable censuses on what many Arab populations look like. What we do have is thirty—two Arab surveys-sixteen from WVS wave six and seven and sixteen from AB wave three and five—that probed four similar questions on trust and gender equality to similar populations[1], which creates exciting opportunities for comparisons.

Ultimately, comparing surveys allows us to lay bare where existing surveys' quality might be lacking. If two surveys targeting the same population return substantially different results, at least one of them must be flawed. If two surveys are comparable, however, this might signify either the quality of both surveys or that both are biased in the same way. Therefore, we focus particularly on where exactly surveys are lacking in quality, as we cannot provide direct evidence for the opposite.

This paper continues as follows. The next section introduces possible reasons for why we could expect to see diversity in the quality of Arab surveys. As we explain there, we focus in particular on whether sample differences, enumerator fraud, and socially desirable answers explain diversity in survey quality, because these capture some of the often presumed defects of Arab surveys that we can address empirically—directly or indirectly—in the current study (Gengler et al., 2021). The sections thereafter start by outlining technical methodological details and our findings on diversity between the surveys. Next, we test whether unrepresentative samples, enumerator fraud, and socially desirable answers might explain any discrepancies between the AB and the WVS, and we end with tips and tricks for substantive scholars. We conclude that there is far too much diversity in survey quality to warrant any generalizations on Arab surveys' caliber.

2 Possible reasons for diversity in Arab survey quality

Although the quality of Arab surveys has hardly been assessed (Benstead, 2018a; Benstead, 2018b; Tessler, Palmer, Farah, and Ibrahim, 2019 [1987]) and the AB and WVS have reached milestones in covering Arab countries that have received awards[2], sometimes it is simply assumed that all Arab surveys are unreliable because "non-Western, op-

[1] The populations are: Algeria in March-April 2013 and in January 2014; Egypt in 2013; Egypt in 2018; Iraq in 2013; Iraq in June 2018 and in December 2018 and January 2019; Jordan in December 2012 and January 2013 and in 2014; Jordan in 2018; Kuwait in 2014; Lebanon in 2013; Lebanon in 2018; Libya in 2014; Morocco in May-June 2011 and in April-June 2013; Palestine in December 2012 and February-March 2013; Tunisia in 2013: Tunisia in October-December 2018 and in April-May 2019; Yemen in November-December 2013 and February 2014. See Sect. 3 for further discussion.

[2] APSA has awarded its Lijphart/Przeworski/Verba Data Set Award to Inglehart for the WVS and Eurobarometers in 2001, and to Tessler and Jamal in 2010 for the Arab Barometer: https://www.apsanet.org/STAFF/MembershipWorkspace/Organized-Sections/Organized-Section-Awards/Organized-Section-Awards/Section-20.

pressed, passive" Arab people would be unable to answer survey questions (Gengler et al., 2021). This paper swaps such non-evidence driven assumptions with Orientalist undertones with a systematic empirical assessment emphasizing diversity. We argue that there is nothing inherently flawed about Arab countries that makes it impossible to conduct survey research in the region. In fact, one substantive working paper on trust globally that tests the quality of its data as a laudable robustness test, concludes that there is diversity in the Arab surveys' quality, but finds no fault with the AB surveys (Nunn, Qian, & Wen, 2018). More generally, we argue that there are real differences between Arab countries which are likely to create diversity in its surveys as well (e.g., Glas & Spierings, 2020; Price, 2015; Owen, 2013). What is needed is a context-dependent view that takes into account the particularities across the region (Clark & Cavatorta, 2018).

This section connects regional insights with the general survey methodology literature. However, we do not (cl)aim to detail every phenomenon that might impact survey quality, such as question order (Nugent, Masoud, & Jamal, 2018), interviewer traits (Corstange, 2014), perceptions of the national affiliations of surveys (Gengler, Le, & Wittrock, 2019), or the appropriateness of the questions asked (Gengler et al., 2021). These issues are beyond the scope of what we can empirically study, and they have been addressed already. Instead, we focus on three less studied phenomena that are often assumed to inhibit survey quality in the Arab region and that we can indirectly or directly test empirically: crooked samples, faked fieldwork, and biased answers provided by people who are too oppressed to report what they really believe (Gengler et al., 2021).

First and perhaps most obviously, diversity between the AB and WVS surveys might be driven by sample differences. Creating a representative sample is a challenge for any population, but the Arab region is especially difficult to represent well, because some censuses are outdated and thus demographic benchmarks are not always known. Additionally, several studies point out that Arab publics are wary to participate in certain surveys (Corstange, 2014), although findings differ on whether "Western-perceived" surveys beget more (Gengler et al., 2019) or less response (Nugent et al., 2018).

Having said that, it is a mistake to assume that sampling errors are omnipresent, and that it is and will always be impossible to decently represent any Arab country. Coverage has gotten far better over the years; for instance, the Pew Research Center in the 2000s only sampled Cairo and extrapolated the results to all of Egypt (Heath, Fisher, & Smith, 2005), whereas nowadays the fresh census of 2017 is used to represent far more different areas within Egypt—although some areas are still underrepresented for security reasons.

Additionally, the representation challenge is bigger for some Arab countries than others. There are Arab countries where censuses are outdated for political reasons—Lebanon being the prime example, with its last census in 1932—but there are also countries for which far more recent censuses are available for the latest waves, such as Egypt's 2017 census[3]. The availability of censuses could explain diversity in survey quality; in 2018, it was far more feasible to represent Egypt than Lebanon well. In the second part of this paper, we empirically test to what extent sample differences drive discrepancies between the AB and WVS surveys by weighing the surveys not by politicized censuses, but by each other. Doing so, we equalize their samples, which lays bare whether discrepancies between surveys are a factor of sample differences.

The second phenomenon concerns fieldwork procedures, which have been shown to be at times flawed in developing countries (Lupu & Michelitch, 2018). Although the AB and WVS ask similar questions, the way that those questions are asked depends on interview training, processes, and supervision. However, again, it is simplistic to assume that fieldwork throughout the Arab region is ubiquitously shotty; there are differences between the data sources and countries. Therefore, in the second part of this paper, we outline fieldwork procedures[4] and empirically test the presence and influence of enumerator fraud. Enumerator fraud might be context-dependent, as enumerators are more likely to fake interviews in rural settings with larger distances between potential respondents' houses and when they are paid per interview. If surveys got the funds they needed to supervise and backcheck all interviews, enumerator fraud could be eradicated. Presently however, this paper assesses to what extent there is evidence for widespread enumerator fraud and whether that explains any differences between surveys.

The third phenomenon that creates diversity in surveys, globally and in Arab countries, concerns socially desirable answers and (partial) nonresponse. Studies have shown that interviewers' characteristics affect respondents' willingness to respond openly, honestly, or at all (Blaydes and Gillum, 2013; L. Benstead, 2014; Lupu and Michelitch, 2018). For instance, when female respondents are not interviewed in privacy, they may provide biased answers to questions concerning their gender attitudes especially (Diop et al., 2017). Unfortunately, in lieu of cross-survey data on interviewer traits as gender or interview circumstances as privacy, we

[3] https://archive.unescwa.org/sub-site/arab-population-housingcensuses.

[4] We note however, as did scholars on "Western" surveys (Brown, Micklewright, Schnepf, & Waldmann, 2007) and on Arab surveys (Gengler et al., 2021), that information is sometimes limited (see Table 2).

cannot test socially desirable answers as directly as we can sampling differences and enumerator fraud.

What we can do here is use the diversity in the region to our advantage. Some Arab countries may be more affected by socially desirable answers and non-response than others. Countries with more authoritarian regimes limit freedom of speech, which could make Arab people wary to provide their (true) attitudes or to participate at all (Fish, 2002; Gengler et al., 2021). If our results show more comparable data in more democratic Arab countries, socially desirable answers may be part of the puzzle. Additionally, we assess missingness patterns; if questions are sensitive, respondents might skip them more often (Tourangeau & Yan, 2007). Relatedly, across the region, gender equality is expected to be perceived as more sensitive than generalized trust, as trust is shown to be a stable value less susceptible to fieldwork differences (Uslaner, 2008). Therefore, if we find more diversity in gender attitudes than in trust when comparing surveys, socially desirable answers might be part of the explanation.

Before moving on to our methods and results, let us note that the tiniest details of fieldwork can affect surveys' results in complex and unpredictable ways. It is highly unlikely that any survey manages to represent any population anywhere perfectly. We will thus probably uncover at least some discrepancies between the AB and the WVS surveys. This might create a somewhat gloomy picture, especially considering that our design is focused on uncovering discrepancies rather than providing evidence for the quality of surveys (which we cannot do, because even comparable surveys might simply be biased in the same way). However, obviously, not all discrepancies are equally severe. Some discrepancies are unlikely to truly shape substantive scholars' main conclusions. Therefore, the question we turn to now concerns the extent of discrepancies, whether discrepancies can be explained by sample differences, enumerator fraud, or socially desirable answers, and what substantive scholars can do to tackle issues.

3 Data and Methods

3.1 Surveys

This paper compares sixteen AB surveys with sixteen WVS surveys. Although some readers might regret our narrow selection (notably the exclusion of AB wave four surveys), we opted to err on the side of caution. The AB wave III and the WVS wave VI collected data within Egypt, Iraq, Kuwait, Lebanon, Libya, and Tunisia in the same year (either 2013 or 2014) and in Algeria, Jordan, Palestine, and Yemen in consecutive years (between December of 2012 and 2014). By far our least comparable case is Morocco, which was sampled in 2011 by the WVS and in 2013 by the AB. Also because the WVS carried out its fieldwork in Morocco in the tumultuous period of the Arab Spring, any diversity between the Moroccan cases is consequently least surprising and hardly indicative for quality concerns. We still included Morocco to test for the influence of differences in the timing of surveys (which did not seem to matter much, see results below). We also include the AB wave V and WVS wave VII data collected in Egypt, Iraq, Jordan, and Lebanon in the same year (either 2018 or 2019) and Tunisia in consecutive years (October-December 2018 and April-May 2019). Next to focusing on similar populations, these surveys are analogous in other respects; for instance, the AB and WVS both a) focus on similar topics in their interviews (social, cultural, and political values); b) target to represent countries' entire populations of eighteen years and older; c) employ stratified random or national full probability sampling; d) identify a similar number of strata, if stratified; e) use face-to-face interviews. We further detail fieldwork procedures—with a special focus on enumerator fraud—in Sect. 4.3.

3.2 Operationalization

General trust was measured in the AB with the item "Generally speaking, do you think most people are trustworthy or not?", which included two substantive answer categories: "most people are trustworthy" and "most people are not trustworthy" (for all original Arab questions, see Appendix 1). In the WVS, respondents were asked "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" and provided with the answer categories "most people can be trusted" and "need to be very careful". Out of the four questions we study (compare below), the general trust item is worded most differently in the AB versus the WVS. This means that if we only find discrepancies between the AB and the WVS in general trust and not trust in police or gender equality, discrepancies could be solely attributable to question wording. However, if our results show differences in the other three attitudes as well and particularly if those are bigger than for general trust, this does indicate discrepancies and that (a small difference in) question wording has little effect.

Trust in police was measured in both surveys with an item in a battery. The AB asked "I will name a number of institutions, and I would like you to tell me to what extent you trust each of them: Public security, the police" [I trust it to a great extent; I trust it to a medium extent; I trust it to a limited extent; I absolutely do not trust it]. The WVS question reads: "I am going to name a number of organizations.

For each one, could you tell me how much confidence you have in them? Is it a great deal of confidence, quite a lot of confidence, not very much confidence, or none at all? The police". Unfortunately, trust in police was not gauged in Egypt in 2018 by the WVS, but in all other surveys, all four attitudes are included in all surveys. Though English versions of the questionnaires provide different wording of the main concept of the question ("trust" vs "confidence"), the Arabic original questions are formulated with the same word, *atthiqa* (see details in Appendix 1).

Gender equality in education was measured in the AB with the statement "University education for males is more important than university education for females" and in the WVS by "A university education is more important for a boy than a girl". *Gender equality in politics* was measured by "In general, men are better at political leadership than women" in the AB and by "On the whole, men make better political leaders than women do" in the WVS. The gender equality items in both surveys provided four substantive answer categories: I strongly agree; I agree; I disagree; I strongly disagree.

General trust, trust in police, gender equality in education and gender equality in politics were coded to run from 0 to 1 (the first being dichotomous, the other three ordinal scales), with 1 indicating greater trust or greater gender equality support. We exclude missing values in our main univariate and multivariate analyses ("Diversity in survey quality") but return to their influence later ("Specks of indirect evidence of socially desirable answers").

We use several demographics as control variables in the regression models (see summary in Table 1). *Gender* distinguishes between men and women. *Age* represents respondents' age at the time of interview in years. *Educational attainment* was measured by respondents' answer to the highest level of education they had obtained: none, primary, secondary, and tertiary. *Marital status* distinguishes between single, married, and other (e.g., divorced or separated—to create similar categories across the AB and WVS). *Employment status* distinguishes between the non-employed and the employed. As the missing data on these variables comprised only 1% of respondents in our sample, we excluded them.

3.3 Analytic Strategy

Our analyses take four steps: a) first, we examine to what extent there are (consequential) differences between surveys, and next, we assess whether there is evidence for those differences being created by b) sample differences, c) enumerator fraud, and d) socially desirable answers. Within each step, several substeps are taken that we introduce here (see Appendix 2 for details).

In our first step, we analyze to what extent there are differences between AB and WVS surveys. Recall that finding no differences does not necessarily mean that surveys are of high quality—as two surveys can be biased in the same way—but uncovering discrepancies does mean that at least one of the surveys is skewed. We first assess univariate statistics—whether the means of the four attitudes differ statistically significantly between the surveys, and how big differences in distributions are (as indicated by Duncan's Dissimilarity Index, DDI) (Duncan and Duncan, 1955; see also Ortmanns and Schneider, 2016; Schneider, 2009). We then assess whether public opinion scholars studying multivariate relations on what drives trust or gender equality would arrive at different conclusions depending on the survey they used. To that end, we estimate regression models with the attitudes as dependent variables and the demographics as independent variables and estimate (separately) whether the effect of a certain demographic differs between (is moderated by) survey types (WVS, coded 0, or AB, coded 1). Evidently, we cannot address these 500 analyses in-depth, but we provide overviews and point out findings most poignant to the public opinion literature.

In our second step, we test whether any discrepancies laid bare in our first analyses are explained by sample differences, by applying innovative weights to the data and reanalyzing them. We weigh the surveys not by (outdated census-informed) demographic distributions in the populations in the Arab countries but by each other. We cannot stress enough here that these analyses are not conducted to and do not show (a closer approximation of) the "actual" attitudes in Arab countries, because that is not the main question of this paper. Rather, we are interested in whether any differences between the AB and WVS surveys in a particular country are due to their samples being different (Vandenplas and Lipps, 2024). By weighing the AB data in a country by the demographic distributions in the WVS data in that country, we can equalize their samples (at least in terms of combinations of gender*age*education*marital status*employment status, 48 strata), which greatly limits sample differences. We then re-estimate the DDIs and regression models and assess whether the results have become more similar (indicating sample differences drive survey differences) or not (indicating sample differences are small or not influential).

Third, we test whether differences between surveys are due to enumerator fraud. To that end, we first calculate the percentage of respondents in each survey whose responses are almost completely (at least 85%) the same as another respondent's, which indicates fraud by enumerators (Kuriakose and Robbins, 2015; Sarracino and Mikucka, 2017). In Arab countries with non-negligible amounts of fraudulent cases (5% or over), we then re-estimate the DDI's and regression models while excluding the fraudulent cases to test

whether enumerator fraud is not only present but explains (part of) the survey differences-interviewer identifiers were not consistently available.

Fourth and finally, we assess the influence of socially desirable answers best we can with the publicly available data. These tests are more indirect than our analyses on sampling differences and enumerator fraud, so our conclusions are less certain here, as noted in Sect. 2. Still, as discussed, we employ three proxy analyses for the extent of socially desirable answers: greater discrepancies (and more missings) in less democratic Arab countries (measured as Freedom House's Freedom in the World index reversed so that higher scores indicate higher levels of democracy); and larger amounts of non-response and greater discrepancies (and more missings) concerning gender equality than trust.

4 Results

4.1 Diversity in survey quality

Do the WVS and the AB surveys lead us to similar conclusions on Arab people's trust and support for gender equality? Generally, the results show that there are discrepancies between the AB and WVS. However, the extent of discrepancies differs across attitudes and countries (a very general overview of findings is provided in Table 4 at the end of this section).

Univariate analyses. Our univariate analyses show, first, that all four attitudes tend to have significantly higher means in the AB than in the WVS, especially in 2013–2014 (see Fig. 1). The AB data thus suggest that Arab people are, on average, significantly more trusting and supportive of gender equality than the WVS data do. In that sense, the AB data thus provide a markedly rosier picture of Arab public opinion than the WVS data.

However, not all statistically significant differences are big discrepancies that will impact opinion scholars' conclusions. Our results show, first, that there are far smaller discrepancies (lower DDI's) on general trust than on trust in police and gender equality (see Fig. 2). To achieve equal distributions between the AB and WVS, about 16–17% of respondents would have to change the answers they provided on the other three attitudes, but only 9% of respondents would have to do the same for equality in general trust. This indicates the stability of general trust, as other studies noted (Uslaner, 2008).

Second, discrepancies are larger for some surveys than others. Particularly noteworthy are the surveys of Algeria '13–'14, Kuwait '14, and Lebanon '13, with remarkably high DDI's (over 0.20 across attitudes). At the other end of the spectrum, across the four attitudes, Iraq '13 and Lebanon '18's DDIs are under 0.1, indicating only 10% of respondents have to change their answers to achieve equality between the WVS and the AB.

Multivariate analyses. Univariate distributions do not tell us much, however, about whether the surveys lead scholars to similar conclusions on what socio-demographic characteristics are associated with higher or lower trust and gender equality. That question is answered by our multivariate analyses. Table 1 summarizes our results, showing whether relations between socio-demographics and attitudes were non-significant (0), positive (+), or negative (–) in regression models. When the data sources returned similar results—i.e., interactions between survey type and demographics were non-significant—the same sign is shown for the AB and the WVS—e.g., both 0. When relations significantly differed between the surveys but only in strength, the Table shows in what survey the relation was more strongly positive (++) or negative (– –). Bolded text indicates that public opinion scholars' conclusions would depend on the data source they used, because a relation is non-significant in one data source (0) but significant in another (+ or –) or because a relation was positive and significant in one data source (+) and negative and significant in the other (–).

Again, Table 1 indicates diversity in discrepancies. On the one hand, most cells show the same signs, which indicates demographics relate to trust or gender equality in the same way in the AB and WVS. The majority of tested cases—65%—show that predictors drive trust and gender equality similarly across data sources, which implies that substantive scholars' conclusions would be the same regardless of whether they opted for the WVS or the AB.

Of course, that also implies a flip side: in 35% of tested cases, relations are statistically significantly different between the WVS and the AB (see also Fig. 3). Although that might seem like a shocking number at first, it is important to note that not all of those cases would meaningfully alter substantive scholars' conclusions on what drives trust or gender equality. Half of the time, relations are merely significantly stronger in one data source than another.

In 17% of tested cases, however, relations differ to such an extent that they would alter substantive scholars' conclusions (bolded cells). In these cases, the AB data often do not find a significant relation between a demographic and an attitude whereas the WVS data do. That is interesting, given the arguments in the survey comparison literature that data sources that find stronger effects might be more adequate (e.g., Kieffer, 2010; Manning et al., 2019). More generally, though, these results show that in about one in five tested cases, substantive scholars' conclusions would completely depend on their choice of data source.

Table 1

Differences in how socio-demographic indicators relate to trust and gender equality in multivariate models

		General trust	Trust in police	Educational gender equality	Political gender equality	# Discrepancies
Algeria '13–'14	Female	0; 0	+; +	++; +	+; +	6 (8)
	Age	**–; 0**	+; +	0; 0	–; –	
	Education	–; —	**–; 0**	+; +	+; +	
	Single	**+; 0**	–; –	0; 0	0; 0	
	Employed	**+; 0**	**–; 0**	**–; +**	0; 0	
Egypt '13	Female	0; 0	0; 0	+; +	+; +	3 (5)
	Age	++; +	0; 0	**–; 0**	0; 0	
	Education	0; 0	**–; 0**	++; +	0; 0	
	Single	0; 0	0; 0	**+; 0**	0; 0	
	Employed	0; 0	–; –	0; 0	0; 0	
Iraq '13	Female	0; 0	+; ++	++; +	+; +	5 (8)
	Age	0; 0	**–; 0**	0; 0	**–; 0**	
	Education	0; 0	0; 0	++; +	+; +	
	Single	0; 0	**+; –**	**+; 0**	**+; –**	
	Employed	0; 0	+; +	0; 0	0; 0	
Jordan '13–'14	Female	0; 0	+; +	+; +	+; ++	3 (5)
	Age	0; 0	0; 0	+; +	0; 0	
	Education	0; 0	–; –	+; +	+; +	
	Single	0; 0	**+; 0**	0; 0	0; 0	
	Employed	**–; 0**	0; 0	0; 0	**+; –**	
Kuwait '14	Female	0; 0	0; 0	++; +	++; +	1 (9)
	Age	0; 0	+; +	0; 0	0; 0	
	Education	0; 0	+; ++	+; +	–; –	
	Single	+; +	**–; 0**	0; 0	++; +	
	Employed	++; +	0; 0	++; +	++; +	
Lebanon '13	Female	0; 0	0; 0	+; +	+; ++	3 (7)
	Age	0; 0	0; 0	+; ++	**–; +**	
	Education	0; 0	0; 0	++; +	++; +	
	Single	0; 0	0; 0	**+; 0**	0; 0	
	Employed	0; 0	0; 0	+; +	**+; 0**	

Table 1

(Continued)

		General trust	Trust in police	Educational gender equality	Political gender equality	# Discrepancies
Libya '14	Female	0; 0	**–; +**	+; +	++; +	1 (3)
	Age	+; +	0; 0	0; 0	0; 0	
	Education	0; 0	–; –	+; +	++; +	
	Single	0; 0	0; 0	–; –	0; 0	
	Employed	0; 0	0; 0	0; 0	0; 0	
Morocco '11–'13	Female	0; 0	+; +	+; ++	+; +	1 (4)
	Age	0; 0	0; 0	0; 0	0; 0	
	Education	0; 0	—; –	+; ++	+; +	
	Single	0; 0	0; 0	**+; 0**	0; 0	
	Employed	0; 0	0; 0	0; 0	0; 0	
Palestine '12–'13	Female	0; 0	0; 0	++; +	+; +	2 (6)
	Age	0; 0	0; 0	+; ++	**–; +**	
	Education	0; 0	0; 0	+; +	**+; 0**	
	Single	0; 0	0; 0	++; +	+; +	
	Employed	0; 0	0; 0	+; ++	+; +	
Tunisia '13	Female	0; 0	++; +	++; +	+; +	6 (10)
	Age	**0; +**	+; +	**0; +**	**–; +**	
	Education	++; +	0; 0	**+; 0**	++; +	
	Single	0; 0	0; 0	0; 0	0; 0	
	Employed	0; 0	**0; +**	**0; +**	0; 0	
Yemen '13–'14	Female	0; 0	+; +	++; +	++; +	3 (6)
	Age	++; +	0; 0	0; 0	**+; 0**	
	Education	0; 0	**–; 0**	+; +	0; 0	
	Single	**–; 0**	0; 0	0; 0	0; 0	
	Employed	+; +	0; 0	0; 0	0; 0	
Egypt '18	Female	0; 0	N/A	+; +	+; ++	1 (5)
	Age	0; 0	N/A	++; +	**0; –**	
	Education	0; 0	N/A	+; ++	+; +	
	Single	0; 0	N/A	+; ++	+; +	
	Employed	0; 0	N/A	+; +	+; +	

Table 1

(Continued)

		General trust	Trust in police	Educational gender equality	Political gender equality	# Discrepancies
Iraq ’18	Female	0; 0	**0; +**	+; +	++; +	6 (9)
	Age	**0; +**	**0; –**	0; 0	**0; –**	
	Education	–; –	—; –	++; +	+; +	
	Single	**+; 0**	**–; 0**	+; +	0; 0	
	Employed	0; 0	0; 0	0; 0	0; 0	
Jordan ’18	Female	0; 0	**0; +**	+; +	++; +	4 (5)
	Age	0; 0	**+; 0**	+; +	+; +	
	Education	**–; 0**	–; –	+; +	0; 0	
	Single	0; 0	**–; 0**	0; 0	0; 0	
	Employed	0; 0	0; 0	0; 0	0; 0	
Lebanon ’18	Female	**0; –**	+; +	+; +	++; +	5 (9)
	Age	0; 0	+; ++	0; 0	0; 0	
	Education	0; 0	++; +	**+; 0**	++; +	
	Single	0; 0	+; +	**+; 0**	0; 0	
	Employed	**0; +**	**0; +**	0; 0	0; 0	
Tunisia ’18–’19	Female	0; 0	+; +	++; +	++; +	5 (10)
	Age	0; 0	**0; +**	**+; 0**	0; 0	
	Education	0; 0	**0; –**	+; ++	0; 0	
	Single	0; 0	**0; –**	—; –	—; –	
	Employed	0; 0	0; 0	0; 0	**–; 0**	
# Discrepancies		11 (16)	21 (28)	12 (35)	12 (30)	

Fig. 1

Differences in average trust and support for gender equality between the AB and the WVS

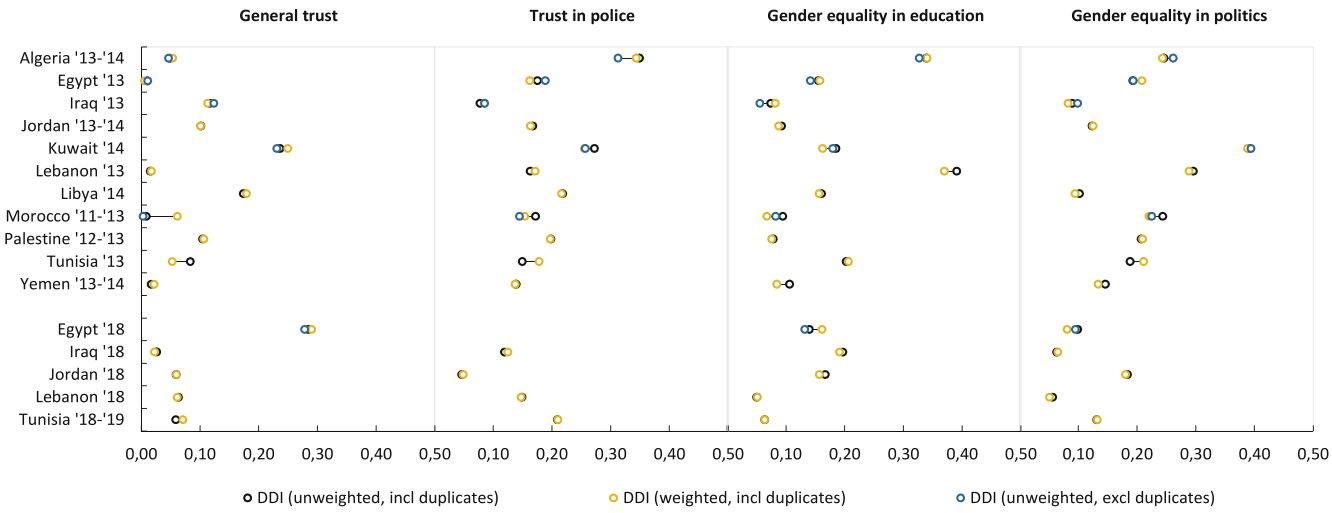

Fig. 2

The extent of univariate differences between the AB and the WVS, as expressed by Duncan's Dissimilarity Index

Having said that, some Arab countries are driving that number up, while other countries' results are pretty robust across data sources (see Fig. 3). The surveys with most discrepancies are Algeria 2013–2014, Iraq 2018, and Tunisia 2013.[5] In those countries, the conclusions of substantive scholars depend completely on the data source they selected in no less than 30% of tested cases. Our analyses however also uncover several countries in which the choice of survey

[5] Tunisia 2013 had 18 out of 24 provinces not sampled in the WVS compared to 0 of 24 in the AB which can explain the difference partly; Iraq 2018 was sampled better in terms of province coverage then in 2013, when 10 out of 19 provinces were not covered by the WVS. No data on province coverage is available for Algeria 13–14 in WVS, but 8 of 58 provinces were not sampled in the AB. More information on these samples problematic in multivariate analyses and others see in Table 2.

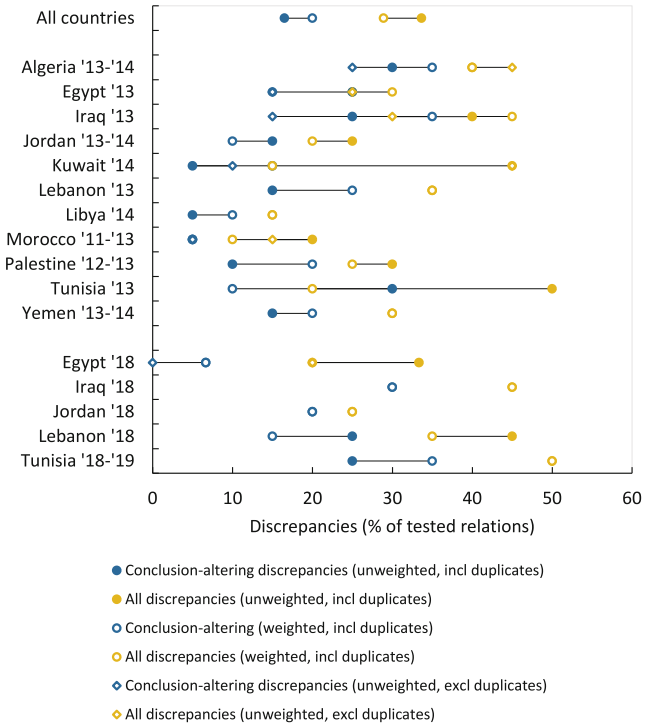

Fig. 3

Discrepancies between the WVS and the AB in multivariate relations between socio-demographics and trust and gender equality

would almost never impact scholars' conclusions, no matter what they studied: Libya 2014, Morocco 2011–2013, and Kuwait 2014 (only 5%).

In the most general terms, these results of the first step of our analyses show a smorgasbord. We have uncovered major discrepancies in the data, but there are also cases where few discrepancies were flagged throughout our close scrutiny. The question we turn to now is what might explain that diverse pattern, starting with sample differences.

4.2 Sample differences do not explain much

This section assesses whether the differences between the AB and WVS surveys laid bare above are due to their samples being different. We weighted the data by each other, so that their demographic distributions are equal (see Sect. 3) and re-estimated the DDIs and regression models. If sample differences are spawning the discrepancies between the data sources, we would see improvements in our new, weighted results. However, in short, we do not.

To elaborate shortly, Fig. 2 also plots the DDIs in the weighted data, but they strongly overlap with the unweighted DDIs. In most cases (44 out of 63), the change in DDI is smaller than 0.01, which is negligible.[6] We only see (over 0.01) smaller discrepancies between the WVS and AB after restricting sample differences in twelve cases (out of 63), and we even find a few (seven) cases where discrepancies got (over 0.01) larger by weighing the data. Generally, univariate differences in trust and gender equality between the AB and WVS thus do not change much when their samples are equalized.

Fig. 3 shows that discrepancies between the data sources in what predicts trust or gender equality are also not alleviated by equalizing the samples. In fact, we find more conclusion-altering discrepancies after weighing in nine countries (out of sixteen). The number of discrepancies stays the same in four countries and only reduces in three. Altogether, these results provide no indication that sample differences caused the discrepancies between the AB and WVS data, because differences between the data sources remain after sample differences are restricted.

4.3 Little evidence of faked data

The second possible explanation for discrepancies between data sources concerns fieldwork procedures, such as a lack of supervision and faked data by enumerators paid per interview (Gengler et al., 2021). This section first describes the fieldwork procedures in the AB and WVS and then empirically tests whether there is evidence of widespread fraud by enumerators.

Both the AB and the WVS use different sampling strategies in varying countries but tend to use multistage stratified probability samples across countries (*Methodology—Arab Barometer*, 2021; *Methodology—World Values Survey*, 2021). Generally, countries were first stratified by regions (e.g., governorates, provinces), and at times stratified again (e.g., by urbanization). Primary sampling units (PSUs) were randomly selected, usually using probability proportional to size (PPS). Next, blocks or clusters of about 200 households tend to be randomly selected within these PSUs. Households are usually selected using random walks with random starting points, and respondents were selected using kish grids or first/last birthday methods. It is notable that the WVS tends to provide detailed descriptions of sampling procedures within each country, while the AB only provides a general statement for all—which might inspire greater confidence in the WVS. However, it should

[6] 0.01 means that the percentage of respondents who would have to change their answers to achieve equality in the distribution of a particular attitude between the AB and WVS is only 1 percentage point different after weighing than before—even though we did find some substantial DDIs before weighing (see Fig. 2).

Table 2

Data on AB and WVS samples collected from the field reports published on their websites (AB III, AB V, and WVS 6, WVS 7)

		Callbacks	Supervised %	Backchecked %	Company	Unsurveyed provinces	Total N of provinces	Percent match	Number of probl.cases
Algeria 13–14	AB	*ND*	_	_	Okba Com	8	58	0.6	8
	WVS	No	22	0	Okba Com	ND		19.5	237
Egypt 13	AB	*ND*	–	–	MADA Center	6	27	19	227
	WVS	Some (undef.)	0	0	TNS Egypt	ND		7.37	112
Egypt 18	AB	*ND*	–	–	Local Res. Org	3	27	0	0
	WVS	Min.3 w/repl	20	25	ERTC	0	27	0.6	7
Iraq 13	AB	*ND*	–	–	IIACSS	0[(a)]	19	7.4	90
	WVS	Min.3	5	15	IIACSS	10	19	2.3	28
Iraq 18	AB	*ND*	–	–	IIACSS	3	19	0	0
	WVS	Min.3 w/repl	10	15	IIACSS	2	19	0.2	2
Jordan 13–14	AB	*ND*	–	–	CSS	0	12	0	0
	WVS	None req	99	40	CSS	0	12	0	0
Jordan 18	AB	*ND*	–	–	CSS	0	12	0	0
	WVS	Min.3	99	90	NAMA	0	12	0	0
Kuwait 14	AB	*ND*	–	–	Gulf Opinions	N/A		0	0
	WVS	No	0	0	Gulf Opinions	N/A		10.2	133
Lebanon 13	AB	*ND*	–	–	Statistics Lebanon	2[(b)]	9	0.9	11
	WVS	Unlimited	0	20	REACH	0	9	1.2	14
Lebanon 18	AB	*ND*	–	–	Statistics Lebanon	1	9	0.3	4
	WVS	Min.3	0	75	Statistics Lebanon	3	9	0.2	2

Table 2

(Continued)

		Callbacks	Supervised %	Backchecked %	Company	Unsurveyed provinces	Total N of provinces	Percent match	Number of probl.cases
Libya 14	AB	*ND*	–	–	RCC,Univ.of Benghazi	0[c]	22	0.6	8
	WVS	3	*ND*	*ND*	RCC,Univ.of Benghazi	0	22	1.8	40
Morocco 11–13	AB	*ND*	–	–	Hassan II M.Uni	0[d]	16	1.5	17
	WVS	No	*ND*	*ND*	SEREC	7	16	16	191
Palestine 12–13	AB	*ND*	–	–	PCPSR	0[e]	16	1.6	20
	WVS	3	15	10	JMCC	0	16	3.4	34
Tunisia 13	AB	*ND*	–	–	Sigma Conseil	0[f]	24	0.5	6
	WVS	Unlimited	0	20	Emrhod	18	24	2.6	32
Tunisia 18–19	AB	*ND*	–	–	1 to 1 FRP	0[g]	24	0	0
	WVS	Min.3 w/repl	10	10 + GPS	ASSF	0	24	0.3	3
Yemen 13–14	AB	*ND*	–	–	YCSO	0[h]	21	0.8	10
	WVS	2	0	20	PERCENT	Not 0[i]	21	1.2	12
			100 for all AB	20 for all AB					

N/A not applicable, *ND Not data available.* Percent match was calculated by the authors.

[a] 1 governorate urban only. [b] o sunni population in El Nabatieh; no Shia population in North; no Christian population in El Nabatieh; no Druze population in Beirut, North, South, and El Nabatieh. [c] Ghat urban only; Wadi al Shatii rural only. [d] Governorates Guelmim-Es Semara, Laayoune-Boujdour-Sak urban only; Oued Ed Dahab rural only. [e] No refugee camps in Salfit and Qalqilya; no rural area in Jabalia, and Deir al-Balah. [f] Governorates Tunis, Ariana, Monastir, Tozeur urban only. [g] Governorates Tunis and Monastir urban only. [h] Governorate Aden urban only. [i] Number of "selected provinces" unclear

also be noted that using multistage probability samples and random walks has been shown to lead to higher unit nonresponse bias (Kohler, 2007).

Table 2 summarizes fieldwork procedures further, again based on the surveys' websites (*Methodology—Arab Barometer*, 2021; *Methodology—World Values Survey*, 2021). It is particularly noteworthy, first, that some crucial information is missing (e.g., payment of interviewers, language of interviews in WVS, NDs in table), especially for the AB, or only provided in a generalized manner (star signs). This evidently does not help get to the root of what explains the diversity between surveys.

Second, the AB and WVS differ in the quality checks they report using. The AB reports all its interviews are supervised and one in five is backchecked, while the WVS only reports supervising between zero and 22% of its interviews and does not report doing any backchecks. Based on this information, we would expect the WVS enumerators to have greater opportunities to commit fraud.

However, a closer look at Table 2 keeps us from drawing strong conclusions based on these data. The response rates provided by the AB III are reported at the same level across countries, and those of the WVS are peculiarly high or absent. Additionally, multiple samples excluded provinces, precluding the country-wide representation that both websites claim. Therefore, we urge scholars not to take the documentation provided at face value.

To more directly assess fieldwork procedures, we calculated the percent match in each survey: the number of respondents whose answers overlap with those of another respondent in at least 85% of all questions asked (Kuriakose and Robbins, 2015). Because duplicates are exceedingly unlikely to happen by chance (Slomczynski, Powalko, and Krauze, 2017) they point to fraud by enumerators. Fig. 4 shows that, encouragingly, hardly any surveys suffer from widespread enumerator fraud. In 25 cases (out of 32), the percentage of duplicated response patterns is under 5, which would hardly impact results. There are only seven cases of non-negligible fraud (percent match ≥5%): Iraq 2013 (AB), Egypt 2013 (AB and WVS), Kuwait 2014 (WVS), Egypt 2018 (WVS), Morocco 2011–2013 (WVS), and Algeria 2013–2014 (WVS). Our results thus indeed show more evidence of fraud in the WVS data than in the AB data. However, it remains remarkable that the AB data also contain some near-duplicates, as their website states all interviews were supervised and the data were checked for percent matches. More generally, though, our results show that, although there are some worrisome cases, enumerator fraud is hardly widespread in the Arab surveys, which undercuts the notion that fieldwork would be commonly faked.

Additionally, the few cases of non-negligible fraud we did uncover do not seem to have biased results. When we exclude the fraudulent cases in the seven flagged surveys and re-estimate their DDIs, only eight (out of twenty-three) DDIs improved somewhat and thirteen did not—and improvements were too small to truly influence the conclusions of substantive scholars (see Fig. 2). Likewise, discrepancies between the AB and WVS in what predicts trust and gender equality by and far (99 out of 115 cases) remained the same when fraudulent cases were excluded (see Fig. 3). Altogether, our results show little evidence of fraud, and the few problematic cases that we do find hardly explain any discrepancies between the AB and the WVS data. Counter to what some might believe, there are no indications of flagrantly faked fieldwork in the Arab region.

4.4 Specks of indirect evidence of socially desirable answers

If sample differences and enumerator fraud can hardly explain the discrepancies between the AB and the WVS, perhaps far-reaching socially desirable answering does. Although we do not have the data necessary to test the extent of socially desirable answers directly, we can review three proxies: widespread missingness, and more aberrant data in less democratic Arab countries and concerning more sensitive topics. Although we have to be more careful in our conclusions here, our indirect tests (once again) provide no consistent indications of socially desirable answers, although we do uncover a few specks.

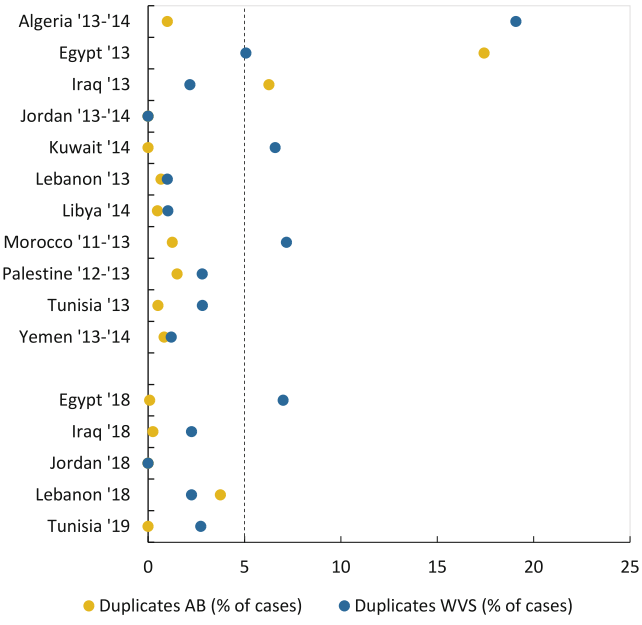

Fig. 4

Percentage of duplicates by sample in the WVS and the AB

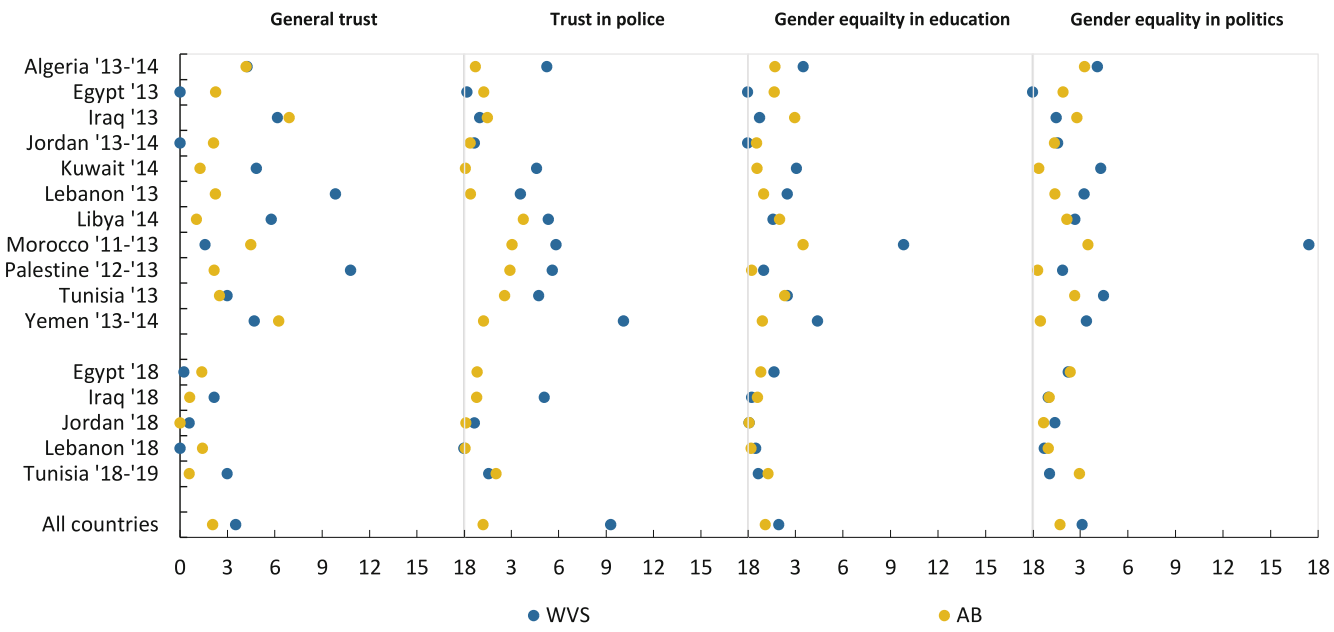

Fig. 5

Percentage of missing values by sample in the WVS and the AB

First, even though we focus on some contentious issues, our data generally do not contain many respondents who declined to answer particular questions (see Fig. 5). Missings total 3% on average, and that average is skewed by a few problematic cases, most notably the remarkably high percentage of missing values on gender equality in the Moroccan WVS survey.

Our results also do not consistently show more peculiar data in less democratic countries. Table 3 shows that correlations between countries' levels of democracy and the extent of their discrepant data are mostly small or positive. More democratic Arab countries do not have fewer missings on or smaller (univariate or multivariate) discrepancies in institutional trust or gender equality in education or politics. Only for generalized trust do our results show less discrepant results in more democratic countries (i.e., correlations are negative). This is unexpected, as general trust was assumed to be the least controversial issue and thus least affected by socially desirable answering. A possible explanation is that all respondents might provide similar, socially desirable answers in more authoritarian countries. This might reduce the discrepancies between data sources in less democratic countries—although the data would still be biased. However, this evidently still does not explain why we only find this pattern for general trust and why we find big discrepancies in democratic countries. More generally, we presently err on the side of caution and only conclude that the data do not seem to be consistently more aberrant in less democratic countries.

Third, tentatively, socially desirable answers might have skewed the answers that respondents provided on more sensitive topics, but not in a way that harshly impacts scholars' conclusions on what drives trust or gender equality. In this assessment, we assume that generalized trust is the least sensitive issue in the region that we study and gender equality the most (see also Uslaner, 2008). The gaps between the AB and WVS data are far smaller on generalized trust (DDI of 0.09) than on both types of gender equality (DDIs of 0.16 and 0.17; see Fig. 2). There are also fewer total discrepancies uncovered by our multivariate analyses on generalized trust (16) than on gender equality (35 and 30; see Table 1). These results are as we would expect if socially desirable answering were present. However, the number of conclusion-altering discrepancies uncovered by our regression models hardly differs for trust (11) and gender equality (12). Therefore, our indirect results seem to be in line with socially desirable answers, but even if that is the case, socially desirable answers do not seem to bias scholars' conclusions on what predicts support for gender equality.

4.5 Overview of results

Making up the final balance, all of our analyses tell a tale of diversity. We find diversity across the countries in the discrepancies uncovered and even diversity within surveys concerning what they perform better or worse on. We can-

Table 3

Correlations between countries' levels of democracy and aberrant data

		Missing, democracy	DDIs, democracy	Multivariate total discrepancies, democracy	Multivariate conclusion-altering discrepancies, democracy
Generalized trust	WVS	–0.06	–	–	–
	AB	–0.39	–	–	–
	Both	–0.25	–0.2	–	–
Institutional trust	WVS	–0.18	–	–	–
	AB	0.33	–	–	–
	Both	0.01	0.13	–	–
Gender equality in education	WVS	0.07	–	–	–
	AB	0.23	–	–	–
	Both	0.15	0.04	–	–
Gender equality in politics	WVS	0.15	–	–	–
	AB	0.37	–	–	–
	Both	0.24	0.14	–	–
All	–	–	–	0.29	0.18

not state outright—and it might even be impossible to ever do so—which surveys are better and which are worse.

What we can say is that our results are too all over the place to warrant generalizations on Arab surveys' quality. What is more, the oft-heard reasons for the presumed difficulties in surveying Arab public do not seem to hold either. Discrepancies were not caused by differences in samples' distributions or by what little enumerator fraud we uncovered. And although there might be some socially desirable answers, they do not seem to impact substantive scholars' conclusions on the causes of Arab people's trust and gender equality (Table 4).

5 Recommendations for Substantive Scholars

What does this all mean for scholars of Arab public opinion? Unfortunately, our results do not permit a blanket sigh of relief. Some surveys return comparable results on some points, but we also uncovered several problematic cases that could potentially impact substantive scholars' findings to such an extent that their conclusions would be entirely dependent on what data source they chose to use. To help scholars avoid that, we now provide a few tips and tricks.

First off, scholars should know that, generally, if they select the AB data, their results probably paint a more optimistic "liberal" picture of the region. They are also less likely to find strong drivers of trust and gender equality attitudes. The WVS data generally paint a more grim picture but tend to find stronger effects. Our most concrete advice is that, if possible, scholars should combine the AB and the WVS data. They can control their models for survey type to adjust for differences in mean values. Pivotally, scholars can then estimate the main relations they are interested in a) while also including a moderation by survey type (see Glas et al., 2019), or b) per survey type separately. This will show scholars whether their conclusions hold across data sources. The big advantage of this approach is that, given the plethora of discrepancies between the AB and the WVS, any conclusion that holds across the data sources can be considered to be remarkably strong evidence.

Additionally, we recommend that substantive scholars pay special attention to Algeria 2013–2014, Iraq 2018, and Tunisia 2013. These cases show such discrepancies concerning public opinion scholars' focus—what predicts certain attitudes—that we strongly urge caution. If scholars are not interested in country-level effects and thus a slightly lower higher-level N would not hinder them much, we would recommend excluding these cases entirely, because we dare not vouch for their quality. If scholars need these cases because they are interested in country-level effects, we recommend that they test the robustness of their models by also (sequentially) excluding Algeria, Iraq, and Tunisia to make sure that survey errors do not bias their conclusions.

Table 4

Levels of univariate and multivariate discrepancies by sample in the WVS and the AB

	Univariate discrepancies	Multivariate discrepancies
Yemen '13–'14	Low	Medium
Morocco '11–'13	Medium	Low
Palestine '12–'13	Medium	Medium
Libya '14	Medium	Low
Egypt '18	Medium	Medium
Egypt '13	Medium	Medium
Jordan '13–'14	Low	Medium
Jordan '18	Medium	Medium
Lebanon '18	Low	Medium
Iraq '13	Low	Medium
Iraq '18	Medium	High
Kuwait '14	High	Low
Tunisia '18–'19	Medium	High
Tunisia '13	Medium	Medium
Lebanon '13	High	Medium
Algeria '13–'14	High	High

6 Conclusion and Discussion

Comparative survey data quality is the foundation stone for thousands of publications in the social sciences. Some countries, especially European and English-speaking ones, have been routinely surveyed by various companies for decades. It is quite safe to rely on the data from those countries, as scholars analyze the datasets looking for fraud, non-compatibility, and spurious trends (Hayford and Morgan, 2008; Dubrow and Tomescu-Dubrow, 2016; Ortmanns and Schneider, 2016; Manning et al., 2019). This is not the case for Arab North African and Middle Eastern countries, which have been involved in the global endeavor of social surveys only since the late 2000s. To our knowledge, little systematic, empirical quality assessment has been conducted on Arab surveys yet (see Benstead, 2018a, p. 224; Benstead, 2018b).

The present study empirically compared two prominent data sources: the Arab Barometer (AB) and the World Values Survey (WVS). We analyzed four attitudes often studied by public opinion scholars—generalized and institutional trust and support for gender equality in education and in politics—in thirty-two Arab surveys in total (e.g., Diop et al., 2017; Fish, 2011; Inglehart and Norris, 2003; Spierings, 2019; Sika, 2020). Our results are diverse. Certain surveys, such as Libya 2014, show relatively few discrepancies across the board, but we uncovered a multitude of discrepancies for others, such as Algeria 2013–2014. Further complicating matters, we also find diversity within surveys; for instance, the AB and WVS surveys on Lebanon in 2013 are far more robust on trust than on gender equality. Especially poignant are our findings that substantive scholars' conclusions on what drives trust or gender equality would be similar in two-thirds of tested cases but would depend on what data source scholars selected in one in five tested cases. This hodgepodge of findings does make one thing very clear; there are far too many differences to claim that all Arab publics are inherently unsurveyable.

As our results point out several problematic cases, the question "what do we do now?" looms. This question is more difficult to answer than might be expected, because some of the most prolific presumed reasons for why Arab survey quality would be lacking—unrepresentative samples, faked fieldwork, and socially desirable answers—hardly explain the discrepancies we find when we addressed them empirically. Because the Arab region is of great interest to politicians, policy-makers, pundits, and scholars alike, and we would never recommend ceasing investigations into the region using all available information.

We provided substantive scholars with tips and tricks on how to utilize these surveys—controlling for moderations with data source and omitting a few outlying cases—so we would now like to turn our attention to the surveys themselves. In many countries in the region, data are still sporadic, with evident gaps due to lack of funding, interest, human resources, or risks of working in some settings.

We are convinced that if surveys received the resources they needed, many issues could be overcome in the future. Presently, however, we remain at somewhat of a loss as to what lies at the root of all this diversity in survey quality, which is a pity, because that understanding could help us use the surveys more effectively. For instance, if we could test interview privacy and that turned out to underlay some discrepancies, substantive scholars could simply control for (moderations with) privacy in their models. However, meta-data on interviews are not consistently publicly available. If they were, we could carry out more thorough methodological evaluations, and substantive scholars could assess whether interview differences substantially altered their conclusions. We therefore encourage the responsible scholars to publish more open and detailed fieldwork documentation and sampling procedures, which can help Arab public opinion scholars get even more out of their vital data.

Having said all that, we still strongly encourage scholars to study the Arab region and use most of the available data on Arab countries. Data are not biased by crooked samples, nor are they figments of faked fieldwork, and evidence of socially desirable answers is sparse. The Arab region presents questions too burning to neglect, and the inclusion of Arab countries into cross-national research adds to sample diversity and better representation of the opinions, mores, and values of the people around the globe. The endeavour of conducting mass social surveys in countries beyond Europe is one of the most important achievements of the social sciences of recent decades, and improving their scope and quality is a brilliant goal for the new generation of scholars.

Acknowledgements We want to thank Niels Spierings for organizing a oneday seminar in Radboud University, Nijmegen, for a team of Europe-based sociologists interested in Arab region. That meeting with Pamela Abbott, Amy Alexander, Lars Berger, and Niels was so fruitful that the authors of the piece got acquainted and conceived this paper right away. We hope that kind of short and focused seminars will gain more popularity, as they really bring people together. We also thank all the discussants of the paper at the April conference 2021 at the Higher School of Economics, Moscow, especially Boris Sokolov, for in-depth comments that helped us revise the text, as well as all the reviewers for their time and diligence.

References

Alexander, A.C., & Welzel, C. (2011). Islam and patriarchy: How robust is muslim support for patriarchal values? *International Review of Sociology*, *21*(2), 249276. https://doi.org/10.1080/03906701.2011.581801.

Benstead (2018a). Survey research in the Arab world. In *The Oxford handbook of polling and survey methods*. https://doi.org/10.1093/oxfordhb/9780190213299.001.0001.

Benstead (2018b). Survey research in the Arab world: Challenges and opportunities. *Political Science Politics*, *51*(3), 535–542.

Benstead, L. (2014). Effects of interviewer-respondent gender interaction on attitudes toward women and politics: Findings from Morocco. *International Journal of Public Opinion Research*, *26*(3), 369–383. https://doi.org/10.1093/ijpor/edt024.

Blaydes, L., & Gillum, R.M. (2013). Religiosity-ofinterviewer effects: Assessing the impact of veiled enumerators on survey response in Egypt. *Politics and Religion*, *6*(3), 459–482. https://doi.org/10.1017/S1755048312000557.

Brown, G., Micklewright, J., Schnepf, S.V., & Waldmann, R. (2007). Cross-national surveys of learning achievement: How robust are the findings? *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *170*(3), 623–646. https://doi.org/10.1111/j.1467-985X.2006.00439.x.

Clark, J.A., & Cavatorta, F. (Eds.). (2018). *Political science research in the Middle East and North Africa: methodological and ethical challenges*. Oxford, New York: Oxford University Press.

Corstange, D. (2014). Foreign-sponsorship effects in developing-world surveys: Evidence from a field experiment in Lebanon. *Public Opinion Quarterly*, *78*(2), 474–484. https://doi.org/10.1093/poq/nfu024.

Diop, A., Tessler, M., Wittrock, J., & Jardina, A. (2017). Antecedents of trust among citizens and non-citizens in Qatar. *Journal of International Migration and Integration*, *18*(1), 183–202. https://doi.org/10.1007/s12134-016-04740.

Dubrow, J.K., & Tomescu-Dubrow, I. (2016). The rise of cross-national survey data harmonization in the social sciences: emergence of an interdisciplinary methodological field. *Quality Quantity*, *50*(4), 1449–1467. https://doi.org/10.1007/s11135-015-0215-z.

Duncan, O.D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210–217. https://doi.org/10.2307/2088328.

Fish, M.S. (2002). Islam and authoritarianism. *World Politics*, *55*(1), 4–37. https://doi.org/10.1353/wp.2003.0004.

Fish, M.S. (2011). *Are muslims distinctive?: A look at the evidence*. Oxford University Press.

Gengler, J.J., Le, K.T., & Wittrock, J. (2019). Citizenship and surveys: group conflict and nationality-ofinterviewer effects in Arab public opinion data. *Political*

Behavior. https://doi.org/10.1007/s11109-019-09583-4.

Gengler, J.J., Tessler, M., Lucas, R., & Forney, J. (2021). Why do you ask?' The nature and impacts of attitudes towards public opinion surveys in the Arab world. *British Journal of Political Science*, *51*(1), 115–136. https://doi.org/10.1017/S0007123419000206.

Glas, S., & Spierings, N. (2020). Connecting contextual and individual drivers of anti-americanism in Arab countries. *Political Studies*, *69*(3), 686–708. https://doi.org/10.1177/0032321720923261.

Glas, S., & Spierings, N. (2021). Rejecting homosexuality but tolerating homosexuals: The complex relations between religiosity and opposition to homosexuality in 9 Arab countries. *Social Science Research*. https://doi.org/10.1016/j.ssresearch.2021.102533.

Glas, S., Spierings, N., & Scheepers, P. (2018). Reunderstanding religion and support for gender equality in Arab countries. *Gender G Society*, *32*(5), 686–712. https://doi.org/10.1177/0891243218783670.

Glas, S., Spierings, N., Lubbers, M., & Scheepers, P. (2019). How polities shape support for gender equality and religiosity's impact in Arab countries. *European Sociological Review*, *35*(3), 299–315. https://doi.org/10.1093/esr/jcz004

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Puranen, B., et al. (2020). *World Values Survey: Round seven—country-pooled datafile*. Madrid, Vienna: JD Systems Institute and WVSA Secretariat. https://doi.org/10.14281/18241.13.

Hayford, S.R., & Morgan, S.P. (2008). The quality of retrospective data on cohabitation. *Demography*, *45*(1), 129–141. https://doi.org/10.1353/dem.2008.0005.

Heath, A., Fisher, S., & Smith, S. (2005). The globalization of public opinion research. *Annual Review of Political Science*, *8*, 297–333. https://doi.org/10.1146/annurev.polisci.8.090203.103000.

Inglehart, R., & Norris, P. (2003). *Rising tide: gender equality and cultural change around the world*. Cambridge: Cambridge University Press.

Inglehart, R., Haerpfer, C., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Puranen, B., et al. (2014). *World Values Survey: round six—country-pooled datafile version*. Madrid: JD Systems Institute. https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

Jamal, A. (2007). When is social trust a desirable outcome?: examining levels of trust in the Arab world. *Comparative Political Studies*, *40*(11), 1328–1349. https://doi.org/10.1177/0010414006291833.

Jamal, A., & Tessler, M. (2008). The democracy barometers (part II): Attitudes in the Arab world. *Journal of Democracy*, *19*(1), 97–110.

Jamal, A., Shikaki, K., Tessler, M., Al-Emadi, D., Eyadat, Z., & Meddeb, Y. (2014). Arab barometer wave iii. Inter-university Consortium for Political and Social Research [distributor]. https://www.arabbarometer.org/surveys/arab-barometer-wave-iii/

Jamal, A., Shikaki, K., Tessler, M., Al-Emadi, D., Eyadat, Z., & Meddeb, Y. (2019). Arab Barometer wave v. Inter-university Consortium for Political and Social Research [distributor]. https://www.arabbarometer.org/surveys/arab-barometer-wave-v/

Kieffer, A. (2010). Measuring and comparing levels of education: methodological problems in the classification of educational levels in the european social surveys and the french labor force surveys. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *107*(1), 49–73. https://doi.org/10.1177/0759106310369974.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, *1*(2), 5567.

Kostenko, V.V., Kuzmuchev, P.A., & Ponarin, E.D. (2016). Attitudes towards gender equality and perception of democracy in the Arab world. *Democratization*, *23*(5), 862–891. https://doi.org/10.1080/13510347.2015.1039994.

Kuriakose, N., & Robbins, M. (2015). *Don't get duped: fraud through duplication in public opinion surveys*. SSRN Scholarly Paper No. ID 2580502. Rochester: Social Science Research Network. https://papers.ssrn.com/abstract=2580502

Lupu, N., & Michelitch, K. (2018). Advances in survey methods for the developing world. *Annual Review of Political Science*, *21*(1), 195–214. https://doi.org/10.1146/annurev-polisci-052115-021432.

Manning, W.D., Joyner, K., Hemez, P., & Cupka, C. (2019). Measuring cohabitation in US national surveys. *Demography*, *56*(4), 1195–1218. https://doi.org/10.1007/s13524019-00796-0.

Methodology—arab barometer (2021). https://www.arabbarometer.org/survey-data/methodology/

Methodology—world values survey (2021). https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

Nugent, E., Masoud, T., & Jamal, A. (2018). Arab responses to western hegemony: experimental evidence from Egypt. *Journal of Conflict Resolution*, *62*(2), 254–288. https://doi.org/10.1177/0022002716648738.

Nunn, N., Qian, N., & Wen, J. (2018). *Distrust and political turnover during economic crises*. National Bureau of

Economic Research. https://www.nber.org/papers/w24187

Ortmanns, V., & Schneider, S.L. (2016). Harmonization still failing? Inconsistency of education variables in crossnational public opinion surveys. *International Journal of Public Opinion Research*, *28*(4), 562–582. https://doi.org/10.1093/ijpor/edv025.

Owen, R. (2013). *State, power and politics in the making of the modern middle east*. Routledge.

Price, A. (2015). How national structures shape attitudes toward women's right to employment in The Middle East. *International Journal of Comparative Sociology*, *56*(6), 408–432. https://doi.org/10.1177/0020715215625494.

Rizzo, H., Abdel-Latif, A.-H., & Meyer, K. (2007). The relationship between gender equality and democracy: a comparison of Arab versus non-Arab muslim societies. *Sociology*, *41*(6), 1151–1170.

Ross, M.L. (2008). Oil, islam, and women. *American political science review*, *102*(1), 107–123. https://doi.org/10.1017/S0003055408080040.

Sarracino, F., & Mikucka, M. (2017). Bias and efficiency loss in regression estimates due to duplicated observations: a monte carlo simulation. *Survey Research Methods*, *11*(1), 17–44. https://doi.org/10.18148/srm/2017.v11i1.7149.

Schneider, S. (2009). *Confusing credentials: The cross-nationally comparable measurement of educational attainment*. https://doi.org/10.13140/RG.2.2.35997.51683. Doctoral dissertation

Sika, N. (2020). Contentious activism and political trust in non-democratic regimes: Evidence from the MENA. *Democratization*, *27*(8), 1515–1532. https://doi.org/10.1080/13510347.2020.1813113.

Slomczynski, K.M., Powalko, P., & Krauze, T. (2017). Nonunique records in international survey projects: the need for extending data quality control. *Survey Research Methods*, *11*(1), 1–16. https://doi.org/10.18148/srm/2017.v11i1.6557.

Spierings, N. (2019). Social trust in the middle east and north africa: the context-dependent impact of citizens' socio-economic and religious characteristics. *European Sociological Review*, *35*(6), 894–911. https://doi.org/10.1093/esr/jcz038.

Tessler, M., & Tout, H. (2017). Religion, trust, and other determinants of muslim attitudes toward gender equality: Evidence and insights from fifty-four surveys in the Middle East and North Africa. *Taiwan Journal of Democracy*, *13*(2), 1–29.

Tessler, M., Palmer, M., Farah, T., & Ibrahim, B. (2019). *The evaluation and application of survey research in the Arab world*. Routledge.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859.

Uslaner, E.M. (2008). Where you stand depends upon where your grandparents sat: the inheritability of generalized trust. *Public Opinion Quarterly*, *72*(4), 725–740. https://doi.org/10.1093/poq/nfn058.

Vandenplas, C., & Lipps, O. (2014). *Robustness of items within and across surveys*. FORS working paper series.