A study of stereotype accuracy in the Netherlands: immigrant crime, occupational sex distribution, and provincial income inequality

- Emil O. W. Kirkegaard, Ulster Institute for Social Research, United Kingdom, emil@emilkirkegaard.dk
- Arjen Gerritsen, independent researcher, The Netherlands, arjengerritsen@icloud.com

Abstract

In this pre-registered study, we gathered two online samples totaling 615 subjects. The first sample was nationally representative with regards to age, sex and education, the second was an online convenience sample with mostly younger people. We measured intelligence (vocabulary and science knowledge, 20 items each) using newly constructed Dutch language tests. We measured stereotypes in three domains: 68 national origin-based immigrant crime rates, 54 occupational sex distributions, and 12 provincial incomes. We additionally measured other covariates such as employment status and political voting behaviors.

Results showed substantial stereotype accuracy for each domain. Aggregate (average) stereotype Pearson correlation accuracies were strong: immigrant crime .65, occupations .94, and provincial incomes .85. Results of individual accuracies found there was a weak general factor of stereotype accuracy measures, reflecting a general social perception ability. We found that intelligence moderately but robustly predicted more accurate stereotypes across domains as well as general stereotyping ability (r's .20, .25, .26, .39, β's 0.17, 0.25, 0.21, 0.37 from the full regression models). Other variables did not have robust effects across all domains, but had some reliable effects for one or two domains.

For immigrant crime rates, we also measured the immigration preferences for the same groups, i.e. whether people would like more or fewer people from these groups. We find that actual crime rates predict net opposition at r = .55, i.e., subjects were more hostile to immigration from origins that had higher crime rates. We examined a rational immigration preference path model where actual crime rates→stereotypes of crime rates→immigrant preferences. We found that about 84% of the effect of crime rates was mediated this way, and this result was obtained whether or not one included Muslim% as a covariate in the model. Overall, our results support rational models of social perception and policy preferences for immigration.

Keywords: stereotype accuracy, immigrants, immigration, Muslim, Islam, sex, gender, provinces, inequality, the Netherlands, intelligence, cognitive ability, preregistered, vocabulary, science knowledge

Introduction

Stereotypes have a poor reputation in science. They are frequently labeled inaccurate, exaggerated, sexist, racist, and harmful (Jussim, 2012; Jussim et al., 2018). Theories of the harmfulness of stereotypes generally involve claims of causality from social perception to social reality. For instance, the very popular model of stereotype threat involves stress from negative stereotypes being activated to actual performance or behavior in line with the stereotype (Shewach et al., 2019). On the other hand, there are some researchers who pursue a more rationalist approach to stereotypes wherein these are mostly accurate approximations of group differences that ordinary humans form based on their observations of other humans, reports in the media, government statistics and so on. From this perspective, social reality is the main cause of stereotypes, not the other way around. In recent years, social and biomedical science has suffered from the revelations of the replication crisis. It has been found that probably most published research is not replicable when other researchers, or sometimes even the same researchers, try to re-do a given experiment or study (Gordon et al., 2020; Kvarven et al., 2020). These empirical findings of replication failure are mostly congruent with the original findings being false positives, i.e., they reported some association in reality that didn't exist, or at least, is so weak as to be indistinguishable from noise unless one has a quite large sample size (e.g. n > 1,000). Like most other research in social psychology, the popular stereotype threat theory has not fared well. In particular, meta-analyses show that larger studies find weaker or no such effects, thus suggesting publication bias causing a misleading picture of the evidentiary status (Shewach et al., 2019). Furthermore, at least one large (n = 2,064 Dutch high school students) pre-registered replication failed to find any evidence of stereotype threat for female and math ability (P. Flore, 2018; P. C. Flore et al., 2018). On the other hand, stereotype accuracy findings have replicated extremely well with several recent, large studies finding approximately the same results as earlier small-scale research had reported (Jussim et al., 2018). Thus, publication bias did not seem to affect this literature, in line with Sesardic' conjecture.[1]

[1] Sesardic's conjecture is that because research that is not friendly to left-wing ideology faces unscientific discrimiantion in academia, the published research that nonetheless make it through the filters (ethics approval, grant applications, editorial and peer review), is consequently of higher than average scientific rigor. In his case, he was thinking of research on genetics (behavioral genetics), but the conjecture is more general (Kirkegaard, 2020a; Sesardić, 2005).

There are some limitations on the existing research on stereotype accuracy, however. First, most research has been done with North American students, the convenience sample of choice for most social science (Henrich et al., 2010). Thus there is a general need to see if findings replicate with samples from other parts of the world, especially those which are more representative. Second, as far as the authors are aware, all prior research has involved only a single domain of estimates. Thus, studies may have asked about stereotypes of ethno-racial groups, sexes, immigrant groups, or age groups, but none have measured stereotype accuracy for multiple groups at a time. For this reason, it is unknown how accuracy or bias in one domain relates to accuracy in another domain. Are people with more accurate stereotypes about sex differences also more accurate about age differences? Do people who exaggerate group differences in one domain also do so in others? Third, there are so far no known strong predictors of stereotype accuracy, e.g., with r > .30. Prior literature has reported positive results mainly for intelligence, educational attainment, male sex (sometimes), older age (sometimes), and some policy preferences. However, these factors have so far explained a quite meagre perfect of the observed variance, always below 10%. Thus, science does not know much about why some people apparently have accurate social perceptions for some domain (or in general) while others don't. The current study sought to partially remedy these shortcomings by studying multiple domains of social perception at once, measuring more potential predictors of individual variation in stereotype accuracy, and employing large, Dutch samples that are more representative than the typical student samples. Furthermore, the study's analyses were pre-registered to a large extent, thus giving confidence that the results were not cherry picked among possible method variations.[2]

Data

We sought to sample 500 Dutch citizens using the Prolific survey company (https://www.prolific.co/; (Palan & Schitter, 2018)). Unfortunately, they did not have sufficient subjects to take part in our research as we had planned, and we ended up with 411 subjects with valid data. For this reason we used an additional company, Survee (https://www.survee.dk/), to sample an additional 200 subjects (204 valid subjects obtained; we decided on 200 to have sufficient data to compare the two data sources). We used the same survey for both recruitment services, except that we had to insert extra questions into the survey for Survee because this service did not provide the same metadata automatically as Prolific does (data about employment status and so on). These questions were inserted for consistency, so that we had full coverage of all variables for both data sources. In total, we collected data for 685 subjects, 60 of which failed our attention checks and were excluded from further analysis, leaving us with 615 valid subjects. See the appendix for details on the attention checks.

[2] Preregistration document https://osf.io/8qhmr/.

Our questionnaire took a median of 22.9 minutes to complete[3] and contained 64-67 questions (3 extra questions for the Survee version). The survey structure was as follows:

1. Description and consent
2. Confirmation of Dutch citizenship
3. Extra questions for Survee if relevant
4. Education
5. Family background
6. Political party support (voting)
7. Vocabulary test (20 questions)
8. Science knowledge test (20 questions)
9. Sex distribution by occupation (54 units)
10. Immigrant crime rates (68 units)
11. Provincial incomes (12 units)
12. Final questions and comments
13. Thank you page

For measuring intelligence, we decided to measure two aspects: vocabulary and science knowledge. These are both crystalized (i.e. accumulated) aspects of intelligence. The reason for the choice of these is that they are faster to measure and have high loadings on the general factor, making for a more reliable test (Kan et al., 2013). The items in these tests were designed for this survey. The vocabulary test was designed after the English test at https://openpsychometrics.org/tests/VIQT/, (Kirkegaard, in review). The design is a select-2-from-5 approach. Each item is a list of 5 words and the subject is asked to pick the two synonyms. The science test concerned knowledge questions about various areas of science in typical multiple choice format (choose 1 from 6-10 options).[4] We sought to maximize the number of distractors (false response options) since this reduces the chance of blindly guessing correctly, and thus should increase the factor loading of the item. The tests are both in Dutch and are freely available to anyone in the supplementary materials for any purpose with no prior permission (public domain). The appendix gives examples of items. The resulting data were analyzed using item response theory. We tried different scoring methods, as set out in the pre-analysis plan. Our primary measure was the single factor item response theory (IRT) model. We additionally scored a 3-factor model, with a general factor and a group factor for each test (vocabulary and science knowledge). We were particularly interested in examining ability tilt effects (Coyle, 2018; Kirkegaard, 2020b), so we wanted an orthogonal tilt factor to include in our regressions. The scores from the 3-factor IRT model did not produce this result, so we tried 3 other methods: creating a tilt score by 1) subtracting the z-scored sum of science items from the z-scored sum of verbal items, 2) subtracting the

[3] The median is preferable here because some subjects leave the tab with the survey open for hours or even days, causing a large tail. The mean was 2,305 minutes (38.4 hours) with a standard deviation of 23,155. The median absolute deviation was 8.6 minutes.

[4] These were obtained from our current pool of about 250 science knowledge questions that are under development. There were 6 biology, 4 math/statistics, 2 economics, 1 history, 2 psychology/psychiatry, 2 linguistics, 2 physics, and 1 geography questions.

science knowledge IRT score from the verbal IRT score, both from the 3-factor model, and 3) computing the non-g residuals of the sum scores of the verbal and science scores, and then subtracting the science score from the verbal score. All of these attempt to quantify the notion of doing relatively better on the verbal part as compared to the science part, while ignoring the overall level of intelligence. Analyzing these scores, we find that the last approach produces appropriately orthogonal scores to g, and we used this for our analysis (V tilt).

For the stereotype measurement, we were especially interested in immigrant stereotypes because of prior research on the topic and the general political relevance (Kirkegaard et al., 2020). We sought two other domains and picked sex differences in employment sections/jobs, as well as provincial income differences. These were selected so as to be as different from each other as possible. For the immigrant crime data, we picked 68 origin countries from a recent study of immigrant crime rates (Kirkegaard & de Kuijper, 2020). These estimates should be highly reliable, as they are based on public data published by the government, and thus suitable as criterion data (Jussim, 2012). The numbers specifically concern the arrest rates for the groups. For the occupation data, we obtained a list of 54 occupations from CBS (Centraal Bureau voor de Statistiek; Statistics Netherlands). For the provincial data, we obtained their average (mean) disposable income (corrected for household size - excl. Student households) for the 12 provinces of the Netherlands.[5] For immigrant preferences data, we reused data from a prior study that concerned the same 68 origins (Kirkegaard & de Kuijper, 2020). The subjects in this prior dataset overlapped with the current ones to some extent, but a prior study found that sample overlap between subjects asked about preference and crime stereotypes did not affect results (Kirkegaard et al., 2020). The specific questions were, with English translations:

- Geef met behulp van de slider aan voor hoeveel procent u denkt dat de volgende beroepen door mannen uitgevoerd worden.
    - Please use the slider to indicate the percentage of men performing this profession.
- Nederland kent vele groepen immigranten. Schuif voor elk land van herkomst de slider naar uw schatting van het misdaadniveau van de immigranten uit dit land die in Nederland leven. 1 is het misdaadniveau van autochtone Nederlanders. Dat betekent bijvoorbeeld dat u de slider op twee (2) zet als u denkt dat het misdaadniveau van deze immigranten twee keer zo hoog is als dat van autochtone Nederlanders.
    - There are many different immigrant groups in the Netherlands. For each of the groups, adjust the slider to your estimation of the crime rate relative to Dutch natives. This means you should adjust the slider to two (2) if you think the crime rate of this group is twice that of natives.
- Besteedbaar inkomen [followed by a slider for each region]
    - Disposable income

[5] https://en.wikipedia.org/wiki/Provinces of the Netherlands

Stereotype scoring methods

Because the scoring methods of stereotypes are not well known, we provide examples here for illustrative purposes. Suppose we gathered data from 3 raters (A, B, C) who rated 3 target groups (X, Y, Z) on some trait. Table X shows example computations of accuracy and bias metrics.

	Estimates			**Deviation**			
Rater	X	Y	Z	X	Y	Z	
A	15	2	25	5	-18	-5	
B	20	12	40	10	-8	10	
C	15	16	17	5	-4	-13	
Mean estimate	16.7	10.0	27.3	6.7	-10.0	-2.7	
Criterion values	10	20	30				
	Statistics						
Rater	**Pearson r**	**Spearman r**	**Mean**	**SD**	**Mean error**	**SD error**	**MAD**
A	0.43	0.50	14.0	11.5	-6.0	1.5	9.3
B	0.69	0.50	24.0	14.4	4.0	4.4	9.3
C	1.00	1.00	16.0	1.0	-4.0	-9.0	7.3
Mean estimate	0.61	0.50	18.0	8.7	-2.0	-1.3	6.4
Criterion values			20	10			

Table X. Calculation example for stereotype accuracy metrics. SD = standard deviation, MAD = mean absolute deviation. Deviation = difference from criterion value.

The table shows both the raw estimates and the computed metrics for accuracy and biases. The real group means are 10, 20, and 30 on some hypothetical trait. Each rater had some level of accuracy for these differences, shown in their correlational accuracies which are all positive, ranging from .43 to 1.00. The difference between the Pearson and Spearman values is that for Spearman, only the order matters, whereas for Pearson, the relative differences matter, though not the scale. The scale consists of the mean and standard deviation of the ratings. Thus, we see that rater C has perfect accuracy in correlational terms (1.00), but actually his scale is widely off the mark in terms of both

central tendency and dispersion. Central tendency and dispersion can be operationalized in different ways, but here we used the (arithmetic) mean and the SD, the most common metrics. C's (implied) estimate of the SD is 1 whereas the true SD is 10, thus earning him an SD error of -9 (i.e., he was 9 too low). Similarly, the raters differ in their estimate of the mean, ranging from 14.0 to 24.0, whereas the true value is 20. Thus, they suffer from mean bias but in different directions, ranging from -6.0 to 4.0. Finally, one can calculate the mean absolute deviation (MAD), which is the all-inclusive measure of accuracy. The various aspects of (in)accuracy may or may not covary. It depends on the structure of errors. If everybody perceives the same signal and is affected by different amounts of random noise, the metrics will tend to be positively related when adjusted for direction (i.e., lower MAD is better, but higher correlations are better). However, if the errors are a mix of many influences that differ between people in complex ways, the various metrics may not relate much or could even show opposite relations. For instance, correlation accuracies might be larger for people who overestimate group differences (i.e., have positive SD errors). The appendix provides a set of results from simulated data that illustrate these points.

In the above case, the scale of the estimates and the criterion values is the same. However, if the scale is not the same, then some of these values cannot be used. It does not make sense to compute deviation scores when they are on different scales, and thus, the metrics derived from these also make no sense. Thus, in practice, when scales are not the same, one is limited to using correlational metrics. In practice, these are arguably the most important metrics as well because for decision making, it is mainly the relative differences between groups that matter, less so whether one gets the scale wrong. However, since there are persistent claims that stereotypes exaggerate real differences, one will have to acquire data on the same scale to compute the dispersion error metrics to examine these claims. Central tendency errors are sometimes of interest. If one was estimating crime rates by group, then tendencies to over- or underestimate crime rates in general may be of interest since this would presumably relate to people's preferred policies in the criminal justice system (e.g., if one overestimates/underestimates the prevalence of crime, then one might support more/less funding for policing than is necessary).

The scoring of aggregate stereotypes works the same way as the individual level, except one first aggregates the individual ratings. In the above, the mean was used. Usually, the average estimates will be more accurate than the average of the individual accuracies. However, this is not necessarily the case, and in fact, not the case in the above for the correlational metrics (since C had perfect scores).

Results

There are results of interest at two levels of analysis: individual level (personal) and aggregate level (consensual) stereotypes. At the individual level, accuracy is generally weaker, and varies among people. This allows one to examine associations between accuracy (and bias) metrics and other individual variables, as well as examine the

distributions of the variables. At the aggregate level, typically the arithmetic mean is used to aggregate the individual stereotypes to a single vector of values, which one then relates to other variables of interest. In this study, we used both approaches, and so the results are necessarily split between these levels of analysis. Because of our use of three domains, the results are complicated to present, and we have split them by domain. The sections on occupations and provinces are less detailed since these were not the primary focus of the study.

To decrease false positive rates, we used a more conservative p-value threshold of 0.01. Due to the number of tests done across the full study, this level should still be taken as suggestive of an association pending replication.

Individual level results

Table X provides descriptive statistics for the main non-stereotype variables in the study.

Variable	Group	N	Mean	Median	SD	MAD	Min	Max	Skew
g	combined	615	0.00	0.00	1.00	0.99	-2.74	3.34	0.11
g	Prolific	411	0.14	0.13	0.97	0.96	-2.20	3.34	0.19
g	Survee	204	-0.28	-0.33	1.01	1.05	-2.74	2.71	0.06
V tilt	combined	607	0.00	-0.01	1.00	1.02	-4.64	2.78	-0.25
V tilt	Prolific	403	-0.21	-0.17	0.98	1.04	-4.64	2.43	-0.27
V tilt	Survee	204	0.41	0.44	0.92	0.94	-2.33	2.78	-0.18
age	combined	607	33.24	29.00	12.84	10.38	18.00	74.00	0.95
age	Prolific	403	28.74	26.00	9.49	7.41	18.00	74.00	1.85
age	Survee	204	42.12	44.00	13.95	16.31	18.00	65.00	-0.15
male	combined	612	0.55	1.00	0.50	0.00	0.00	1.00	-0.18
male	Prolific	408	0.58	1.00	0.49	0.00	0.00	1.00	-0.34
male	Survee	204	0.47	0.00	0.50	0.00	0.00	1.00	0.12
time taken	combined	615	25.53	22.92	11.05	8.60	6.00	54.71	1.15
time taken	Prolific	411	26.02	23.47	10.37	8.56	7.73	54.71	1.17
time taken	Survee	204	24.54	20.83	12.26	8.91	6.00	54.71	1.18
education	combined	615	4.53	4.00	0.95	1.48	1.00	6.00	-0.22
education	Prolific	411	4.66	5.00	0.92	1.48	1.00	6.00	-0.32
education	Survee	204	4.27	4.00	0.95	1.48	2.00	6.00	-0.02

student	combined	610	0.38	0.00	0.49	0.00	0.00	1.00	0.51
student	Prolific	406	0.48	0.00	0.50	0.00	0.00	1.00	0.09
student	Survee	204	0.18	0.00	0.38	0.00	0.00	1.00	1.68
vote PvdD	combined	615	0.05	0.00	0.19	0.00	0.00	1.00	4.16
vote PvdD	Prolific	411	0.06	0.00	0.22	0.00	0.00	1.00	3.66
vote PvdD	Survee	204	0.02	0.00	0.14	0.00	0.00	1.00	5.96
vote Groenlinks	combined	615	0.19	0.00	0.36	0.00	0.00	1.00	1.54
vote Groenlinks	Prolific	411	0.24	0.00	0.40	0.00	0.00	1.00	1.17
vote Groenlinks	Survee	204	0.09	0.00	0.25	0.00	0.00	1.00	2.84
vote SP	combined	615	0.06	0.00	0.22	0.00	0.00	1.00	3.49
vote SP	Prolific	411	0.04	0.00	0.17	0.00	0.00	1.00	4.51
vote SP	Survee	204	0.11	0.00	0.30	0.00	0.00	1.00	2.41
vote D66	combined	615	0.11	0.00	0.26	0.00	0.00	1.00	2.41
vote D66	Prolific	411	0.13	0.00	0.28	0.00	0.00	1.00	2.07
vote D66	Survee	204	0.06	0.00	0.19	0.00	0.00	1.00	3.47
vote PvDA	combined	615	0.06	0.00	0.22	0.00	0.00	1.00	3.56
vote PvDA	Prolific	411	0.05	0.00	0.20	0.00	0.00	1.00	3.94
vote PvDA	Survee	204	0.08	0.00	0.26	0.00	0.00	1.00	2.97
vote VVD	combined	615	0.13	0.00	0.32	0.00	0.00	1.00	2.12
vote VVD	Prolific	411	0.09	0.00	0.26	0.00	0.00	1.00	2.74
vote VVD	Survee	204	0.22	0.00	0.39	0.00	0.00	1.00	1.35
vote Christenunie	combined	615	0.03	0.00	0.16	0.00	0.00	1.00	5.52
vote Christenunie	Prolific	411	0.03	0.00	0.14	0.00	0.00	1.00	5.66
vote Christenunie	Survee	204	0.03	0.00	0.18	0.00	0.00	1.00	5.08
vote PVV	combined	615	0.06	0.00	0.21	0.00	0.00	1.00	3.80

vote PVV	Prolific	411	0.03	0.00	0.15	0.00	0.00	1.00	5.58
vote PVV	Survee	204	0.11	0.00	0.29	0.00	0.00	1.00	2.41
vote CDA	combined	615	0.02	0.00	0.14	0.00	0.00	1.00	6.06
vote CDA	Prolific	411	0.01	0.00	0.09	0.00	0.00	1.00	7.81
vote CDA	Survee	204	0.05	0.00	0.20	0.00	0.00	1.00	4.26
vote FvD	combined	615	0.06	0.00	0.21	0.00	0.00	1.00	3.59
vote FvD	Prolific	411	0.06	0.00	0.20	0.00	0.00	1.00	3.76
vote FvD	Survee	204	0.07	0.00	0.22	0.00	0.00	1.00	3.28
vote SGP	combined	615	0.00	0.00	0.05	0.00	0.00	1.00	19.84
vote SGP	Prolific	411	0.00	0.00	0.02	0.00	0.00	0.50	20.13
vote SGP	Survee	204	0.00	0.00	0.07	0.00	0.00	1.00	14.07
vote 50Plus	combined	615	0.01	0.00	0.07	0.00	0.00	1.00	10.95
vote 50Plus	Prolific	411	0.00	0.00	0.03	0.00	0.00	0.50	14.18
vote 50Plus	Survee	204	0.02	0.00	0.12	0.00	0.00	1.00	7.18
vote DENK	combined	615	0.00	0.00	0.02	0.00	0.00	0.50	24.68
vote DENK	Prolific	411	0.00	0.00	0.02	0.00	0.00	0.50	20.13
vote DENK	Survee	204	0.00	0.00	0.00	0.00	0.00	0.00	NA

Table X. Descriptive statistics for the main numerical variables (not including the stereotype metrics). Vote variables refer to the average votes for that party in the last election and hypothetical election today (i.e., values can be 0, .5, and 1).

The Survee dataset was intended to be nationally representative, and the descriptive statistics bear this out. The Prolific group was younger (means 29 vs. 42, average age for the country is 42), more male (58% vs. 47%, average for the country is about 50%), much more likely to be students (48% vs. 18%), somewhat smarter (0.43 d), and much more likely to vote for left-wing parties (e.g., Green-left/Groenlinks 24% vs. 9%, the party received 9.1% of the vote in the 2017 general election, see appendix). Furthermore, Table X provides summary statistics for the voters by party for some variables.

Party	**N**	**g**	**V tilt**	**Male**	**Age**

D66	65	0.46	-0.18	0.61	30.37
Groenlinks	118	0.21	-0.14	0.42	28.94
PvdD	29.5	0.15	0.09	0.34	32.70
SGP	1.5	0.08	-1.17	0.69	38.17
Christenunie	18	0.08	-0.32	0.52	35.50
SP	39.5	0.03	0.58	0.31	42.47
VVD	82.5	0.01	0.17	0.67	35.53
PvDA	38.5	-0.06	0.19	0.61	33.68
FvD	37	-0.09	-0.11	0.75	34.81
CDA	15	-0.31	0.83	0.31	43.09
DENK	0.5	-0.55	-1.23	-0.55	14.80
PVV	34.5	-0.63	0.47	0.60	42.00
50Plus	4.5	-0.91	-0.36	0.10	43.11

Table X. Summary statistics for political party voters. N refers to the complete persons count, so a person who voted for the party in 2017, but would not vote for today counts as 0.5 (and vice versa).

Values for variables were calculated by extrapolating to the expected value at complete support for the party (voted for the party in 2017 and would vote for today). We see that in terms of average intelligence, the social-liberal parties had the highest levels. This confirms previous research using both scales to measure social liberalness and party votes (Carl, 2014, 2015; Deary et al., 2008; Kirkegaard et al., 2017). The ability tilt seems to be mostly related to the average age of the voters, unsurprising since these variables are correlated r = .42 in this dataset, that is, older people do relatively better on verbal tests than science knowledge.

Immigrants and crime rates

Immigrant crime rates were the primary domain of interest for the stereotypes in this study. Overall, there was a fair amount of accuracy in the stereotypes. Figure X shows the distributions of four main metrics, and Table X shows summary statistics for the metrics.

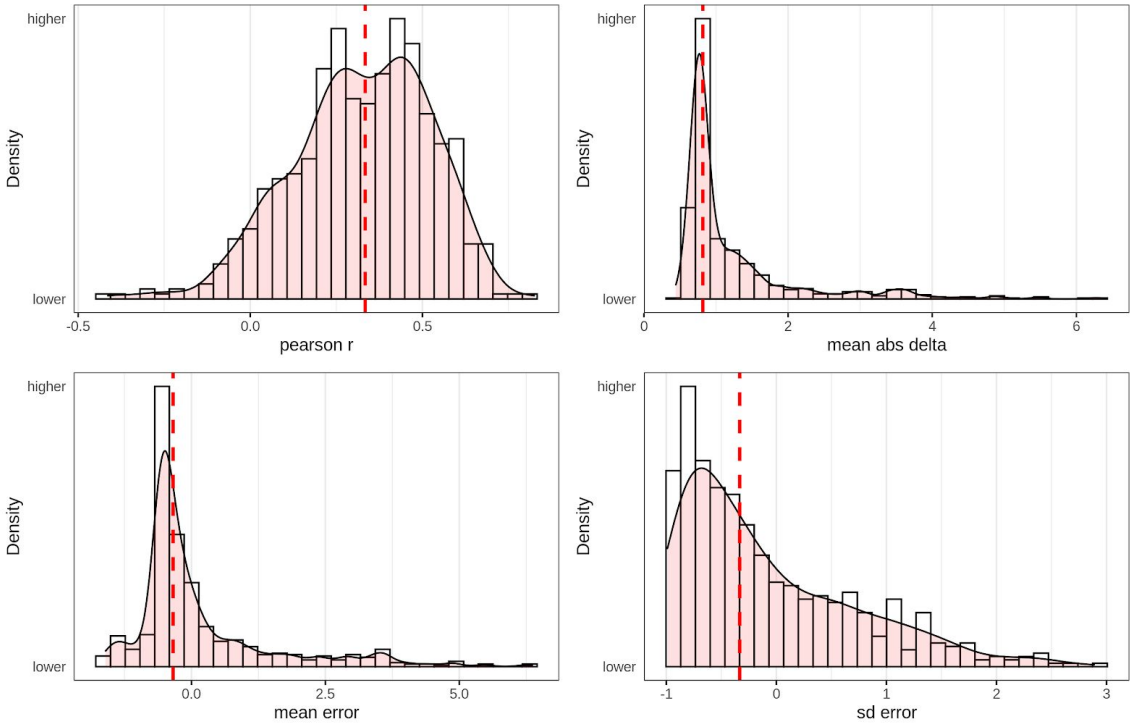


Figure X. Distribution of stereotype accuracy metrics. Vertical line is the median.

Group	N	Mean	Median	SD	MAD	Min	Max	Skew
pearson r	598	0.32	0.33	0.20	0.21	-0.41	0.83	-0.45
rank r	598	0.35	0.37	0.21	0.22	-0.49	0.81	-0.53
mean abs error (MAE)	615	1.23	0.82	0.90	0.19	0.44	6.36	2.56
sd	615	0.92	0.66	0.80	0.67	0.00	3.88	1.11
sd error	615	-0.07	-0.33	0.80	0.67	-0.99	2.89	1.11
sd error abs	615	0.67	0.63	0.45	0.37	0.00	2.89	1.44
mean	615	1.75	1.26	1.32	0.41	0.00	7.96	1.93
mean error	615	0.15	-0.34	1.32	0.41	-1.60	6.36	1.93
mean error abs	615	0.88	0.54	1.00	0.37	0.00	6.36	2.45

Table X. Summary statistics of immigrant crime stereotype metrics. The lower sample sizes for correlations is due to some people assigning the same value to every group, resulting in 0 variance, and thus undefined correlations.

The distribution of correlational accuracy shows the same long tail into negative values as seen in prior studies. It appears that some people provide reverse ratings on purpose, since such large negative values are unlikely to result by chance. Still, this did not affect the mean much, since in this case the mean and medians are nearly identical (.33 and .32). In terms of elevation error, there is disagreement with the mean and median values (0.15 vs. -0.34). This is due to a long term of people who grossly overestimated crime levels relative to natives, but the median person (and the majority of persons) actually underestimated non-native levels of crime involvement. With regards to dispersion error, we see the same pattern general distribution but here even the mean person underestimated true differences in crime rates between groups. We were particularly interested in the association between intelligence and stereotype accuracy. Figure X shows the scatterplots of intelligence, and the stereotype accuracy metrics.

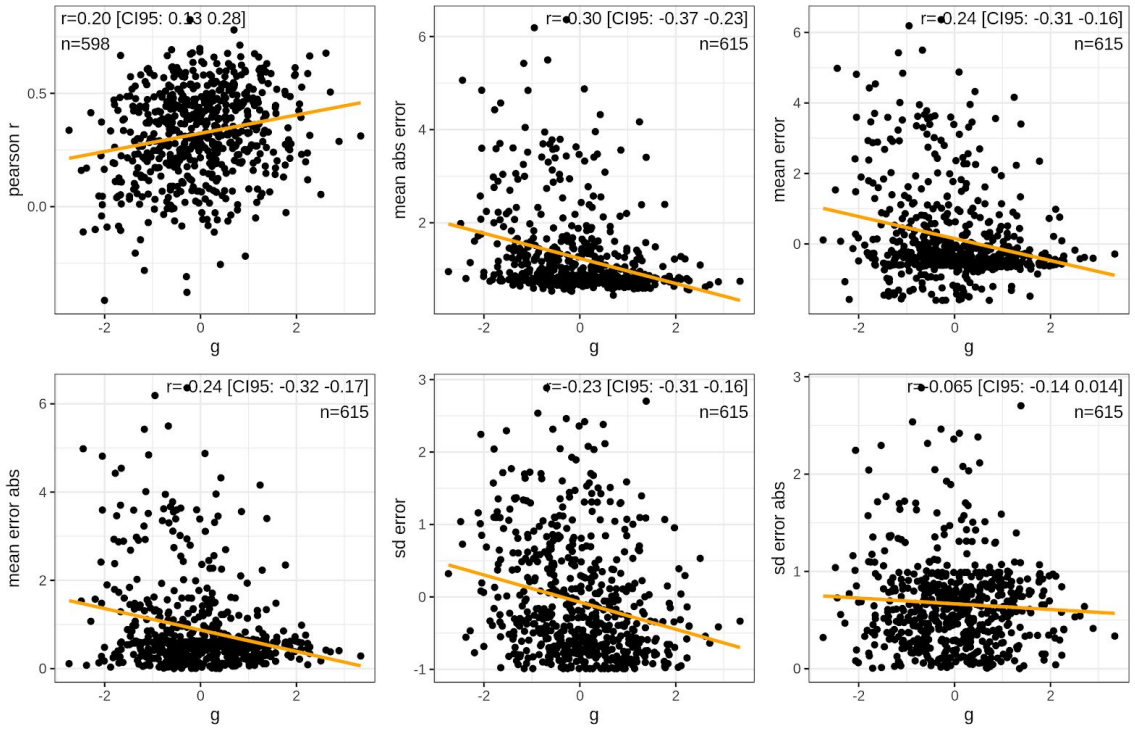


Figure X. Scatterplots of intelligence and the metrics of stereotype accuracy. Abs = absolute, sd = standard deviation.

Each measure tells its own story. For the two main measures of accuracy, we see notable relationships, r = .20 and -.30 (plots 1-2), for correlational accuracy and mean absolute error (lower values better), respectively. In other words, smarter people are better at getting the relative differences right and also better at getting the absolute values right. Looking at the scaling metrics, we see that smarter people tend to underestimate the overall crime rate a bit (plot 3, expected value at high g is negative), but not enough such that they become more inaccurate when we disregard directions (plot 4, expected value of high g is about 0). With regards to dispersion, we see strong evidence that smarter people underestimate real differences (plot 5, negative values for higher g), and this is enough that they aren't more accurate when we disregard direction of error (plot 6, correlation is near zero, though possibly slightly negative). Expanding to

the entire set of predictors, Table X gives the correlations between accuracy metrics and all the quantitative metrics.

Predictor	Pearson r	Mean abs error (MAE)	Mean error	Mean error abs	SD error	SD error abs
g	0.20***	-0.30***	-0.24***	-0.24***	-0.23***	-0.07
V tilt	0.07	0.08	0.10	0.04	0.20***	0.11*
Age	0.12**	0.22***	0.25***	0.20***	0.24***	0.09
Time taken	0.02	-0.04	-0.04	-0.03	0.00	0.00
Education	0.16***	-0.21***	-0.17***	-0.16***	-0.18***	-0.07
Vote PvdD	-0.10	-0.05	-0.07	-0.04	-0.09	0.02
Vote Groenlinks	-0.16***	-0.10*	-0.15***	-0.07	-0.20***	-0.04
Vote SP	0.04	0.10	0.08	0.09	0.08	0.03
Vote D66	0.04	-0.10	-0.10	-0.07	-0.18***	-0.05
Vote PvDA	0.00	0.00	0.02	-0.01	0.01	-0.06
Vote VVD	0.01	0.09	0.12**	0.09	0.09	-0.02
Vote Christenunie	-0.01	-0.06	-0.07	-0.04	-0.10	-0.04
Vote PVV	0.03	0.19***	0.18***	0.14***	0.30***	0.19***
Vote CDA	0.03	0.02	0.06	0.02	0.06	0.01
Vote FvD	0.13**	0.03	0.07	-0.01	0.17***	0.02
Vote SGP	0.01	0.08	0.08	0.09	0.05	0.06
Vote 50Plus	0.00	0.04	0.06	0.04	0.09	0.02
Vote DENK	-0.04	0.02	0.03	0.01	0.05	0.02

Table X. Correlations for immigrant crime stereotype accuracy metrics and quantitative predictors. * = p < .01, ** = p < .005, *** = p < .001.

The table reveals a number of findings. First, it can be noted that there are sometimes opposite findings for Pearson r and MAE, for instance, age shows a weak positive relationship, r = .11, while MAE shows a stronger relationship, r = .22 (negative is better). Hence, it appears to be the case that older persons exaggerate differences (positive SD errors), and while this increases their pearson r a bit (which ignores scaling errors), it results in overall worse accuracy (by MAE, which takes into account everything). We also see that verbal tilt is related positively to SD errors, and similarly to overall accuracy

metrics, but not to the mean errors. Education is interesting in that the pattern is very similar to that for intelligence, yet comparatively weaker. The two major players in terms of immigration stances are the PVV (Party for Freedom, the party of nationalist Geert Wilders), and Groenlinks (green left party). PVV voters tend to exaggerate both overall rates of crime and the differences between groups (positive errors for mean and SD, including absolute variants), while green party voters tend to underestimate the same, but not so much it causes worse accuracy when direction is ignored (absolute versions are negative, i.e., higher accuracy). However, the predictor variables are correlated, so it is not clear which is an effect of which, or simply due to confounding. Table X shows the main regression results from OLS (ordinary least squares).

Predictor	**Small model**	**Full model**
Intercept	-0.16 (0.060*)	-0.45 (0.365)
g	0.13 (0.045**)	0.17 (0.046***)
Verbal tilt	0.08 (0.045)	0.06 (0.046)
age	0.07 (0.046)	-0.02 (0.054)
male	0.28 (0.082***)	0.23 (0.086*)
education	0.12 (0.043*)	0.15 (0.048**)
time taken	-0.01 (0.041)	-0.02 (0.041)
First Language=Dutch		(ref)
First Language=non-Dutch		-0.09 (0.158)
Birth=Netherlands		(ref)
Birth=Non-Western		-0.08 (0.165)
Birth=Western		0.05 (0.324)
student		-0.35 (0.110**)
employment status		(included)
vote PvdD		-0.68 (0.232**)
vote Groenlinks		-0.53 (0.149***)
vote SP		0.01 (0.207)
vote D66		-0.11 (0.186)
vote PvDA		-0.08 (0.200)

vote VVD		-0.26 (0.162)
vote Christenunie		-0.41 (0.274)
vote PVV		0.03 (0.217)
vote CDA		0.10 (0.307)
vote FvD		0.39 (0.206)
vote SGP		0.17 (0.850)
vote 50Plus		0.15 (0.540)
vote DENK		-1.79 (1.890)
R2 adj.	0.07	0.133
N	581	572

Table X. Main regression results. Voting variables not standardized, other quantitative variables standardized. * = p < .01, ** = p < .005, *** = p < .001.

The regression models clarify some things. Age is no longer a useful predictor (β = 0.00 in full model), the validity seen prior due to association with other predictors included in the model. Most importantly, intelligence was still a notable predictor, and it was about equal with education (β's = 0.17 and 0.15). Males had somewhat better accuracy (β = 0.23 in full model; β denotes the slope for the binary or proportional predictors), as has been found previously. Student status was associated with quite worse accuracy (β = -0.35), which is interesting. Party voting was mostly unrelated to accuracy, except that voters for PvdD and GL (both are left-wing parties) were associated with quite worse accuracy, consistent with bivariate results. The full model above contains a larger number of variables without detectable validity. To see to which degree these could be left out without affecting the model validity, we used lasso regression to find a good subset of the variables. Table X gives the results.

Term	**Beta**
(Intercept)	0.07
g	0.15
Verbal tilt	0.03
male	0.15
education	0.10

student	-0.29
vote PvdD	-0.49
vote Groenlinks	-0.39
vote VVD	-0.11
vote Christenunie	-0.14
vote CDA	0.00
vote FvD	0.34
vote DENK	-0.56
First Language=non-Dutch	-0.02
Born=Non-Western	-0.03
Employment Status: Part Time	0.08
Employment Status: Unemployed and job seeking	0.19

Table X. Lasso results for prediction of immigrant crime stereotype accuracy. Model trained with **tidymodels** via **glmnet**, and tuned with 10-fold cross validation selecting using the RMSE metric. Betas above from best penalty value (Friedman et al., 2017; Kuhn et al., 2020). Further details can be found in the technical output.

The lasso results replicate most of the findings from OLS, namely that intelligence predicts well (β = 0.15), male status (β = 0.15), education also (β = 0.10). Curiously, lasso selects a number of variables considered mostly useless in the OLS results, such as 5 additional voting variables, employment status, and verbal tilt (which were all $p > .01$ originally).

Since writing the analysis plan for our paper, we became aware of another method for variable selection and comparison, which is arguably more suitable here, namely Bayesian model averaging (BMA) (Goenner, 2004; Hinne et al., 2020). Since our analysis plan included the lasso, we could not switch to this alternative approach, so we instead present it here as an additional approach along the lasso. Briefly, BMA works by exploring the model space of possible models. If it is possible to explore all of them, this is done, otherwise, they are sampled at random or explored using an adaptive approach (similar to forward selection). Among the models explored, the best are chosen based on model fit criteria. Among this subset of best models, the summary statistics for betas are summarized by the mean and SD, weighted by the model fit (best fitting models providing the most weight). This approach results in more stable results than picking a

single best model (as lasso does), while still providing more interpretable results. We used the **BMA** package for this using the default settings (Raftery et al., 2020).[6] Results are shown in Table X.

Predictor	**PIP (%)**	**Post mean (beta)**	**Post SD**
Intercept	100.00	0.19	0.08
g	100.00	0.19	0.04
Verbal tilt	0.00	0.00	0.00
age	1.00	0.00	0.01
male	32.80	0.06	0.10
time taken	0.70	0.00	0.00
education	37.10	0.04	0.06
First Language: non-Dutch	0.80	0.00	0.02
Birth: Non-Western	1.80	0.00	0.03
Birth: Western	0.00	0.00	0.00
student	100.00	-0.33	0.09
Employment Status: Full-Time	2.40	0.00	0.03
Employment Status: Not in paid work	0.80	0.00	0.01
Employment Status: Other	0.90	0.00	0.02

[6] We tried 3 alternative R packages and found this to be the best. See https://rpubs.com/EmilOWK/BMA_examples.

Employment Status: Part-Time	1.50	0.00	0.02
Employment Status: Unemployed and job seeking	13.00	0.03	0.09
vote PvdD	87.10	-0.56	0.29
vote Groenlinks	100.00	-0.46	0.12
vote SP	0.00	0.00	0.00
vote D66	0.00	0.00	0.00
vote PvDA	0.00	0.00	0.00
vote VVD	8.30	-0.02	0.07
vote Christenunie	1.10	0.00	0.04
vote PVV	0.00	0.00	0.00
vote CDA	0.00	0.00	0.00
vote FvD	38.80	0.17	0.24
vote SGP	0.00	0.00	0.00
vote 50Plus	0.00	0.00	0.00
vote DENK	0.80	-0.01	0.22

Table X. Results from Bayesian model averaging. PIP = posterior inclusion probability (probability that predictor is included in best models).

The results are similar to the prior. Notable is that 100% of the best models included intelligence as a predictor (β = 0.19, not much different from r = .20 in bivariate results in Table X). In contrast to the full regression results (in Table X), and the lasso results (in Table X), BMA does not think education is as important as intelligence, including it only in 37% of the best models, and assigning it a notably smaller beta (β = 0.04). Student status (β = -0.33) and voting for certain left-wing parties (PvdD and GL) are still strong predictors included in almost all models, while the remaining variables are more

sporadic. An exception to this is the FvD party (Forum for Democracy, a nationalist party headed by Thierry Baudet), which had a moderate positive beta (β = 0.17). However, all the parties with notably betas had high variance in the estimates, so the present study is not large enough to estimate their values.

Muslim bias

Of particular interest was potential Muslim bias in the stereotypes (Kirkegaard & Bjerrekær, 2016). Muslim bias can be thought of conceptually in multiple ways, but all of them involve groups and countries with higher percentage of Muslims having larger errors in some sense. As detailed in the prior study (Kirkegaard et al., 2020), there were 3 metrics to use: 1) muslim error r, 2) muslim standardized error r, and 3) Muslim elevation error. In the first, the deviation from criterion values are computed for each estimated, and these are correlated with Muslim% in the groups. In the second, a regression model is fit, the residuals saved, and then correlated with the Muslim% values. The difference here is that the first approach forces the scaling to be on the true scale, whereas the second does not. The third method involves computing the deviations as in the first, but then computing two weighted means, one with Muslim% and one with 1-Muslim% as weights, and then subtracting the second from the first. Empirically, the first and third metrics have been found to be very highly correlated (r = .96 in prior study), but the latter has the advantage of being given in natural units, whereas the correlation is unitless [-1 to 1]. Table X shows summary statistics for the metrics and Figure X shows the distributions.

Variable	N	Mean	Median	SD	MAD	Min	Max	Skew
Muslim bias r	615	-0.18	-0.27	0.25	0.22	-0.50	0.66	0.97
Muslim bias r abs	615	0.20	0.23	0.18	0.19	-0.45	0.54	-0.84
Muslim bias wmean	615	-0.28	-0.52	0.71	0.42	-1.48	3.73	1.92
Muslim bias wmean abs	615	0.39	0.40	0.46	0.32	-1.32	2.75	0.63
Muslim resid r	598	-0.34	-0.36	0.09	0.09	-0.49	0.08	1.22
Muslim resid r abs	598	-0.03	-0.03	0.06	0.05	-0.27	0.16	0.01

Table X. Summary statistics for Muslim bias metrics. Abs = absolute value (smaller is better). Wmean = weighted mean, r = Pearson correlation, resid = residual, abs = absolute value.

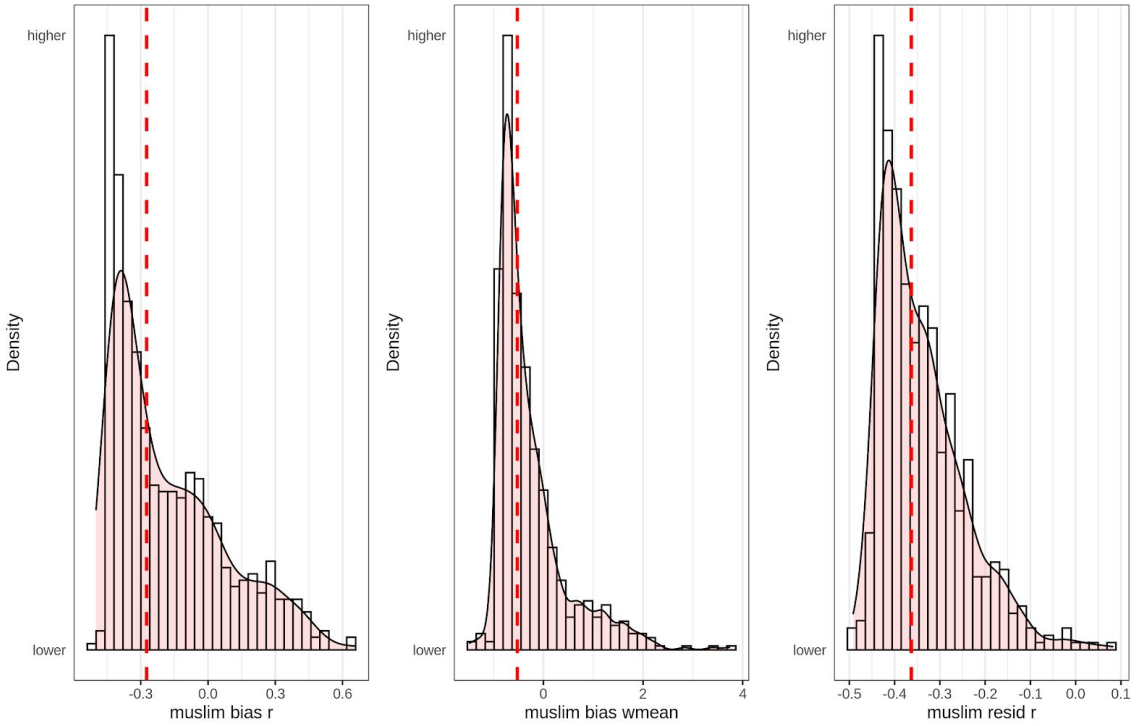


Figure X. Distributions of Muslim bias metrics. Vertical line is the median.

In terms of direction, all the metrics show the same conclusion, namely that subjects had biases in favor of the more Muslim groups in the sense that they tended to underestimate their criminal involvement (negative values mean that more negative errors are associated with Muslim%, i.e., underestimation bias). The weighted mean metric shows the median error is about -0.52, meaning that the median subject underestimated the relative crime rate by 0.52 for more Muslim groups as compared to the less Muslim groups. In the same way, the median correlation between their errors and Muslim% is -.27. Figure X shows the most representative case in the dataset across bias metrics (see the appendix for details of the method).

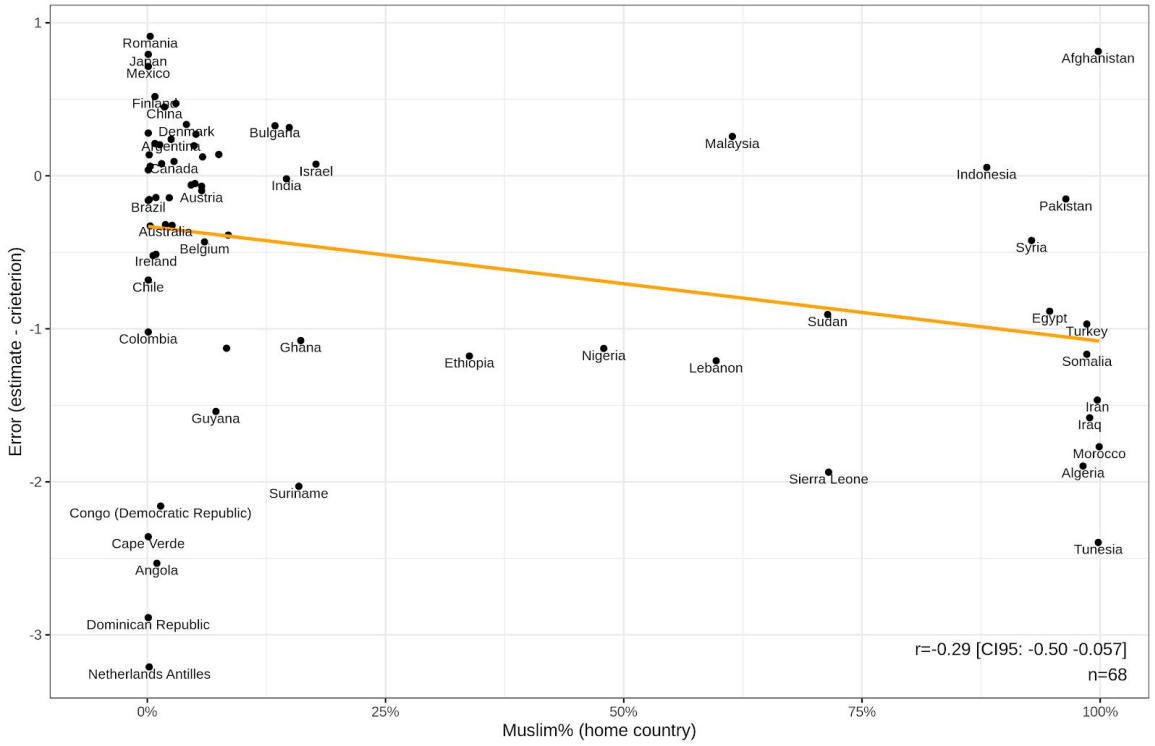


Figure X. Representative case (49) of Muslim bias in estimation.

For this case, we see their correlation is -.29 (compared to the median of -.27). Their estimation errors for the Muslim groups are more negative on average. For instance, Tunesia is estimated to have an average crime rate (RR = 1.1), but actually has a high one (RR = 3.5), thus giving a large negative error (-2.4). Table X shows the correlations among the metrics as well as the primary accuracy metrics.

	Muslim bias r	**Muslim bias r abs**	**Muslim bias wmean**	**Muslim bias wmean abs**	**Muslim resid r**	**Muslim resid r abs**	**Pearson r**	**Mean abs error**	**Mean error**	**Mean error abs**	**SD error**	**SD error abs**
Muslim bias r	1	-0.12**	0.94***	0.30***	0.59***	0.34***	0.37***	0.42***	0.52***	0.30***	0.86***	0.32***
Muslim bias r abs	-0.12**	1	0.08	0.84***	0.07	-0.13**	0.12**	-0.29***	-0.42***	-0.20***	-0.23***	0.41***
Muslim bias wmean	0.94***	0.08	1	0.52***	0.60***	0.31***	0.42***	0.41***	0.48***	0.30***	0.82***	0.51***
Muslim bias wmean abs	0.30***	0.84***	0.52***	1	0.23***	0.01	0.21***	0.02	-0.08	0.04	0.24***	0.66***
Muslim resid r	0.59***	0.07	0.60***	0.23***	1	0.42***	0.68***	-0.05	0.07	-0.06	0.24***	0.11*

Muslim resid r abs	0.34***	-0.13**	0.31***	0.01	0.42***	1	0.17***	0.01	0.06	0.00	0.14***	0.01
Pearson r	0.37***	0.12**	0.42***	0.21***	0.68***	0.17***	1	-0.18***	-0.03	-0.17***	0.19***	0.07
Mean abs error	0.42***	-0.29***	0.41***	0.02	-0.05	0.01	-0.18***	1	0.89***	0.96***	0.61***	0.39***
Mean error	0.52***	-0.42***	0.48***	-0.08	0.07	0.06	-0.03	0.89***	1	0.83***	0.67***	0.32***
Mean error abs	0.30***	-0.20***	0.30***	0.04	-0.06	0.00	-0.17***	0.96***	0.83***	1	0.46***	0.36***
SD error	0.86***	-0.23***	0.82***	0.24***	0.24***	0.14***	0.19***	0.61***	0.67***	0.46***	1	0.49***
SD error abs	0.32***	0.41***	0.51***	0.66***	0.11*	0.01	0.07	0.39***	0.32***	0.36***	0.49***	1

Table X. Correlation matrix for Muslim bias scores and the primary accuracy measures. Wmean = weighted mean, r = Pearson correlation, resid = residual, abs = absolute value. * = p < .01, ** = p < .005, *** = p < .001.

As in the prior study, the error correlation and weighted mean approaches are in near perfect agreement, both in directional and absolute variants (r's .94, and .84), whereas the scale-free residual approach only shows strong correlations with the directional metrics (r's .59 and .60), and near zero with the absolute variants (r's -.13, and .01). As with correlations with accuracy, there are seemingly paradoxical findings. First, there is a positive correlation between Muslim bias r and Pearson r accuracy, r = .37, but also a positive with the mean absolute error, r = .42 (again, smaller is better, so one would expect this to be negative). How is this possible? Here it should be recalled that the metrics are directional and the best value is 0, not 1 (or infinite). As such, greater directional Muslim bias values may be related to correlational accuracy, it depends on where the distribution is located. Figure X shows the scatterplot of the two variables.

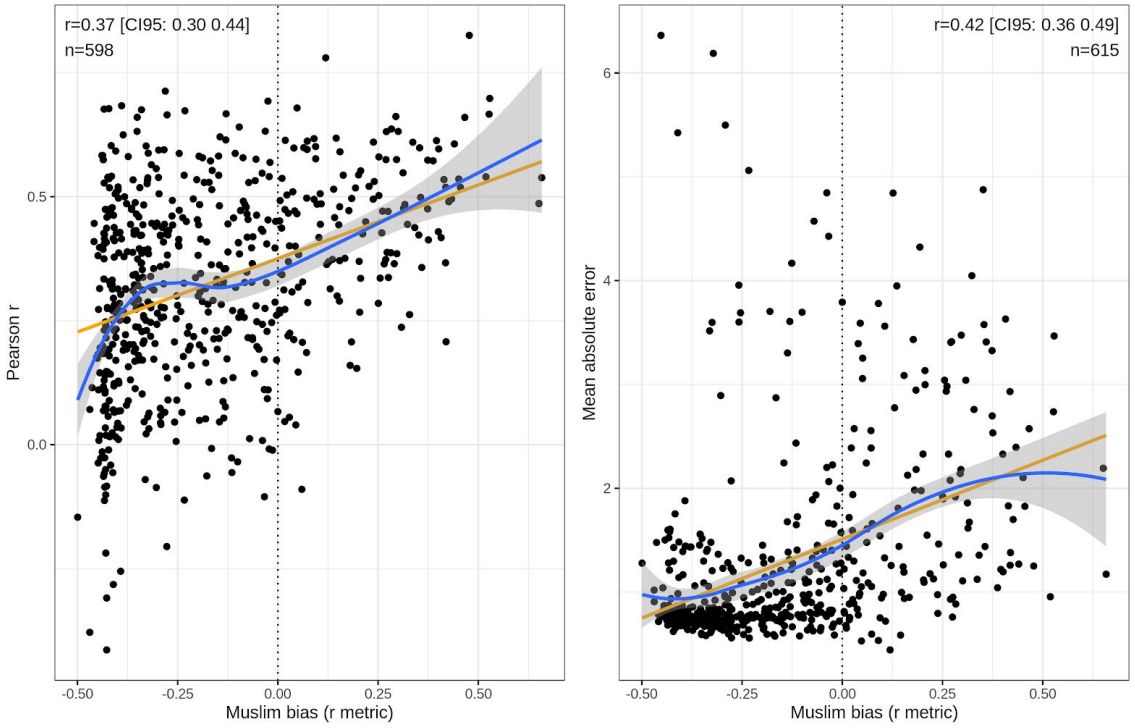


Figure X. Scatterplot for Muslim bias (directional) and stereotype accuracy metrics. Blue line is LOESS fit.

In the left plot, we see that most people are located on the left side (i.e., underestimate Muslim crime rates), and those who overestimate them (on the right side) tend to have higher correlational accuracies, possibly because they exaggerate real differences. To clarify this, one can examine the other primary accuracy metric, in the right plot. Here we see that people with greater errors (in any direction), do show relatively positive Muslim biases. So perhaps those with large Pearson r accuracy attain this by drastically overestimating Muslim crime rates. In fact, the correlation between SD error (i.e., tendency to over or underestimate true differences) and Pearson accuracy is .19, so this appears not to be (much) the case. Table X shows the correlations with the quantitative predictors, and Figure X shows the scatterplots for the relationship to intelligence.

	Muslim bias r	**Muslim bias r abs**	**Muslim bias wmean**	**Muslim bias wmean abs**	**Muslim resid r**	**Muslim resid r abs**
g	-0.20***	0.20***	-0.16***	0.07	0.02	-0.09
Verbal tilt	0.13**	-0.05	0.14***	0.06	-0.03	0.02
Age	0.14***	-0.11*	0.12**	0.00	-0.06	-0.03
Time taken	-0.01	0.06	-0.01	0.04	-0.03	-0.03
Education	-0.14***	0.12**	-0.12**	0.03	0.06	-0.04

Vote PvdD	-0.14***	0.01	-0.13**	-0.03	-0.14***	-0.05
Vote Groenlinks	-0.22***	0.09	-0.20***	-0.01	-0.12**	-0.08
Vote SP	0.08	-0.04	0.08	0.01	0.03	0.05
Vote D66	-0.16***	0.05	-0.14***	-0.04	0.01	-0.04
Vote PvDA	0.02	-0.01	0.01	-0.01	-0.02	0.01
Vote VVD	0.09	-0.09	0.05	-0.05	0.01	0.04
Vote Christenunie	-0.09	0.03	-0.08	-0.02	-0.06	-0.08
Vote PVV	0.23***	-0.03	0.22***	0.10	0.02	0.09
Vote CDA	0.05	-0.10	0.06	-0.02	0.03	0.05
Vote FvD	0.24***	-0.01	0.21***	0.05	0.21***	0.01
Vote SGP	0.06	0.03	0.07	0.07	0.03	0.01
Vote 50Plus	0.11*	-0.04	0.09	0.00	0.03	0.06
Vote DENK	0.01	-0.03	-0.01	-0.02	-0.04	-0.01

Table X. Correlations between Muslim bias metrics and quantitative predictors.

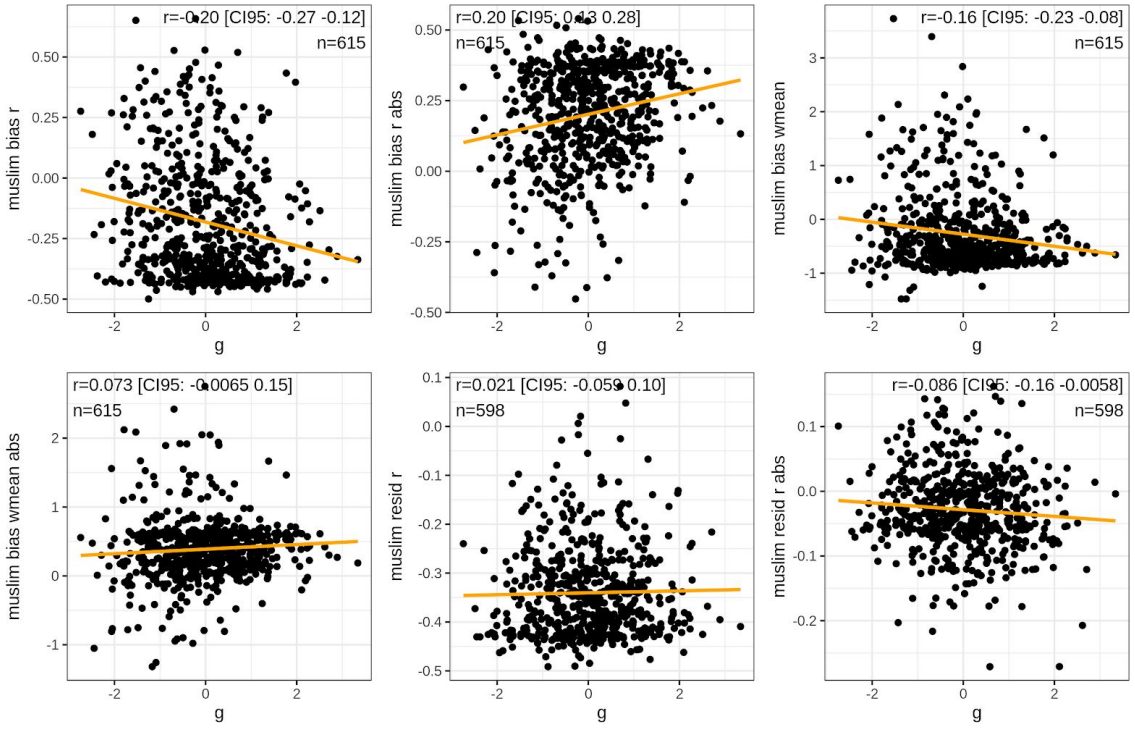


Figure X. Scatterplots between intelligence and Muslim bias metrics.

Starting with the first metric (bias r), we see that smarter people tend to underestimate Muslim crime rates (plot 1, expected values for high g is negative), and this tendency is strong enough that it results in greater undirectional errors (plot 2, expected value for high g is positive). Interestingly, the same pattern is seen for the second metric (wmean, plots 3-4), but much weaker, especially for the absolute variant. The reason for the discrepancy is not obvious as the metrics are correlated quite strongly, and should ideally quantify the same concept. The weaker third metric is only included here for comparison purposes (plots 5-6), and shows only very weak patterns. Turning to the predictors, we see that the left-wing parties tend to have similar bias patterns to the highly intelligent and highly educated, thus illustrating the connection between voting left-wing and being upper class. The nationalist parties show the opposite patterns. Still, when we look at the undirectional errors (bias r abs, and wmean abas), few variables are related. It is only intelligence and education, but again, only to the bias r variant, not the wmean variant. It therefore appears that various kinds of people make approximately equal sized errors with estimation of Muslim groups, but differ in their directions of error along a partisan split of globalist vs. nationalist.

Occupations and sex

Next, we turn to the stereotypes about sex differences in occupation. We have data for 54 occupations, thus a similar number of groups as in the prior section (68 origin groups). As before, we score these for accuracy using a variety of metrics. Distributions of select metrics are shown in Figure X, while Table X shows the summary statistics.

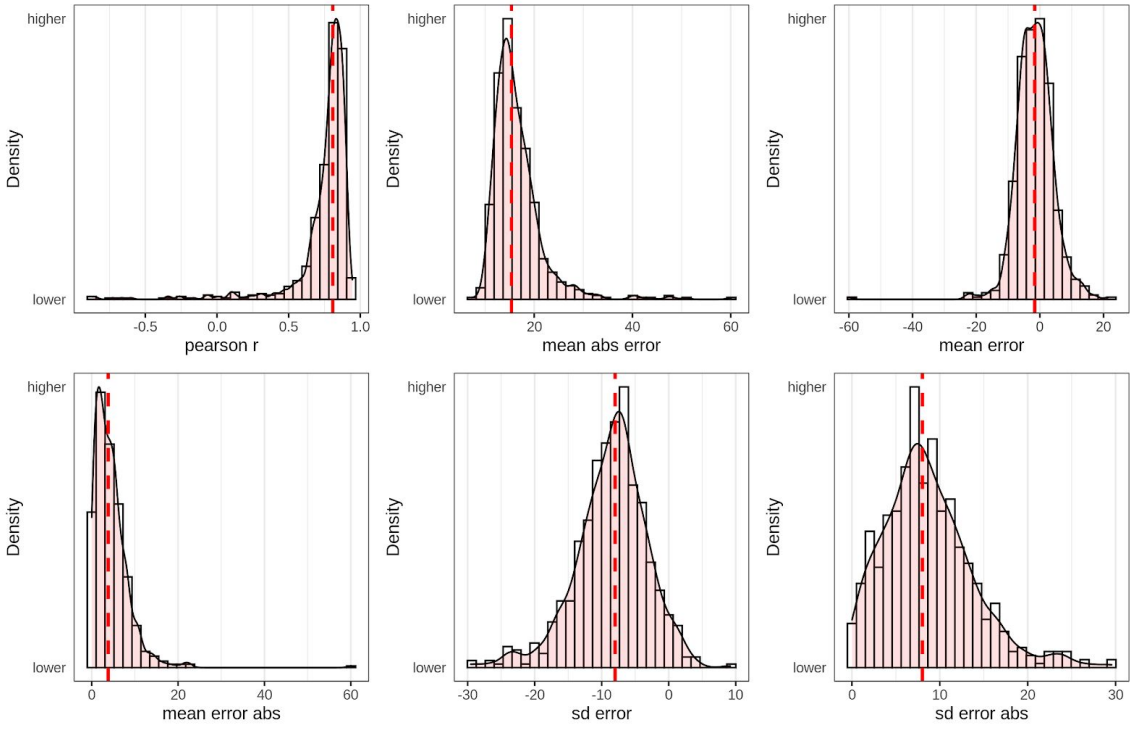


Figure X. Distributions of accuracy metrics for sex differences in occupation.

Variable	N	Mean	Median	SD	MAD	Min	Max	Skew
pearson r	615	0.75	0.81	0.22	0.08	-0.87	0.94	-4.06
rank r	615	0.73	0.79	0.22	0.09	-0.88	0.93	-3.80
mean abs error (MAE)	615	16.57	15.37	5.34	3.49	7.25	60.10	3.02
SD	615	22.02	22.56	5.36	4.69	1.00	39.73	-0.59
SD error	615	-8.53	-7.99	5.36	4.69	-29.54	9.19	-0.59
SD error abs	615	8.69	8.01	5.09	4.66	0.01	29.54	0.88
mean	615	59.97	59.94	6.04	5.16	1.48	83.02	-1.34
mean error	615	-1.61	-1.63	6.04	5.16	-60.10	21.44	-1.34
mean error abs	615	4.55	3.81	4.28	3.20	0.00	60.10	4.48

Table X. Summary statistics for stereotype accuracy metrics for sex differences in occupation.

Overall, we see very high levels of accuracy. The median Pearson r is .75, among the highest value seen for any stereotype study (Jussim, 2018). Unlike the case with crime rates, the rank r is slightly lower than the Pearson r (.73 vs. .75), suggesting that for crime rates, people had some trouble with the scaling, but not the rank orders. For sex differences, this was not the case. The median MAE is 15.37, meaning that the median guess was about 15%points from the right value. Considering that values can span from 0 to 100%, and thus a random guess would lead to a median MAE of about 32 (cf. the simulations in the appendix), this is a quite high level of absolute accuracy. The median mean error was about -1.6, meaning that people slightly underestimated the number of men in an occupation, but it was very close to the true average across the occupations (61.6 vs. median guess of 60.0). For dispersion, there is substantial *underestimation* of sex inequality of the job market, with the median SD being 22.6, while the true value is 30.5, thus a median SD error of -8.00. In relative terms, this is quite large, -26.2% underestimation. This finding flies in the face of repeated claims of exaggerated sex steroetypes (Jussim, 2018). Turning to the relationship with intelligence, Figure X shows the scatterplots.

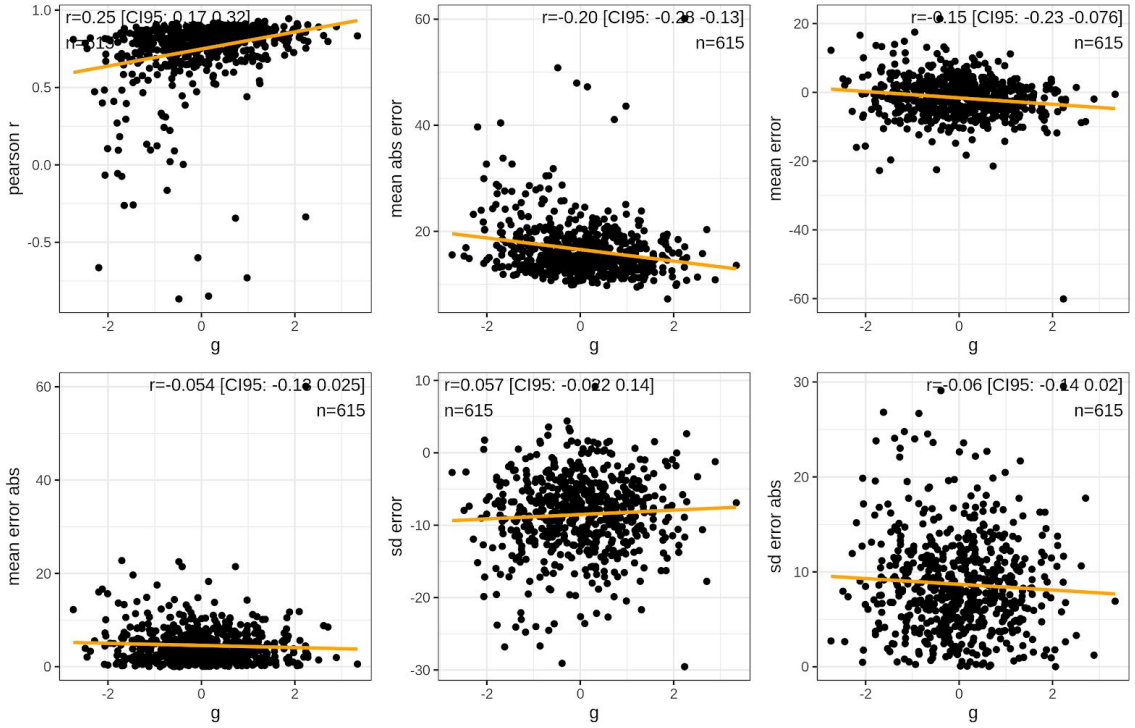


Figure X. Scatterplots for intelligence and stereotype accuracy metrics for sex differences in occupation.

As with crime rates, we see notable accuracy levels for individual estimates, with correlations with Pearson r accuracy of .25, and MAE of -.20 (lower better). For the componential errors, we see that smarter people tend to underestimate the proportion of men in the occupations, though not by much, and the tendency is only moderate (r = -.15). The Pearson r accuracies have a number of outlying values. If one removes values below 0, the correlation with intelligence increases to r = .32. The other associations are too small to care about (|r| < .10). Table X expands this approach to our OLS regression models with the other predictors.

Predictor	**Small model**	**Full model**
Intercept	0.04 (0.060)	-0.32 (0.355)
g	0.27 (0.045***)	0.25 (0.048***)
Verbal tilt	0.07 (0.045)	0.06 (0.047)
age	-0.08 (0.046)	-0.10 (0.056)
male	-0.09 (0.082)	-0.08 (0.089)
education	0.02 (0.044)	0.03 (0.050)
time taken	-0.01 (0.041)	0.01 (0.042)

First Language=Dutch		(ref)
First Language=non-Dutch		-0.60 (0.163***)
Birth=Netherlands		(ref)
Birth=Non-Western		0.31 (0.167)
Birth=Western		0.15 (0.321)
student		-0.09 (0.113)
employment status		(included)
vote PvdD		0.05 (0.232)
vote Groenlinks		-0.01 (0.152)
vote SP		0.11 (0.211)
vote D66		0.07 (0.191)
vote PvDA		0.31 (0.207)
vote VVD		0.20 (0.167)
vote Christenunie		0.38 (0.282)
vote PVV		-0.08 (0.224)
vote CDA		0.20 (0.319)
vote FvD		0.44 (0.213)
vote SGP		0.33 (0.887)
vote 50Plus		0.51 (0.563)
vote DENK		0.57 (1.973)
R2 adj.	0.062	0.068
N	598	589

Table X. Regression results for Pearson r stereotype accuracy for sex differences in occupation.

Intelligence keeps its position as the dominant quantitative predictor (β = 0.25), while we also see a substantial effect (β = -0.60) of being a non-Dutch native speaker. This second finding could be interpreted as a lack of language skill with the survey, however, this predictor was not strong for crime rates (β = -0.09, p > .05), which makes this

explanation implausible. No other predictor reaches detectable utility (as $p < .01$, unadjusted for multiple testing), so it appears the model can be substantially simplified. Table X gives the results from the lasso regression, while Table X gives the results from BMA.

Predictor	**Beta**
g	0.17
vote FvD	0.00
First Language: non-Dutch	-0.28

Table X. Lasso results for Pearson r stereotype accuracy for sex differences in occupation. Intercept left out.

Predictor	**PIP (%)**	**Post mean (beta)**	**Post SD**
g	100	0.22	0.04
Verbal tilt	2.4	0	0.01
age	3.4	0	0.01
male	4.8	-0.01	0.03
time taken	0	0	0
education	0	0	0
First Language: non-Dutch	95.9	-0.49	0.18
Birth: Non-Western	14.9	0.05	0.13
Birth: Western	0	0	0
student	0	0	0
Employment Status: Full-Time	2.5	0	0.02
Employment Status: Not in paid work	0	0	0
Employment Status: Other	3.6	0.01	0.04
Employment Status: Part-Time	0	0	0
Employment Status: Unemployed and job seeking	0	0	0
vote PvdD	0	0	0
vote Groenlinks	3.5	0	0.03

vote SP	0	0	0
vote D66	0	0	0
vote PvDA	2.9	0	0.04
vote VVD	0	0	0
vote Christenunie	4	0.01	0.08
vote PVV	2.9	-0.01	0.05
vote CDA	0	0	0
vote FvD	5.9	0.02	0.08
vote SGP	0	0	0
vote 50Plus	2.2	0.01	0.09
vote DENK	0	0	0

Table X. Bayesian modeling averaging results for Pearson r stereotype accuracy for sex differences in occupation. Intercept left out.

The approaches are mostly in agreement: there are only 2 notable predictors, intelligence and being a non-Dutch native speaker. Lasso oddly includes a political party but with a near-zero beta, so this should probably be seen as a fluke. Our models are not much predictive of variation in this stereotype (full model adj. R2 = 6.8%, compared to 13.0% for crime rates), despite the inclusion of many demographics variables and political variables that should capture aspects of feminist ideology, which conceivably would be related to stereotype accuracy in this domain.

Provincial incomes

Finally, subjects estimated the mean incomes of the 12 provinces of the Netherlands. Table gives the summary statistics of the accuracy metrics, while Figure X gives the distributions.

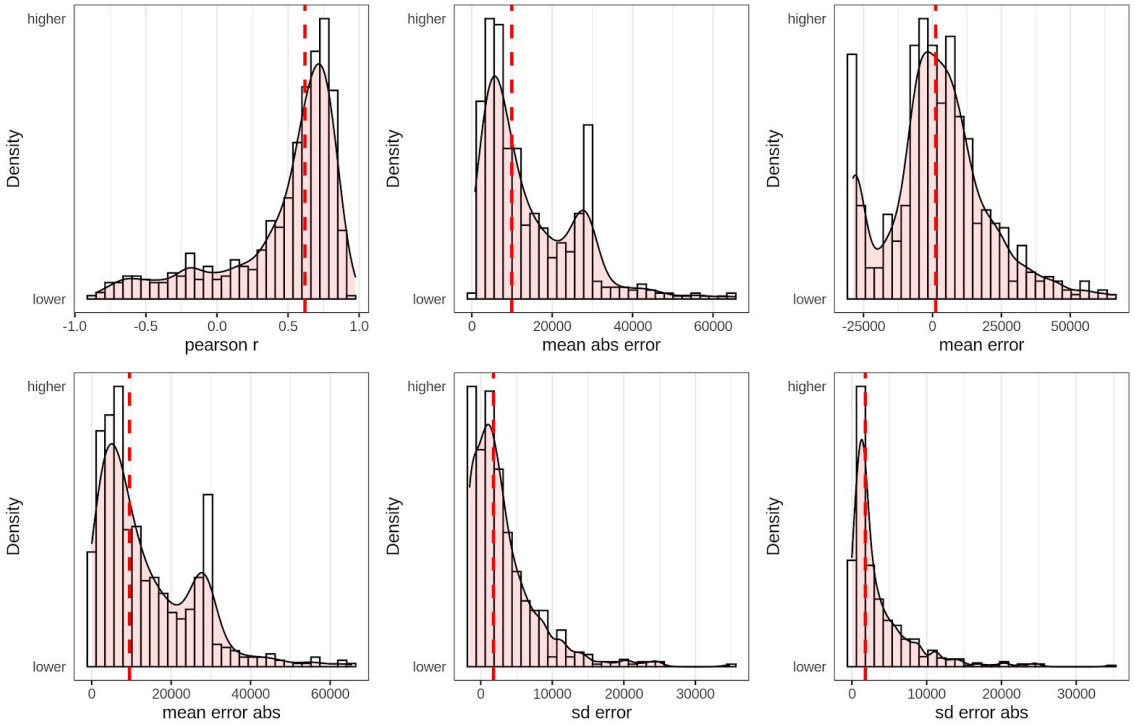


Figure X. Distributions of province income stereotype accuracy metrics.

Variable	N	Mean	Median	SD	MAD	Min	Max	Skew
pearson r	585	0.46	0.62	0.41	0.22	-0.85	0.97	-1.44
rank r	585	0.46	0.61	0.41	0.25	-0.90	0.97	-1.42
mean abs error (MAE)	615	14168.21	9975.00	11633.89	8870.89	816.67	65358.33	1.31
sd	615	4640.01	3342.79	4782.39	3211.54	0.00	36271.28	2.15
sd error	615	3084.69	1787.47	4782.39	3211.54	-1555.32	34715.97	2.15
sd error abs	615	3644.91	1787.47	4369.79	1774.10	1.68	34715.97	2.63
mean	615	30382.05	30000.00	17987.94	13590.50	0.00	94166.67	0.40
mean error	615	1573.71	1191.67	17987.94	13590.50	-28808.33	65358.33	0.40
mean error abs	615	13612.91	9525.00	11850.45	9439.22	58.33	65358.33	1.23

Table X. Summary statistics of province income stereotype accuracy metrics. sd = standard deviation.

In correlational terms, there is substantial accuracy, with a median Pearson r of .62, and rank r of .61. The means, however, are quite reduced (both r's .46), in agreement with the long tail towards -1 seen in the distributions. This tail is curious, as it seems unlikely

some people are purposefully filling out the questionnaire reverse of their real beliefs, as was seen previously with immigrant stereotypes (Kirkegaard & Bjerrekær, 2016). The median mean error was close to 0 (1,192, or about 4% off the true mean value of 28,808), indicating that the scale was understood by the subjects despite its somewhat technical nature (disposable income). Interestingly, the median SD error was positive (1,787), showing that subjects overestimated real differences. In relative terms, this effect is very large, the median estimated SD was 3343 but the true was only 1555, so the estimate was 115% too large! Apparently, the public believes provincial income inequality in disposable income as much, much larger than it really is. Moving on to prediction, Figure X shows the scatterplots with intelligence.

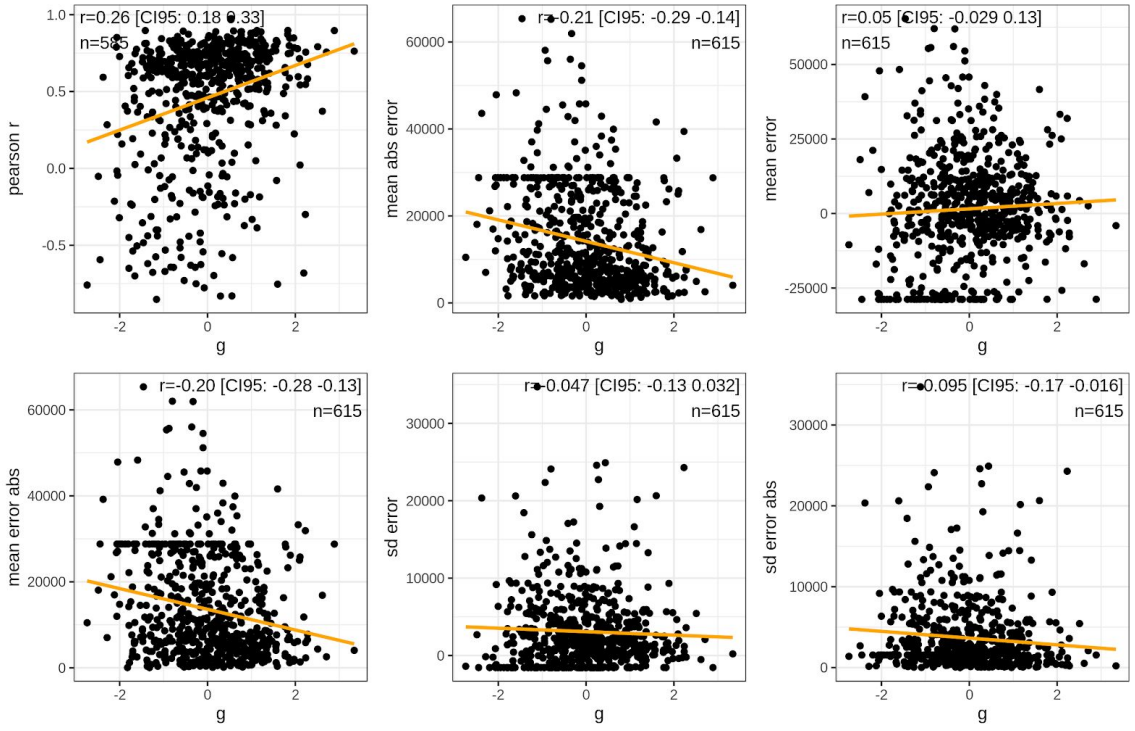


Figure X. Scatterplots of intelligence and province income stereotype accuracy metrics.

The scatterplots reveal the presence of outlying values. Presumably, these represent people who grossly misunderstood the task and somehow gave results opposite of reality. These people are mostly clustered among below average intelligence subjects, giving rise to a correlation of .26 with intelligence. If subjects below Pearson r accuracy of 0 are removed, the correlation is reduced to .18. This is in contrast with the occupational stereotypes where removing outliers (below 0) resulted in a stronger correlation (from .25 to .32). Looking at the plots also reveals a cluster of people with mean errors around -26k. These are people who filled in very small values, relative to this scale. Inspection of these cases showed that some of them are probably lazy responding (e.g., 30 people filled in responses with zero variance), and some are people who assumed they were giving values in the 1000's (e.g., one person filled in varying values between 32 and 45). As a robustness test, we removed all subjects who filled in values with a mean below 1000 (i.e., more than factor 26 off), and those with no

variance, n = 536 cases remaining of 615, or 87%. The relationship to intelligence was mostly unaffected, r = .25 with Pearson r, r = -19 with MAE. The correlations with the mean and SD metrics became a bit stronger, but overall, this exploratory analysis did not change much. Furthermore, a large fraction of the outliers with Pearson r < 0 remained after this filtering (15.6% before, and 14.9% after). Table X gives the regression results that expands the analysis to the other predictors. This analysis was done on the full dataset, as the above exploratory analyses did not reveal serious issues because of the data problems.

Predictor	**Small model**	**Full model**
Intercept	-0.15 (0.060)	-0.29 (0.349)
g	0.22 (0.045***)	0.21 (0.048***)
Verbal tilt	-0.02 (0.045)	-0.02 (0.047)
age	0.08 (0.047)	0.07 (0.059)
male	0.27 (0.082***)	0.31 (0.089***)
education	0.08 (0.044)	0.05 (0.050)
time taken	-0.09 (0.041)	-0.09 (0.043)
First Language=Dutch		(ref)
First Language=non-Dutch		0.12 (0.169)
Birth=Netherlands		(ref)
Birth=Non-Western		0.17 (0.171)
Birth=Western		0.12 (0.314)
student		-0.12 (0.116)
employment status		(included)
vote PvdD		0.11 (0.238)
vote Groenlinks		-0.05 (0.152)
vote SP		-0.17 (0.213)
vote D66		-0.15 (0.188)
vote PvDA		-0.36 (0.203)
vote VVD		-0.07 (0.168)

vote Christenunie		-0.10 (0.285)
vote PVV		-0.58 (0.224*)
vote CDA		-0.06 (0.339)
vote FvD		-0.18 (0.210)
vote SGP		-1.01 (0.867)
vote 50Plus		-0.15 (0.551)
vote DENK		-0.77 (1.927)
R2 adj.	0.062	0.068
N	598	589

Table X. Regression results for province income stereotype accuracy. All available data used.

The full regression results reveal only 3 useful predictors: intelligence (β = 0.21), being male (β = 0.31), and voting for the PVV party (β = -0.58, p < .01). The first is not surprising, given the prior findings. Men have a stronger interest in economic matters and greater scientific knowledge, including economics (Caplan & Miller, 2010; Tran et al., 2014), so the second finding is not surprising. The latter finding is odd, as there is no obvious reason why voting for this nationalist party should be related to stereotype accuracy of provinces, and not other nationalist parties, or the opposite effect for anti-nationalist parties. As before, the model had many variables of doubtful importance, and we used lasso regression and BMA to prune it, results are given in Tables X and X.

Predictor	**Beta**
g	0.15
male	0.13
education	0.03
student	-0.02
vote PVV	-0.13
Employment Status: Other	-0.02

Table X. Lasso regression results for province income stereotype accuracy.

Predictor	PIP (%)	Post mean (beta)	Post SD
g	100.00	0.23	0.04
Verbal tilt	0.00	0.00	0.00
age	5.20	0.00	0.02
male	97.40	0.27	0.09
time taken	27.40	-0.03	0.05
education	11.20	0.01	0.03
First Language: non-Dutch	3.60	0.01	0.05
Birth: Non-Western	3.00	0.01	0.04
Birth: Western	0.80	0.00	0.04
student	20.60	-0.04	0.09
Employment Status: Full-Time	3.60	0.00	0.03
Employment Status: Not in paid work	0.00	0.00	0.00
Employment Status: Other	12.30	-0.03	0.10
Employment Status: Part-Time	0.70	0.00	0.01
Employment Status: Unemployed and job seeking	1.90	0.00	0.03
vote PvdD	0.90	0.00	0.03
vote Groenlinks	0.00	0.00	0.00
vote SP	0.00	0.00	0.00
vote D66	0.00	0.00	0.00
vote PvDA	4.80	-0.01	0.07
vote VVD	1.10	0.00	0.02
vote Christenunie	0.00	0.00	0.00
vote PVV	36.40	-0.17	0.26
vote CDA	0.70	0.00	0.03
vote FvD	0.00	0.00	0.00
vote SGP	1.00	-0.01	0.12
vote 50Plus	0.00	0.00	0.00
vote DENK	0.00	0.00	0.00

Table X. Bayesian model averaging results for province income stereotype accuracy.

The methods were again mostly in agreement. Both find intelligence to be the most important predictor, in the case of BMA, including it in 100% of the best models. Male status was afterwards the most important being included in 97% of the best models. Both approaches included voting for PVV to some degree. With BMA, the influence is uncertain, the beta SD is large, and it is only included in 36% of the best models. BMA furthermore sporadically included other variables (e.g., time taken, 27% of models), but not much can be made of this.

Domain general accuracy

In the last part of the individual-level analysis, we looked for evidence of a general factor of stereotype accuracy. Figure X shows the heatmap of correlations for all metrics across domains.

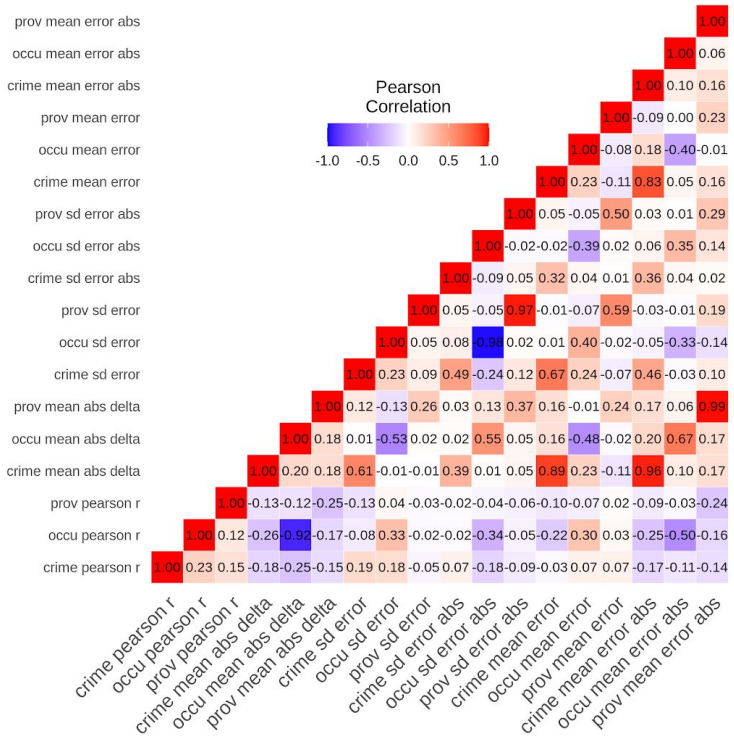


Figure X. Heatmap for stereotype accuracy metrics. Crime = crime rates, occu = occupational distributions by sex, prov = provincial incomes, abs = absolute, sd = standard deviation.

The results show that there are many associations between the metrics both within and between domains. However, most of the cross-domain correlations are weak. For instance, the correlations among the correlational accuracy metrics are only .12 to .23, i.e., there is little shared variance. The same pattern is seen when one looks at the MAE where the correlations range from .18 to .20. The same is seen when one looks at the more specific sources of error. If we look at dispersion errors, the correlations range from

.05 to .23. In other words, there is only a very weak tendency for people who overestimate differences (i.e., have positive SD error) in one domain to do so also in other domains. For mean errors, there was no consistency across domains, with correlations ranging from -.08 to .23. In other words, people who overestimated (or underestimated) values in one domain had no tendency to do so in other domains. Looking at the cross-metric within-domain associations, we find correctly signed relations between correlational accuracy and MAE of -.18, -.92 and -.25. It's not obvious why the correlation is so much stronger for the occupations than for the two other domains. In similar fashion, the correlation between correlational accuracy and mean error absolute was also correctly signed by weak, at -.17, -.50, and -.24. Again, the outlier is for the occupations. In contrast, the associations with SD error abs. were unimpressive and centered around 0 (i.e., people who were more inaccurate about the differences between groups were not the same across domains).

Though the correlations are weak across domains, they are there. How do the factor loadings look like, if we postulate a general factor of stereotype accuracy? While one could use just the MAE metric, which is all inclusive (any deviation from truth), we decided to use both Pearson r and MAE alone, and in combination. Table X shows the results.

Indicator	**Factor loading EFA**	**Factor loading UWFA**
Crime pearson r	-0.30	0.55
Occu pearson r	-0.94	0.75
Prov pearson r	-0.19	0.50
Crime mean abs error	0.28	-0.55
Occu mean abs error	0.92	-0.74
Prov mean abs error	0.24	-0.53

Table X. Factor analysis results of primary stereotype accuracy metrics. EFA = exploratory factor analysis, as implemented in **psych**'s fa() function. UWFA = unit-weighted factor analysis, which is the same as the row-wise sum of z-scores.

Unexpectedly, the planned factor analysis produced poor results, in that one domain had extreme influence on the factor. This is because the two metrics of accuracy in the occupational domain are strangely strongly correlated (r = -.92). This resulting factor is thus mostly just the occupational accuracy score, which is not what was desired. Deviating from our planned analysis, we employed unit-weighted factor analysis. This is a more robust alternative to the more common differential weights factor analysis. In this method, the indicators are given equal weights, and the loading loadings resulting are

the correlations to this resulting score. This method is better in some edge cases where factor loadings vary unexpectedly or sample sizes are too small for stable results (Figueredo et al., 2000; Gorsuch, 2015, sec. 12.2.2). In our case, we see that UWFA produced more even loadings, i.e., more even influences from the three domains. We furthermore scored each metric for its separate UWFA score. Table X shows the correlations among the factor scores and indicators.

	g intelligence	g accu pearson	g accu MAE	g accu UWFA	g accu EFA	Crime pearson r	Occu pearson r	Prov pearson r	Crime mean abs error	Occu mean abs error	Prov mean abs error
g intelligence	1	0.35	0.35	0.39	0.26	0.20	0.25	0.26	-0.30	-0.20	-0.21
g accu pearson	0.35	1	0.60	0.89	0.73	0.69	0.69	0.65	-0.29	-0.65	-0.27
g accu MAE	0.35	0.60	1	0.90	0.74	0.29	0.66	0.25	-0.68	-0.68	-0.67
g accu UWFA	0.39	0.89	0.90	1	0.83	0.55	0.75	0.50	-0.55	-0.74	-0.53
g accu EFA	0.26	0.73	0.74	0.83	1	0.28	0.98	0.20	-0.26	-0.96	-0.27
Crime pearson r	0.20	0.69	0.29	0.55	0.28	1	0.23	0.15	-0.18	-0.25	-0.15
Occu pearson r	0.25	0.69	0.66	0.75	0.98	0.23	1	0.12	-0.26	-0.92	-0.17
Prov pearson r	0.26	0.65	0.25	0.50	0.20	0.15	0.12	1	-0.13	-0.12	-0.25
Crime mean abs error	-0.30	-0.29	-0.68	-0.55	-0.26	-0.18	-0.26	-0.13	1	0.20	0.18
Occu mean abs error	-0.20	-0.65	-0.68	-0.74	-0.96	-0.25	-0.92	-0.12	0.20	1	0.18
Prov mean abs error	-0.21	-0.27	-0.67	-0.53	-0.27	-0.15	-0.17	-0.25	0.18	0.18	1

Table X. Correlation matrix of general stereotype accuracy factor scores and their indicators. g = general factor, accu = accuracy of stereotypes, crime = crimes rates, occu = sex differences in occupation, prov = provincial incomes, abs = absolute, r = Pearson correlation.

We see that the UWFA produced stronger correlations, while the EFA scores were essentially just a duplicate of the occupational scores. Thus, we used the UWFA scores

for further analysis (this is a deviation from our pre-analysis plan). The resulting correlation with intelligence was impressive, r = .39, shown in Figure X.

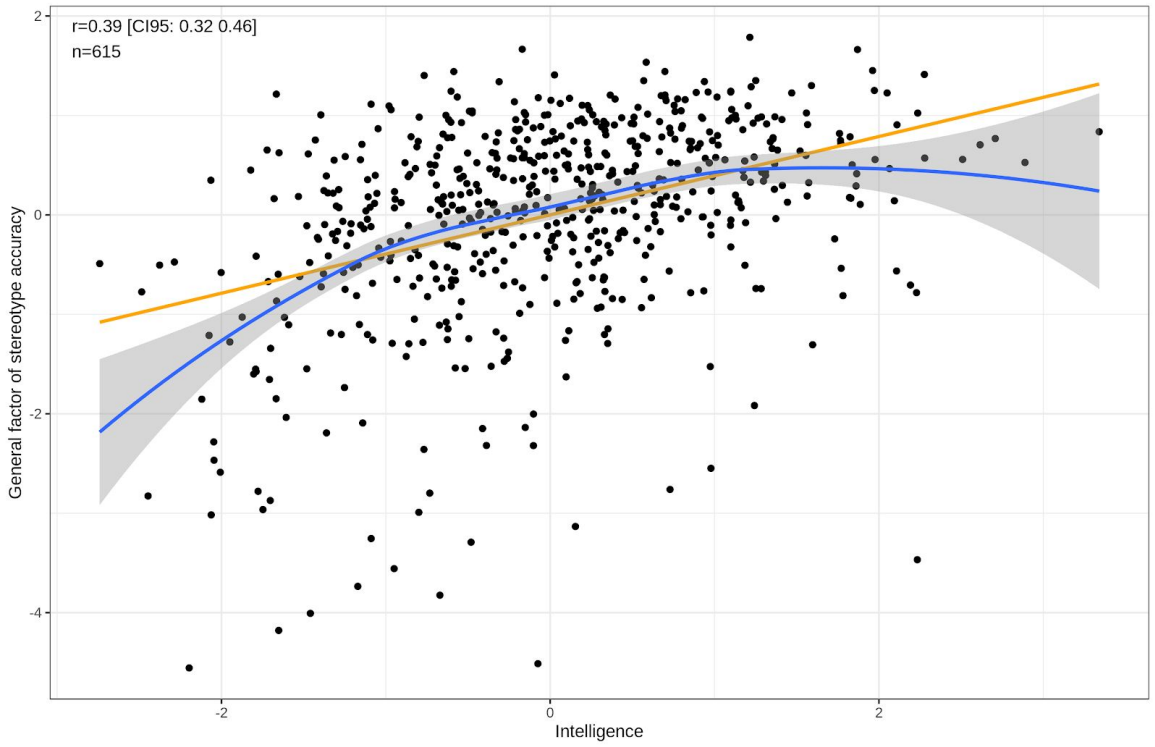


Figure X. Scatterplot of general stereotype accuracy (based on 3 domains and 6 indicators) and intelligence. Orange line is linear fit and blue blue is LOESS fit.

The scatterplot has some outlying values, a result of the outliers in the components. The LOESS fit seems to indicate some nonlinearity, which is confirmed by a model comparison of a linear model vs. a natural spline model ($p < .0001$; adj. R^2's 15.4% and 18.2%). Tabel X shows the regression results.

Predictor	**Small model**	**Full model**
Intercept	-0.08 (0.055, 0.157)	-0.30 (0.329, 0.366)
g	0.39 (0.042, <0.001***)	0.37 (0.044, <0.001***)
Verbal tilt	0.12 (0.042, 0.004**)	0.12 (0.044, 0.006*)
age	-0.12 (0.042, 0.005**)	-0.15 (0.052, 0.004**)
male	0.13 (0.076, 0.079)	0.14 (0.082, 0.096)
education	0.10 (0.040, 0.012)	0.11 (0.046, 0.015)

time taken	-0.05 (0.038, 0.187)	-0.04 (0.039, 0.294)
First Language=Dutch		(ref)
First Language=non-Dutch		-0.40 (0.151, 0.008*)
Birth=Netherlands		(ref)
Birth=Non-Western		0.20 (0.154, 0.185)
Birth=Western		0.04 (0.297, 0.891)
student		-0.21 (0.104, 0.05)
employment status		(included)
vote PvdD		-0.26 (0.215, 0.235)
vote Groenlinks		-0.16 (0.141, 0.269)
vote SP		-0.07 (0.195, 0.704)
vote D66		-0.05 (0.176, 0.767)
vote PvDA		-0.05 (0.191, 0.802)
vote VVD		-0.09 (0.154, 0.581)
vote Christenunie		0.19 (0.261, 0.477)
vote PVV		-0.44 (0.207, 0.033)
vote CDA		-0.02 (0.295, 0.952)
vote FvD		0.25 (0.197, 0.204)
vote SGP		-0.16 (0.821, 0.849)
vote 50Plus		0.37 (0.521, 0.481)
vote DENK		-0.37 (1.826, 0.84)
R2 adj.	0.18	0.187
N	598	589

Table X. Regression results for general stereotype accuracy. Values in parentheses are standard errors and p values.

Overall the model is sparse. Intelligence (β = 0.37) and being a non-Dutch native speaker (β = -0.40) are the most important variables, but others also cross the p = .01 barrier: verbal tilt (β = 0.12), and age (β = -0.15). The fact that both of these cross and are positively correlated and yet have opposite betas is surprising. These variables mostly had same direction betas in the prior analyses. This pattern indicates suppression effects (i.e., where direct and indirect effects differ, presumably of age). As before, we attempt to simplify this model with lasso regression and BMA, the results are shown in Tables X and X.

Predictor	**Beta**
g	0.35
Verbal tilt	0.09
age	-0.11
male	0.09
time taken	-0.02
education	0.10
student	-0.13
vote PvdD	-0.14
vote Groenlinks	-0.07
vote Christenunie	0.16
vote PVV	-0.32
vote FvD	0.23
vote 50Plus	0.18
First Language=non-Dutch	-0.33
Birth=Non-Western	0.11
Employment Status: Part Time	0.08
Employment Status: Unemployed and job seeking	0.15

Table X. Lasso regression results for general stereotype accuracy. Intercept left out.

Predictor	PIP (%)	Post mean (beta)	Post SD
g	100.00	0.36	0.05
Verbal tilt	27.10	0.03	0.05
age	46.50	-0.06	0.08
male	3.10	0.00	0.02
time taken	1.20	0.00	0.01
education	67.60	0.08	0.06
First Language: non-Dutch	55.20	-0.20	0.21
Birth: Non-Western	1.00	0.00	0.02
Birth: Western	0.00	0.00	0.00
student	19.50	-0.05	0.10
Employment Status: Full-Time	0.00	0.00	0.00
Employment Status: Not in paid work	0.80	0.00	0.02
Employment Status: Other	0.70	0.00	0.01
Employment Status: Part-Time	0.50	0.00	0.01
Employment Status: Unemployed and job seeking	3.80	0.01	0.04
vote PvdD	0.90	0.00	0.03
vote Groenlinks	1.50	0.00	0.02
vote SP	0.00	0.00	0.00
vote D66	0.00	0.00	0.00
vote PvDA	0.00	0.00	0.00
vote VVD	0.00	0.00	0.00
vote Christenunie	0.80	0.00	0.03
vote PVV	20.40	-0.08	0.18
vote CDA	0.00	0.00	0.00
vote FvD	12.30	0.04	0.13
vote SGP	0.00	0.00	0.00
vote 50Plus	0.50	0.00	0.04

vote DENK	0.00	0.00	0.00

Table X. Bayesian model averaging results for general stereotype accuracy. Intercept left out.

Surprisingly, lasso regression found that most (17) predictors were needed, while BMA was more parsimonious in the conclusions. Only intelligence was included in all the best models, while some others were also in the majority of models (education 68% of models, non-Dutch first language, 55%), as well as some with more sporadic appearance (e.g. PVV voting, 20% of models).

Data source effects

Though we did not plan to collect data from two different pollsters, the fact that we did so lets us examine whether they produced different results with regards to stereotype accuracy. We saw earlier that Prolific subjects were younger, more left-wing, more likely to be students, smarter, so it is possible they also differ on stereotype accuracy, even beyond the effects of the measured variables. Here we formally test this by adding a source dummy to the regression models. Table X shows the results.

Predictor/Model	**Crime pearson r**	**Occu pearson r**	**Prov pearson r**	**Crime mean abs error**	**Occu mean abs error**	**Prov mean abs error**	**g accu UWFA**
g	0.17 (0.047***)	0.21 (0.049***)	0.19 (0.050***)	-0.26 (0.045***)	-0.15 (0.049**)	-0.20 (0.049***)	0.33 (0.045***)
source=Survee	0.01 (0.111)	-0.35 (0.113**)	-0.22 (0.114)	0.42 (0.105***)	0.28 (0.114)	0.13 (0.112)	-0.38 (0.105***)
R2 adj.	0.131	0.082	0.090	0.226	0.064	0.053	0.204
R2 adj. without source	0.133	0.068	0.086	0.205	0.056	0.052	0.187
Change in R2 adj.	-0.002	0.014	0.004	0.021	0.008	0.001	0.017
N	572	589	559	589	589	589	589

Table X. Data source effects in regression model results (abbreviated results, full results in statistical output). * = p < .01, ** = p < .005, *** = p < .001.

We find evidence of source effects for 3 of the 7 models tested, and most important, for the final model with general stereotype accuracy. Surprisingly, the effect is seen for two different outcomes, but not on one metric: occupational sex differences with Pearson r, and crime rates with MAE. The effect size on the general stereotype accuracy score is quite large at $\beta = -0.38$, with stronger accuracy seen for Prolific users. It is not clear why this is the case, as we statistically controlled here for many of the things that differ between the survey sources (as mentioned in Table X).

Aggregate level results

Having seen the complexity of the individual level results, we are now ready to examine the aggregated results. When aggregating results, opposite-direction errors cancel out, which usually results in a much stronger signal, and in the case of stereotypes, much higher accuracy. However, this is not necessarily the case, but depends crucially on the structure of the errors in estimation. Insofar as these go in the same direction, it can lead to large overall errors in stereotypes (Surowiecki, 2004).[7]

Immigrants and crime rates

As before, we begin with the primary domain of immigrant crime rates. There are many ways to aggregate estimates to a single set of values by taking into account the prior history of and the correlations among estimators (Atanasov et al., 2016; Lyon & Pacuit, 2013; Navajas et al., 2018). Surprisingly, the simplest is among the best: take the arithmetic mean. Table X shows the stereotype accuracy metrics across 4 aggregation methods.

Method	**Pearson r**	**Rank r**	**Mean abs error**	**SD**	**SD error**	**Mean**	**Mean error**
mean	0.65	0.68	0.60	0.50	-0.48	1.75	0.15
10% trimmed mean	0.65	0.70	0.57	0.46	-0.53	1.43	-0.17
median	0.59	0.65	0.66	0.27	-0.72	1.19	-0.41
log mean	0.65	0.69	0.62	0.37	-0.62	1.20	-0.40

[7] The book summarizes the conceptual argument: “The market was smart that day because it satisfied the four conditions that characterize wise crowds: diversity of opinion (each person should have some private information, even if it’s just an eccentric interpretation of the known facts), independence (people’s opinions are not determined by the opinions of those around them), decentralization (people are able to specialize and draw on local knowledge), and aggregation (some mechanism exists for turning private judgments into a collective decision). If a group satisfies those conditions, its judgment is likely to be accurate. Why? At heart, the answer rests on a mathematical truism. If you ask a large enough group of diverse, independent people to make a prediction or estimate a probability, and then average those estimates, the errors each of them makes in coming up with an answer will cancel themselves out. Each person’s guess, you might say, has two components: information and error. Subtract the error, and you’re left with the information.”

Table X. Stereotype accuracy for immigrant crime rates across 4 aggregation methods. Log mean consists of taking the log of the values, taking the mean, and taking the exponential to return to the same scale. The mean/SD of the criterion data are 1.60/0.99.

Here we find that the trimmed mean (at 10%) and the untrimmed mean do about the same, and both do better than the median (which is the same as the 50% trimmed mean). Overall, however, the correlational accuracy is substantial with a Pearson r of .65 and rank (Spearman) r of .70. All methods substantially *underestimate* the true variability between groups (the negative values in SD error), estimating the SD in relative crime rates to be about 50% of its true value. The estimation of the mean value is, however, quite accurate, being slightly too high using the untrimmed mean and slightly too low using the 10% trimmed (0.15 vs. -0.17), whereas the median fares very poorly with a substantial underestimation (-0.41). For simplicity's sake, however, we will be using the untrimmed mean for further analysis (stated in our pre-analysis plan). Figure X shows the scatterplot between the average estimates and the criterion values.

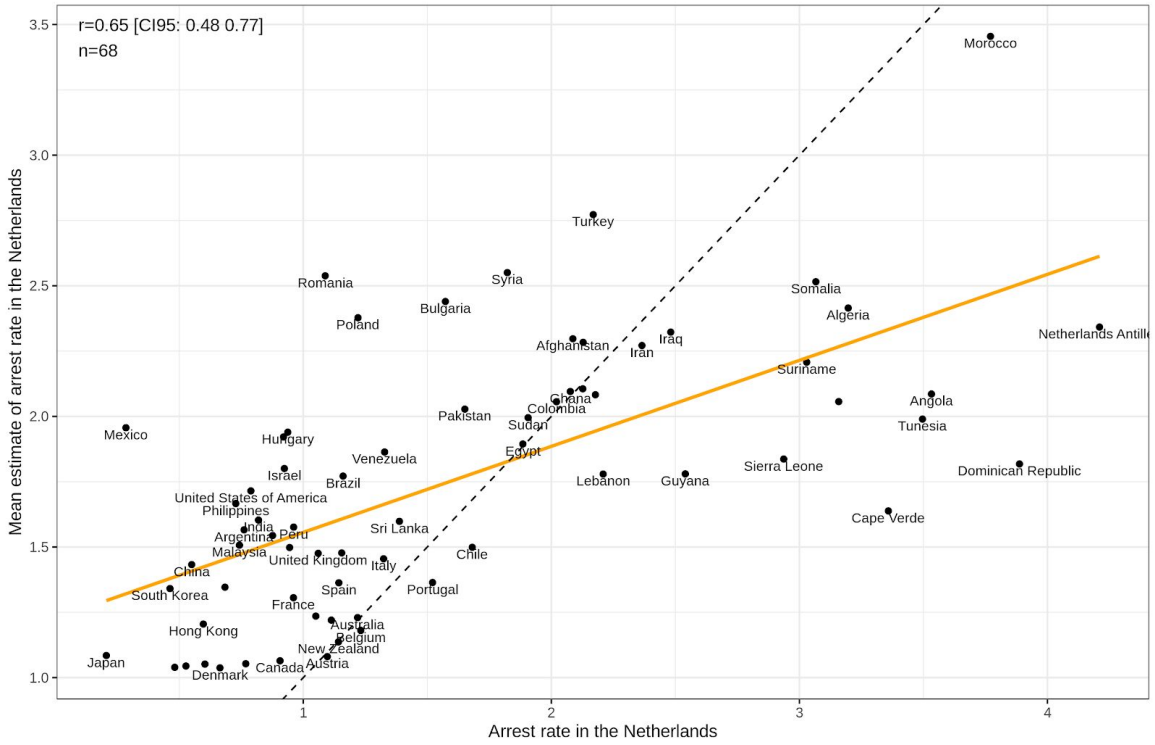


Figure X. Scatterplot of immigrant group crime rates and average stereotypes. The stippled line shows the slope = 1 of perfect calibration, the orange line is the linear fit.

One striking finding is that no country has a below Dutch crime rate estimate, but about 35% of countries actually have a below Dutch rate. In this sense, the crime rates of these countries are all overestimated. At the same time, however, the crime rates of the high crimes are underestimated, sometimes substantially so. The Netherlands Antilles (a Caribbean former colonial possession) has an actual rate of 4.2, while the estimate is only 2.3. If we inspect the distribution of estimates for some countries, we get an idea of why this may be so, shown in Figure X.

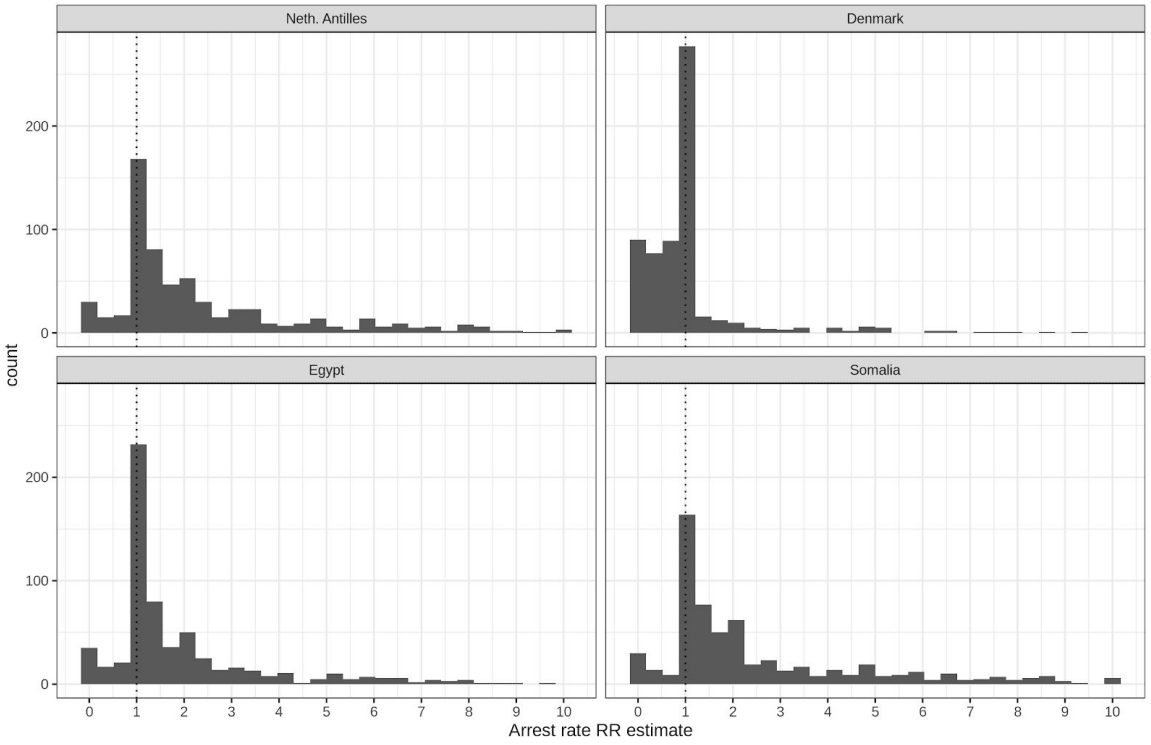


Figure X. Distributions of stereotypes of relative rate (RR) of crime rates for 4 example countries.

For each country, the most chosen value is 1, i.e. the estimate is that immigrants from that country are the same as Dutch natives in crime rate. However, the countries differ in the length of the right tail, thus producing the differences in means. The problem here is that values above 1 have a greater influence than values below 1, even though they are both, in a sense, equally distant from the value of 1. To see this, imagine if we inverted the scale to be the number of times less criminal than Dutch natives (i.e., we took the reciprocal, 1/x). A low crime origin might then be assigned a value of 3 (commits crimes at 1/3 the rate of Dutch natives), and so on. There is a way to avoid this positive bias inherent in the scale, namely to convert the numbers to log scale, take the average, and convert back. This is because, on the log scale, 1/3 and 3 are both equally distant from 1 (i.e., -1.1, and 1.1), so the average of two estimates who describe a group as 1/3 and 3 times as criminal as the natives is 1 (same rate as natives). Figure X shows the results when the mean is taken of these log-converted values (log mean method in Table X).

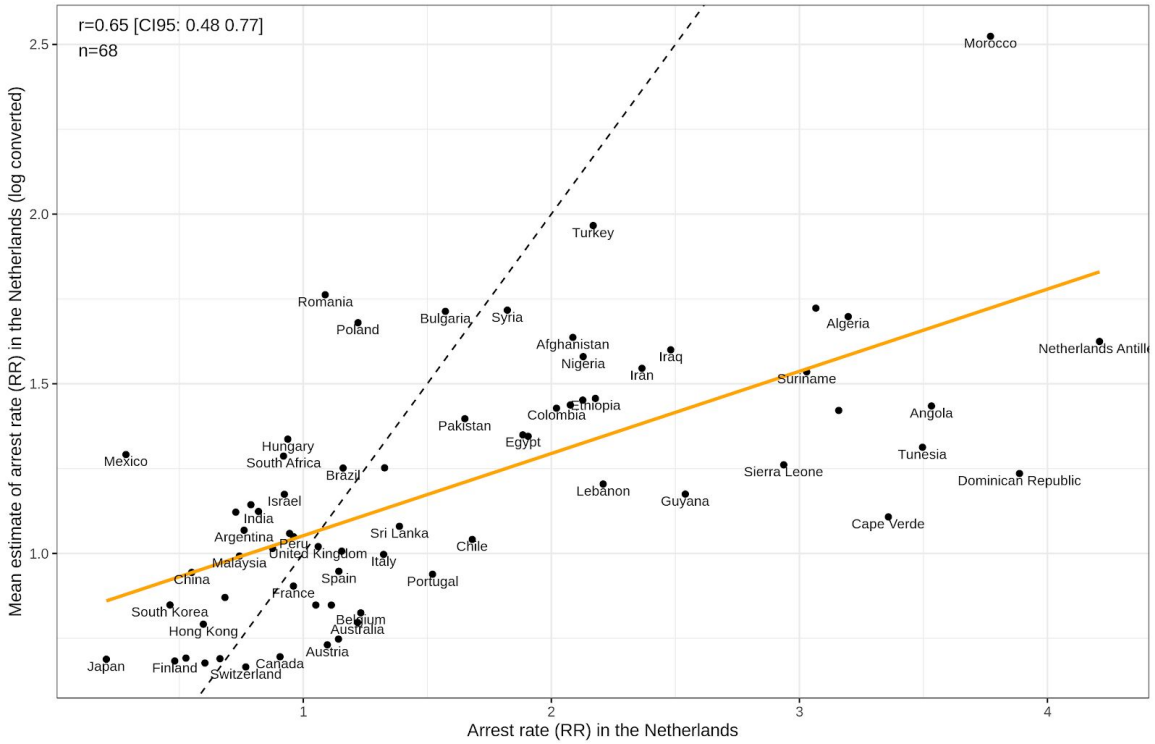


Figure X. Scatterplot of immigrant group crime rates and average stereotypes, using the log-conversion. The stippled line shows the slope = 1 of perfect calibration, the orange line is the linear fit.

Thus, we see that this approach helps with the below 1 values, but also reduces the estimates for the high crime groups. In fact, if we look back at Table X, this approach produces worse overall results in terms of even more severely underestimating the real differences (by about 60%) and the overall mean is also much too low. The result is that the linear fit (orange) deviates further from the perfect calibration fit (stippled line) than before. Beating the simple mean isn't easy. Turning to the question of Muslim bias in the ratings, Figure X shows the scatterplot of Muslim% and the estimation error.

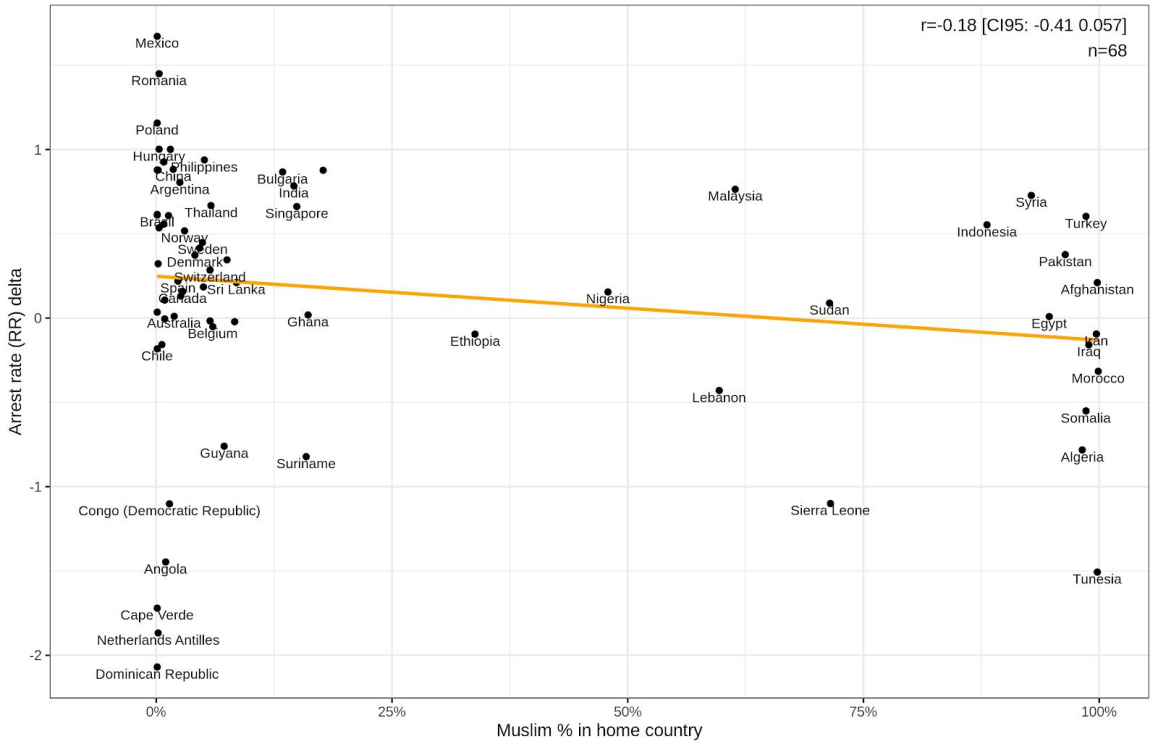


Figure X. Scatterplot of Muslim% and the estimation error from the average stereotype.

In line with the individual level results, we see a tendency to underestimate the crimes rates of the more Muslim groups. There are a number of interesting outliers in the bottom left, countries that had strong underestimation of values, yet do not have many Muslims. This weak pattern in the aggregate data is similar to prior studies with Danish data (Kirkegaard et al., 2020; Kirkegaard & Bjerrekær, 2016). In terms of immigration opinions, a prior study measured the preferences for the same origin groups in a sample of 200 people living in the Netherlands (Kirkegaard & de Kuijper, 2020), partially overlapping with the present. Figure X shows the results.

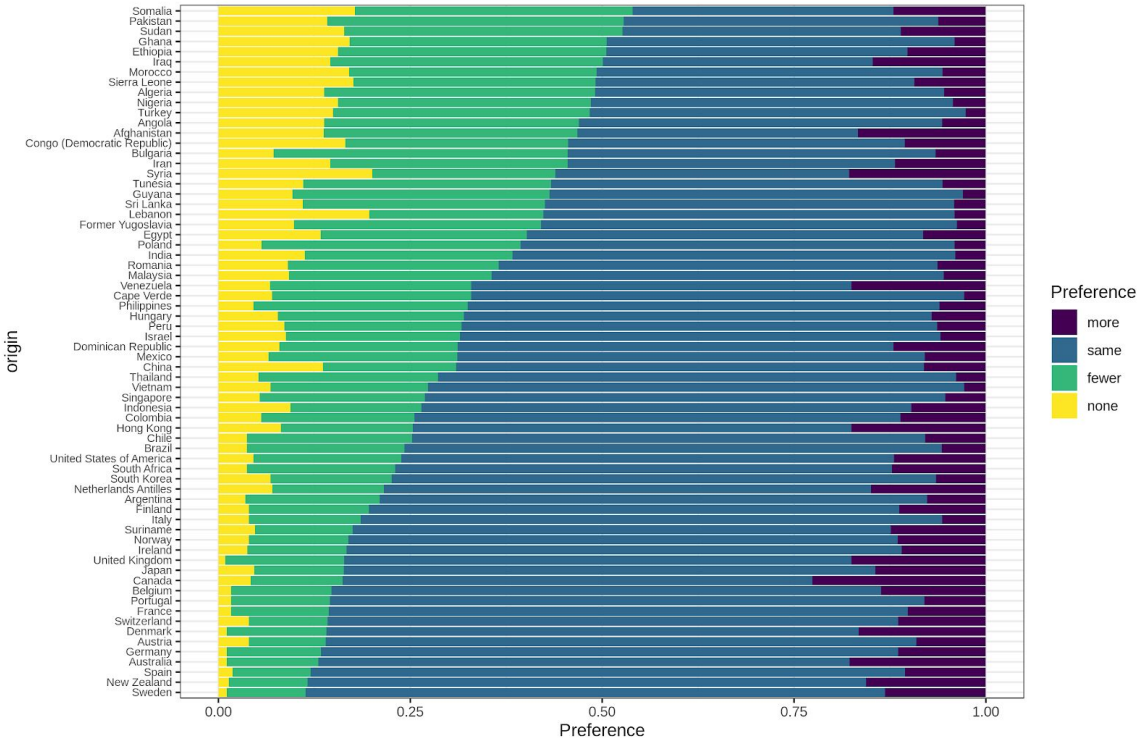


Figure X. Distribution of immigrant preferences by origin group.[8]

Prior research has found that measured stereotypes mediated the link between real crime rates and immigration preferences, a hypothesis suggested by (Carl, 2016). In other words, people are more opposed to immigration from high crime origins, and their preferences are in line with what one would expect based on their actual beliefs (Kirkegaard et al., 2020). To test this model further, we carried out the same mediation tests in the present dataset as done previously for a Danish dataset. Table X shows the correlation matrix for the variables in question.

	Muslim%	**Arrest rate**	**Mean estimate**	**Estimate error**	**Net opposition**
Muslim%	1	0.43 [0.22 0.61]	0.57 [0.38 0.71]	-0.18 [-0.41 0.06]	0.66 [0.50 0.77]
Arrest rate	0.43 [0.22 0.61]	1	0.65 [0.48 0.77]	-0.86 [-0.91 -0.79]	0.55 [0.36 0.70]
Mean estimate	0.57 [0.38 0.71]	0.65 [0.48 0.77]	1	-0.18 [-0.40 0.07]	0.77 [0.65 0.85]
Estimate error	-0.18 [-0.41 0.06]	-0.86 [-0.91 -0.79]	-0.18 [-0.40 0.07]	1	-0.20 [-0.42 0.04]

[8] The survey question was: “"Thinking about people who want to come and live in the Netherlands from different countries, to what extent should people from the following countries be allowed to come and live in the Netherlands?”.

Net opposition	0.66 [0.50 0.77]	0.55 [0.36 0.70]	0.77 [0.65 0.85]	-0.20 [-0.42 0.04]	1

Table X. Correlation matrix between primary variables. Values in brackets are 95% confidence intervals. Net opposition is (none% + fewer% - same% - more%).

Mediation analysis was done using the **mediation** package for R. Results showed that an estimated 84% of the effect of actual crime rates to net opposition was mediated by the stereotypes, and if Muslim% was included as a covariate, this value was 85% (both mediation p's < .0001). As such, the prior findings are strongly confirmed (prior mediation % was about 100%). Next, we fit the path model with Muslim as an indepedent predictor of net opposition. Results are shown in Figure X.

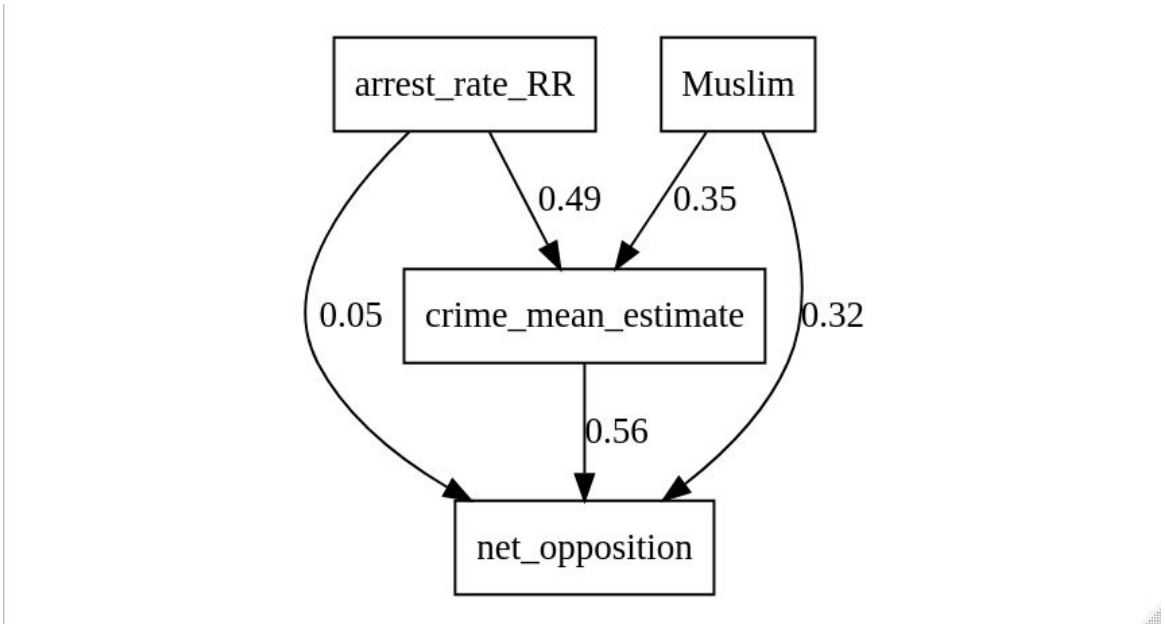


Figure X. Path model for immigration related variables. Paths are standardized. All variables except the direct path from arrest rate to net opposition have p < .001.

The path model shows what the mediation analysis finds, that the crime rate itself does not have much validity on net opposition directly, it's effect is through the stereotypes. In contrast to the prior study, we find a notable effect of Muslim% on net opposition, the prior study found this path to be p > .05 and with a beta of 0.03 (Kirkegaard et al., 2020). Thus, in the Dutch data, we see that the public is against Muslim immigration beyond the effect of the above-average crime rate of the more Muslim groups (r = .43).

The stereotypes and the immigrant preference data were collected from a partially overlapping subset of Prolific subjects. Though the data were collected months ago, this overlap might nonetheless bias results upwards due to the common method variance factor of being collected from the same subjects (Chang et al., 2010). If subjects who provided both data realized the link between them, which is a main hypothesis of this study, then they potentially made their responses more consistent with each other than if

asked independently. The prior study on Danish data looked for evidence of this and found none: the correlations between preferences and stereotypes were the same whether they were aggregated from the same subjects or not (Kirkegaard et al., 2020). In this study we took a further step, since we had collected data from two different pollsters, so we were able to compute two sets of stereotypes. These correlated r = .97. Rerunning the mediation analyses with only the stereotype data from Survee did not produce any notable changes.

Occupations and sex

For the sex differences in occupation, we scored the estimates in the same way as before, except that we left out the log-conversion approach. Figure X shows the scatterplot of the mean estimates and the true values, while Table X gives the summary statistics.

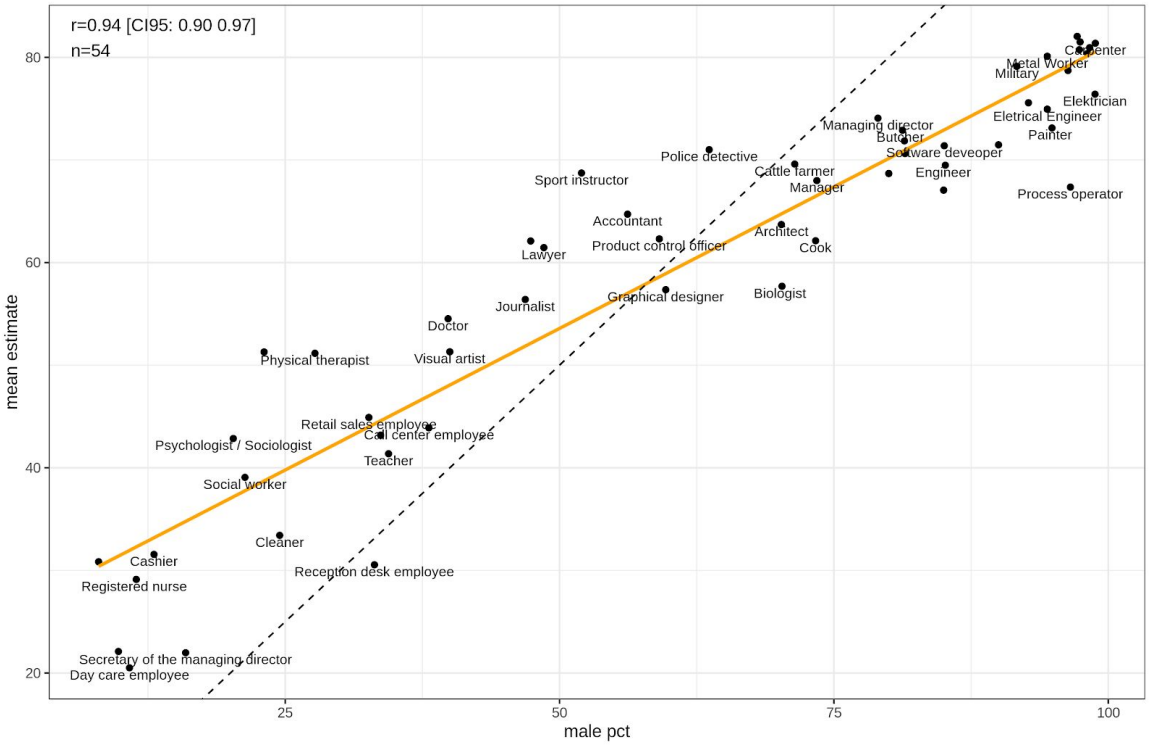


Figure X. Scatterplot of the mean stereotype of and real sex differences in occupation. The stippled line shows the slope = 1 of perfect calibration.

Method	Pearson r	Rank r	Mean abs error	SD	SD error	Mean	Mean error
mean	0.94	0.95	13.51	17.93	-12.62	59.97	-1.61
10% trimmed	0.94	0.94	12.77	19.03	-11.52	60.13	-1.45

mean							
median	0.94	0.94	12.36	19.54	-11.01	60.17	-1.41

Table X. Stereotype accuracy metrics for aggregated stereotypes of sex differences in occupation. The true mean/SD are 61.6/30.5.

The data shows near perfect accuracy in correlational terms, each method producing r = .94. Despite this accuracy, the MAE is not near 0, but is in fact around 13, meaning that the estimate is on average 13%points off the mark. The reason for this divergence is that the estimates are not extreme enough, suffering from a large negative SD bias of 12.6%points, or 41% in relative terms (12.62/30.5). In other words, the estimates substantially underestimate true variability between occupations. This aggregated error is larger than the individual-level one, where the median implied SD estimate was about 26% too small. The stereotypes did not differ by sex, as shown in Figure X.

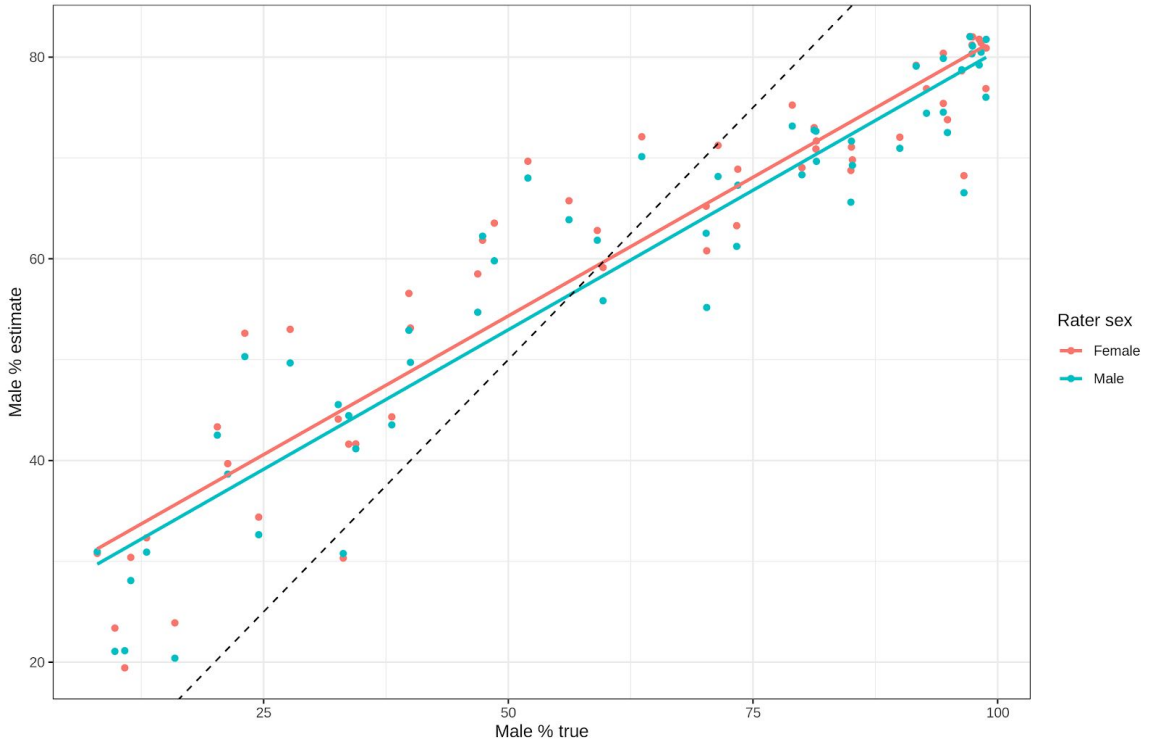


Figure X. Scatterplot of the mean stereotype of and real sex differences in occupation, as rated by men and women. The stippled line shows the slope = 1 of perfect calibration.

The stereotype accuracy metrics of the two set of estimates are essentially identical (e.g., Pearson r accuracies were .94 for both, SD error was -.12.6 for both), showing that stereotypes do not differ by sex. In the same, these two sets of estimates correlated at r = 1.00, regression slope = 1.00 (intercept -1.2, p > .05), so there was no statistical difference between them in any way. Men and women hold on average exactly the same stereotypes of the sex distribution of occupations.

Provincial incomes

Finally, we turn to the provincial incomes. As before, we scored these using 3 aggregation approaches, results shown in Table X, and Figure X shows the scatterplot of the true values and the mean estimates.

Method	Pearson r	Rank r	Mean abs error	SD	SD error	Mean	Mean error
mean	0.85	0.88	1573.71	2457.88	902.56	30382.05	1573.71
10% trimmed mean	0.87	0.88	910.96	2486.67	931.35	29477.87	669.54
median	0.79	0.76	1575.00	2454.12	898.81	30250.00	1441.67

Table X. Stereotype accuracy metrics of stereotypes of provincial income differences. True mean/SD are 28808/1555.

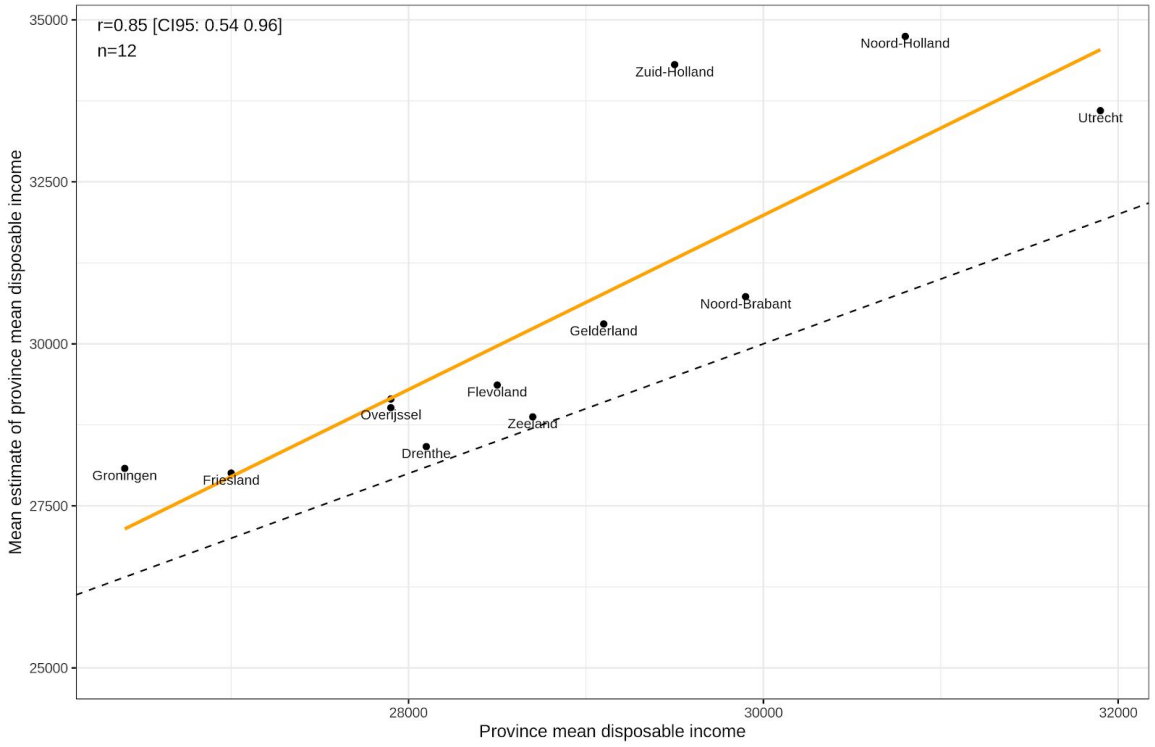


Figure X. Scatterplot of the mean estimate of and real provincial differences in disposable income. The stippled line shows the slope of 1, i.e., perfect calibration.

The overall correlational accuracy is very high, Pearson r = .85 for the mean estimates. The 10% trimmed estimates were slightly more accurate across the metrics. In contrast to the prior two sections, the provincial estimates were actually more dispersed than

reality, giving a positive SD error of about 900, or 58% too large. There was also a small positive mean error, i.e., estimates were too high on average by about 1600, or about 5%.

Discussion

This lengthy study had many findings of interest. First, we found strong accuracies overall. We find this across different measures of accuracy, across data sources, and across domains. This shows that stereotype accuracy is a strong, replicable and general phenomenon. This is furthermore in line with the large majority of other stereotype accuracy studies reviewed in reviews by Lee Jussim and colleagues across many years (Jussim, 2012, 2018; Jussim et al., 2009; Lee et al., 1995).

Second, we find that there are notable Muslim related biases in stereotypes. Going against popular narratives, we confirm prior findings, namely that biases are in favor of Muslim groups since the biases are towards *underestimating* the crime proneness of these groups (Kirkegaard et al., 2020; Kirkegaard & Bjerrekær, 2016).

Third, we found that stereotypes statistically mediate attitudes expressed towards immigration from the same groups in a rational way: groups with higher actual crime rates have stereotypes with higher crime rates, and face more opposition. Insofar as the public are trying to avoid increased crime in their countries, they appear to express crime-minimizing preferences. One British survey found that crime is the most important criterion that the public uses when evaluating which countries should be allowed to send immigrants (Carl, 2016).

Fourth, we found very clear evidence that stereotype accuracy metrics were predictable by intelligence, and this was also the case when we statistically controlled a large list of confounders including education level, age, sex, voting behavior and intentions, student status, employment status, and country of origin. The direct effect of intelligence was usually only somewhat smaller in the full regression model compared to the correlation, and thus the effect of intelligence was not mediated by any of these variables to a notable degree. This seems somewhat surprising. Variable selection methods, in our case lasso regression and Bayesian model averaging, also found that intelligence was always a useful predictor.

Fifth, accuracy of stereotypes was correlated across domains and across metrics, but surprisingly, not strongly so (correlations about r = .20, akin to typical item-correlations in an intelligence test). It was nonetheless possible to speak of a general factor of accuracy. Once extracted, this general factor showed a stronger relationship with intelligence (close to .40, without adjustment for reliability issues), as expected from psychometric theory. In a hierarchical or bifactor model of intelligence, one may posit various group factors, which are other broad abilities that are not general intelligence, but which contribute to the variation of performance on some subset of tests. Since stereotypes are essentially just a type of knowledge of demographics and regional

statistics, this is a kind of general knowledge, and thus may be subsumed under the previously identified knowledge factor (Carroll, 1993; Jensen, 1998; McGrew, 2009). Social psychology offers a number of older studies that also support such general factors, and a relationship to own ability (Landy & Farr, 1980):

> Several studies have found that the performance level of the rater affects the nature of the ratings assigned to others by that rater. D. E. Schneider and Bayroff (1953) and Bayroff, Haggerty, and Rundquist (1954) reported that peers who received high aptitude test scores and were rated positively during training gave ratings of their fellow trainees that were more valid in predicting subsequent job performance. Mandell (1956) found no difference in central tendency between good and poor job performers but did find that those raters who were poor performers tended to disagree more with consensus ratings of subordinates than did the more favorable performers. Kirchner and Reisberg (1962) found that the ratings given to subordinates by supervisors high in job performance were characterized by greater range, less central tendency, and by more emphasis being placed on the independent action of subordinates as the basis for ratings. In a related study Mullins and Force (1962) obtained evidence for a generalized ability to rate others accurately. Peer raters who were more accurate in judging one skill of their co-workers also were accurate in judging another performance dimension. (Accuracy was assessed by comparing the ratings with scores on pencil-and-paper tests.)

Sixth, overall, stereotype accuracy was not well-predicted despite a large collection of potentially relevant predictors. Typically, we were able to account for about 10% variance. One major part of this puzzle may be due to the unknown reliability of stereotypes. As far as we know, no prior study of stereotypes computed a test-retest correlation, so it is unknown whether individual-level stereotypes are simply unreliable, and that is why they are hard to predict well. If that is the case, this would mean the current findings about the role of intelligence are actually greatly underestimated, since adjusting for unreliability of the dependent variable would result in large increases in the betas of all reliable predictor variables. We suggest further research should clarify this question.

Supplementary materials

Supplementary materials, including data, R code, and figures, are available on the study's OSF repository: https://osf.io/aexk9/. The statistical output may also be found at https://rpubs.com/EmilOWK/Dutch_stereotype_study_2020.

References

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2016). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. Management Science, 63(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2016). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Caplan, B., & Miller, S. C. (2010). Intelligence makes people think like economists: Evidence from the General Social Survey. *Intelligence*, *38*(6), 636–647. https://doi.org/10.1016/j.intell.2010.09.005

Carl, N. (2014). Verbal intelligence is correlated with socially and economically liberal beliefs. *Intelligence*, *44*, 142–148. https://doi.org/10.1016/j.intell.2014.03.005

Carl, N. (2015). Cognitive ability and political beliefs in the United States. *Personality and Individual Differences*, *83*, 245–248. https://doi.org/10.1016/j.paid.2015.04.029

Carl, N. (2016). Net opposition to immigrants of different nationalities correlates strongly with their arrest rates in the UK. *Open Quantitative Sociology & Political Science*. https://openpsych.net/paper/48

Chang, S.-J., van Witteloostuijn, A., & Eden, L. (2010). From the Editors: Common method variance in international business research. *Journal of International Business Studies*, *41*(2), 178–184. https://doi.org/10.1057/jibs.2009.88

Coyle, T. R. (2018). Non-g Factors Predict Educational and Occupational Criteria: More than g. *Journal of Intelligence*, *6*(3), 43. https://doi.org/10.3390/jintelligence6030043

Deary, I. J., Batty, G. D., & Gale, C. R. (2008). Childhood intelligence predicts voter turnout, voting preferences, and political involvement in adulthood: The 1970 British Cohort Study. *Intelligence*, *36*(6), 548–555. https://doi.org/10.1016/j.intell.2008.09.001

Figueredo, A. J., McKnight, P. E., McKnight, K. M., & Sidani, S. (2000). Multivariate modeling of missing data within and across assessment waves. *Addiction*, *95*(11s3), 361–380. https://doi.org/10.1046/j.1360-0443.95.11s3.6.x

Flore, P. (2018). *Stereotype threat and differential item functioning: A critical assessment*. https://research.tilburguniversity.edu/en/publications/stereotype-threat-and-differential-item-functioning-a-critical-as

Flore, P., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Friedman, J., Hastie, T., Simon, N., Qian, J., & Tibshirani, R. (2017). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models* (2.0-13) [Computer software]. https://cran.r-project.org/web/packages/glmnet/index.html

Goenner, C. F. (2004). Uncertainty of the Liberal Peace. *Journal of Peace Research*, *41*(5), 589–605. https://doi.org/10.1177/0022343304045977

Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Twardy, C., Wang, J., & Pfeiffer, T. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science*, *7*(7), 200566. https://doi.org/10.1098/rsos.200566

Gorsuch, R. L. (2015). *Factor analysis* (Classic edition). Routledge, Taylor & Francis

Group.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A Conceptual Introduction to Bayesian Model Averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Jussim, L. (2012). *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy*. Oxford University Press.

Jussim, L. (2018). *The Accuracy of Demographic Stereotypes*. https://doi.org/10.31234/osf.io/beaq3

Jussim, L., Stevens, S. T., & Honeycutt, N. (2018). Unasked questions about stereotype accuracy. *Archives of Scientific Psychology*, *6*(1), 214–229. https://doi.org/10.1037/arc0000055

Kan, K.-J., Wicherts, J. M., Dolan, C. V., & van der Maas, H. L. J. (2013). On the Nature and Nurture of Intelligence and Specific Cognitive Abilities: The More Heritable, the More Culture Dependent. *Psychological Science*, *24*(12), 2420–2428. https://doi.org/10.1177/0956797613493292

Kirkegaard, E. O. W. (in review). *An examination of the openpsychometrics.org vocabulary test*. https://openpsych.net/forums/2/thread/236/

Kirkegaard, E. O. W. (2020a, January 31). Sesardić's conjecture: Preliminary evidence in favor. *Clear Language, Clear Mind*. https://emilkirkegaard.dk/en/2020/01/sesardics-conjecture-preliminary-evidence-in-favor/

Kirkegaard, E. O. W. (2020b, May 13). The verbal tilt model. *Clear Language, Clear*

Mind. https://emilkirkegaard.dk/en/2020/05/the-verbal-tilt-model/

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016). Country of origin and use of social benefits: A large, preregistered study of stereotype accuracy in Denmark. *Open Differential Psychology*. https://openpsych.net/paper/49

Kirkegaard, E. O. W., Bjerrekær, J. D., & Carl, N. (2017). Cognitive ability and political preferences in Denmark. *Open Quantitative Sociology & Political Science*, *1*(1). https://openpsych.net/paper/51

Kirkegaard, E. O. W., Carl, N., & Bjerrekær, J. D. (2020). Are Danes' Immigration Policy Preferences Based on Accurate Stereotypes? *Societies*, *10*(2), 29. https://doi.org/10.3390/soc10020029

Kirkegaard, E. O. W., & de Kuijper, M. (2020). Public Preferences and Reality: Crime Rates among 70 Immigrant Groups in the Netherlands. *Mankind Quarterly*.

Kuhn, M., Wickham, H., & RStudio. (2020). *tidymodels: Easily Install and Load the "Tidymodels" Packages* (0.1.0) [Computer software]. https://CRAN.R-project.org/package=tidymodels

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lyon, A., & Pacuit, E. (2013). The Wisdom of Crowds: Methods of Human Judgement Aggregation. In P. Michelucci (Ed.), *Handbook of Human Computation* (pp. 599–614). Springer. https://doi.org/10.1007/978-1-4614-8806-4_47

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126–132. https://doi.org/10.1038/s41562-017-0273-4

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments.

Journal of Behavioral and Experimental Finance, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2020). *BMA: Bayesian Model Averaging* (3.18.14) [Computer software]. https://CRAN.R-project.org/package=BMA

Sesardić, N. (2005). *Making sense of heritability*. Cambridge University Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=241083

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514–1534. https://doi.org/10.1037/apl0000420

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations* (1st ed). Doubleday.

Tran, U. S., Hofer, A. A., & Voracek, M. (2014). Sex Differences in General Knowledge: Meta-Analysis and New Data on the Contribution of School-Related Moderators among High-School Students. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0110391

Appendix

Attention checks

We placed 4 attention check questions throughout the survey. These asked the subject to select a given point using a slider. Attention was then scored simply as whether the subject had picked the right number or not. However, when we were collecting data from Survee, the pollster alerted us to the possibility that many subjects were failing by small amounts. The slider had a range of 100, and it was difficult to hit the exact number using mobile devices for some subjects. This issue was mostly evident for older subjects who are generally less technically competent. To get around this, we improvised a new attention check scoring where we computed the total deviance score per subject, defined

as the sum of many much of their select value on the sliders deviated from the value the attention check asked. Using this approach, a small deviance would not indicate lack of attention but rather trouble hitting the right value. A person who was not paying attention would result in very high deviance scores. Figure SX shows the validation of this approach using the total survey time use as indicator of inattentive responding.

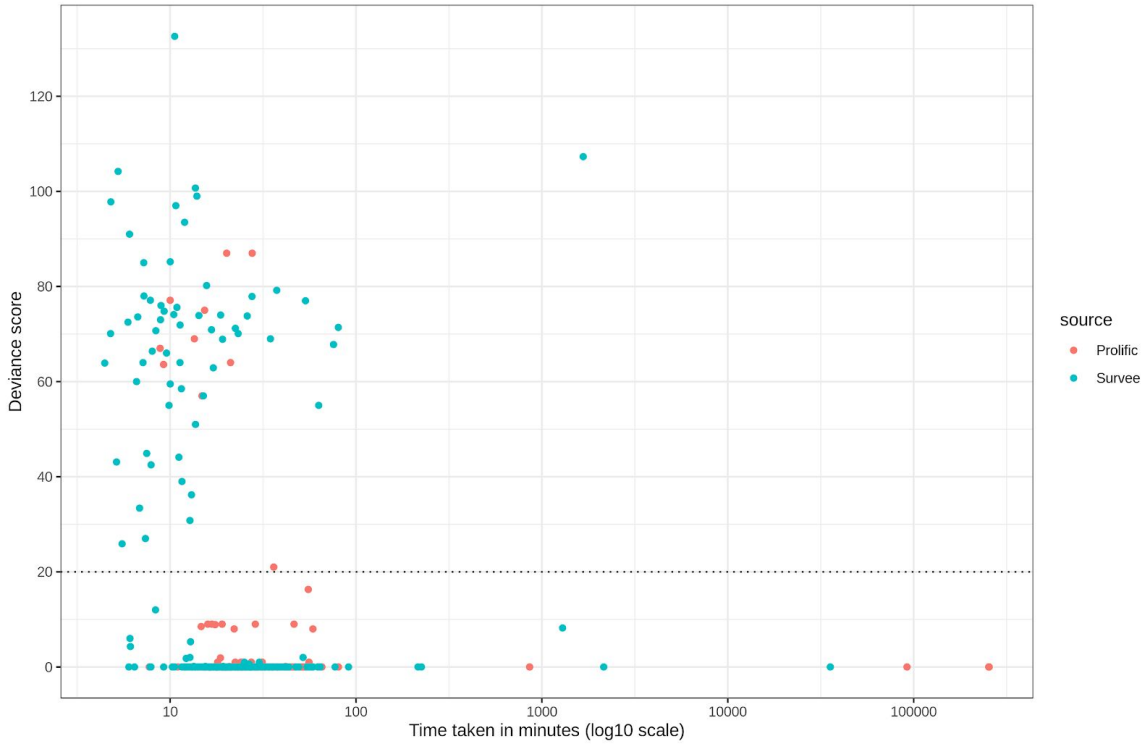
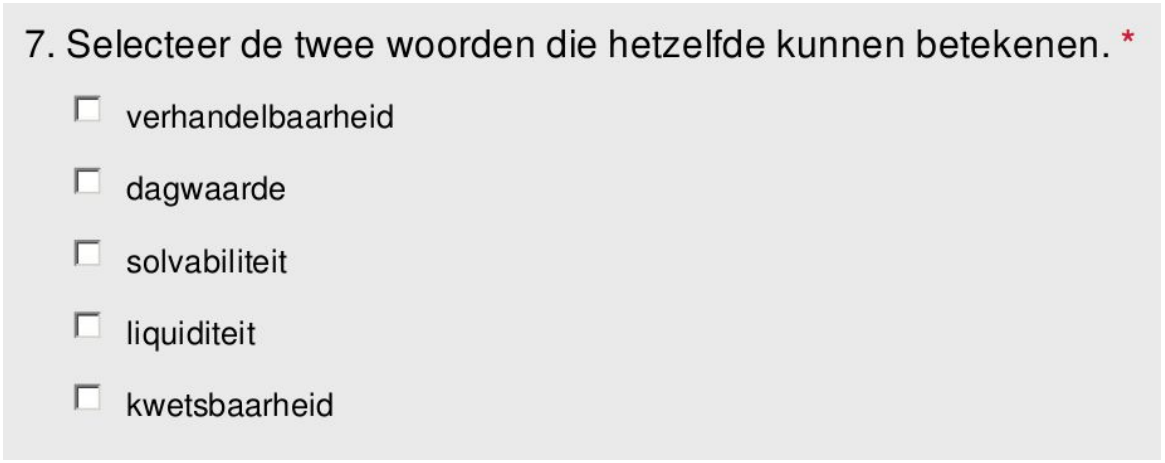


Figure SX. Time spent vs. deviance score by data source. Line fits by LOESS.

The figure shows that most people obtained deviance scores of 0. Those that did not were mostly persons who finished the survey quickly. By agreement with Survee, we used the threshold of deviance score 20 to delineate between attentive and inattentive subjects. This resulted in acceptance rates of 97.6% and 77.2% for Prolific and Survee subjects, respectively. Our planned method was more strict and resulted in rates of 91.9% and 71.6%, respectively. This method choice did not affect our results notably.

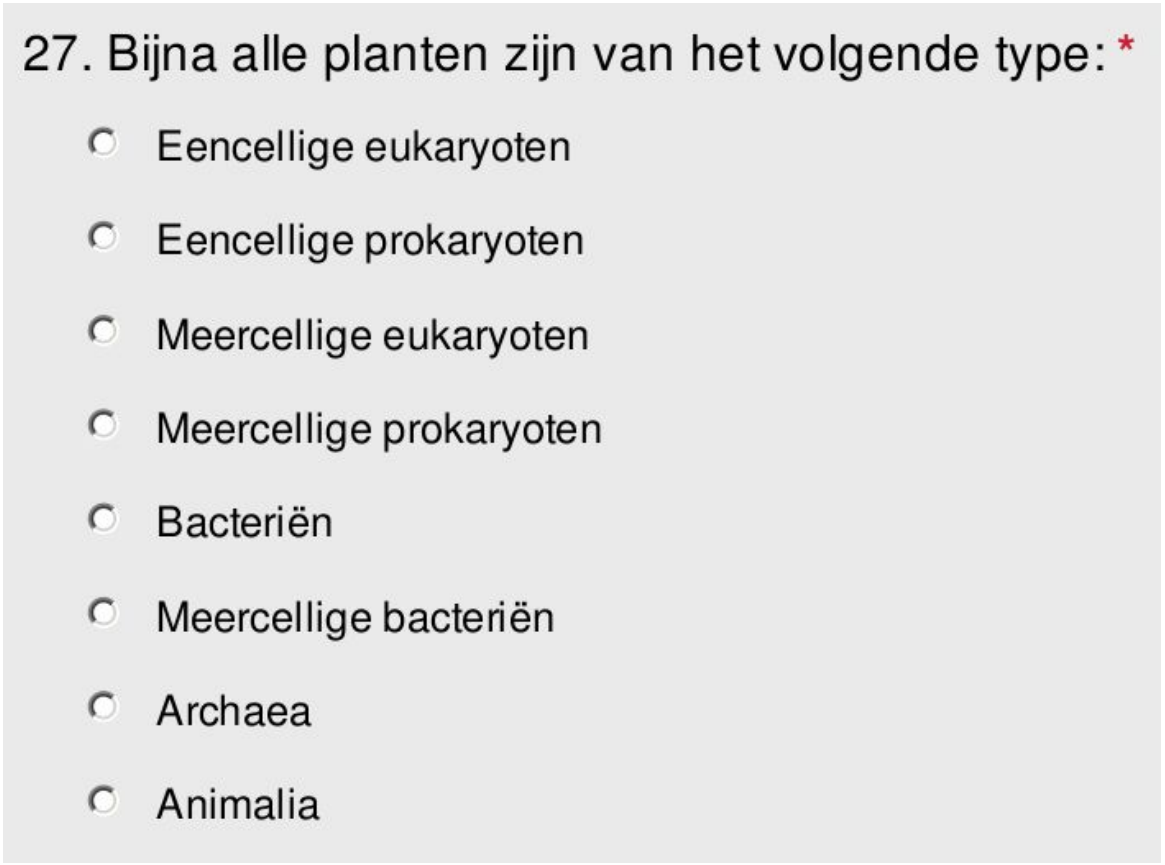
Intelligence example items

The first vocabulary item is:



(*Select the two words that could mean the same*, correct: (1, 3) verhandelbaarheid, liquiditeit; roughly meaning tradability, liquidity)

The first science knowledge item is:



(Question: *Almost all plants are of the following type*, correct: (3) Meercellige eukaryoten, multi-cell eukaryotes)

Both screenshots are from the exported survey file (*Questionnaire Prolific.pdf*).

Main correlation matrix

	g	V tilt	Age	Male	Time taken	Education	Student	Vote PvdD	Vote Groenlinks	Vote SP	Vote D66	Vote PvdA	Vote VVD	Vote Christenunie	Vote PVV	Vote CDA	Vote FvD	Vote SGP	Vote 50Plus	Vote DENK
g	1	0	0.21	0.08	0.17	0.36	-0.11	0.03	0.09	0.01	0.13	-0.01	0	0.01	-0.14	-0.04	-0.02	0	-0.07	-0.01
V tilt	0	1	0.42	-0.22	0.07	-0.04	-0.4	0.02	-0.06	0.14	-0.05	0.05	0.06	-0.05	0.1	0.12	-0.03	-0.05	-0.03	-0.02
Age	0.21	0.42	1	-0.07	0.12	-0.02	-0.75	-0.01	-0.14	0.17	-0.07	0.01	0.07	0.03	0.15	0.11	0.03	0.02	0.06	-0.03
Male	0.08	-0.22	-0.07	1	0.04	0.03	0.01	-0.11	-0.14	-0.14	0.05	0.04	0.12	-0.01	0.03	-0.08	0.12	0.02	-0.1	-0.13
Time taken	0.17	0.07	0.12	0.04	1	-0.01	-0.09	0.01	-0.04	0.01	-0.06	0	-0.01	0.01	0.05	-0.09	0	-0.01	-0.04	-0.02
Education	0.36	-0.04	-0.02	0.03	-0.01	1	-0.19	0.02	0.06	-0.08	0.12	-0.04	0.18	0.05	-0.12	-0.01	-0.08	-0.03	-0.08	0.02
Student	-0.11	-0.4	-0.75	0.01	-0.09	-0.19	1	-0.03	0.23	-0.2	0.1	0.04	-0.09	-0.01	-0.22	-0.21	-0.18	-0.01	-0.1	0.13
Vote PvdD	0.03	0.02	-0.01	-0.11	0.01	0.02	-0.03	1	-0.11	-0.03	-0.07	-0.06	-0.1	-0.05	-0.07	-0.04	-0.06	-0.01	-0.02	-0.01
Vote Groenlinks	0.09	-0.06	-0.14	-0.14	-0.04	0.06	0.23	-0.11	1	-0.13	-0.12	-0.12	-0.22	-0.09	-0.14	-0.09	-0.15	-0.03	-0.05	-0.02
Vote SP	0.01	0.14	0.17	-0.14	0.01	-0.08	-0.2	-0.03	-0.13	1	-0.11	-0.06	-0.12	-0.04	-0.07	-0.05	-0.06	-0.02	0	-0.01
Vote D66	0.13	-0.05	-0.07	0.05	-0.06	0.12	0.1	-0.07	-0.12	-0.11	1	-0.06	-0.15	-0.08	-0.1	-0.04	-0.12	-0.02	-0.02	-0.02
Vote PvdA	-0.01	0.05	0.01	0.04	0	-0.04	0.04	-0.06	-0.12	-0.06	-0.06	1	-0.11	-0.05	-0.08	-0.05	-0.08	-0.02	-0.03	-0.01
Vote VVD	0	0.06	0.07	0.12	-0.01	0.18	-0.09	-0.1	-0.22	-0.12	-0.15	-0.11	1	-0.07	-0.1	-0.06	-0.09	-0.02	-0.04	0.05
Vote Christenunie	0.01	-0.05	0.03	-0.01	0.01	0.05	-0.01	-0.05	-0.09	-0.04	-0.08	-0.05	-0.07	1	-0.05	-0.01	-0.04	0.05	-0.02	-0.01
Vote PVV	-0.14	0.1	0.15	0.03	0.05	-0.12	-0.22	-0.07	-0.14	-0.07	-0.1	-0.08	-0.1	-0.05	1	-0.05	0.01	-0.01	0	-0.01
Vote CDA	-0.04	0.12	0.11	-0.08	-0.09	-0.01	-0.21	-0.04	-0.09	-0.05	-0.04	-0.05	-0.06	-0.01	-0.05	1	-0.04	-0.01	-0.02	-0.01
Vote FvD	-0.02	-0.03	0.03	0.12	0	-0.08	-0.18	-0.06	-0.15	-0.06	-0.12	-0.08	-0.09	-0.04	0.01	-0.04	1	-0.02	-0.03	-0.01
Vote SGP	0	-0.05	0.02	0.02	-0.01	-0.03	-0.01	-0.01	-0.03	-0.02	-0.02	-0.02	-0.02	0.05	-0.01	-0.01	-0.02	1	-0.01	0
Vote 50Plus	-0.07	-0.03	0.06	-0.1	-0.04	-0.08	-0.1	-0.02	-0.05	0	-0.02	-0.03	-0.04	-0.02	0	-0.02	-0.03	-0.01	1	0
Vote DENK	-0.01	-0.02	-0.03	-0.13	-0.02	0.02	0.13	-0.01	-0.02	-0.01	-0.02	-0.01	0.05	-0.01	-0.01	-0.01	-0.01	0	0	1

Table X. Correlation matrix between all quasi-numerical variables. Latent correlations used whenever appropriate. More details are given in the statistical output.

Dutch political parties

Party	Party (English)	Leader	%	Seats	Position (Wikipedia)

VVD	People's Party for Freedom and Democracy	Mark Rutte	21.3%	33	Center-right
PVV	Party for Freedom	Geert Wilders	13.1%	20	Right-wing to far-right
CDA	Christian Democratic Appeal	Sybrand van Haersma Buma	12.4%	19	Center to center-right
D66	Democrats 66	Alexander Pechtold	12.2%	19	Fiscal: Center to center-right Social: Center-left
GL	GroenLinks	Jesse Klaver	9.1%	14	Center-left to left-wing
SP	Socialist Party	Emile Roemer	9.1%	14	Left-wing
PvdA	Labour Party	Lodewijk Asscher	5.7%	9	Center-left
CU	Christian Union	Gert-Jan Segers	3.4%	5	Fiscal: Center to center-left Social: Center-right
PvdD	Party for the Animals	Marianne Thieme	3.2%	5	Left-wing
50+	50PLUS	Henk Krol	3.1%	4	Center
SGP	Reformed Political Party	Kees van der Staaij	2.1%	3	Right-wing
DENK	Denk	Tunahan Kuzu	2.1%	3	Center-left to left-wing
FvD	Forum for Democracy	Thierry Baudet	1.8%	2	Right-wing to far-right

Table X. List of Dutch political parties by their performance in the 2017 general election (https://en.wikipedia.org/wiki/2017_Dutch_general_election).

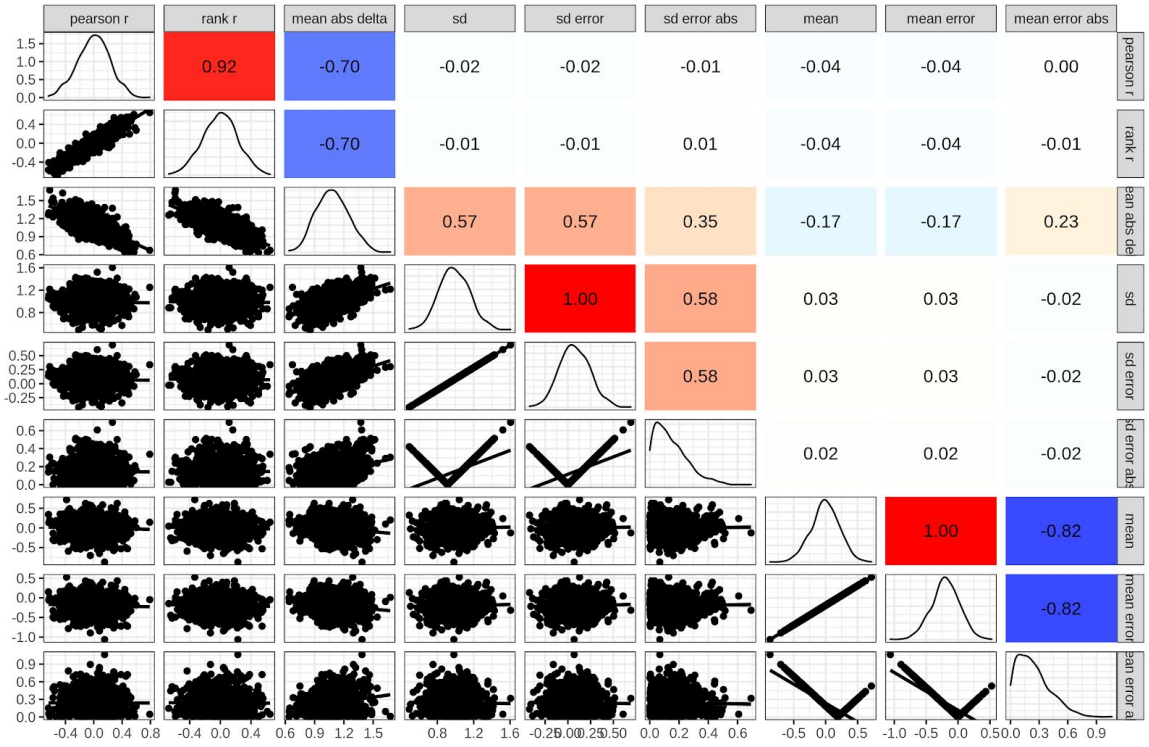
Accuracy metrics in simulated data

To get a better intuitive understanding of the accuracy metrics, it can be useful to simulate data under various conditions and examine the metrics and their interrelations. For this purpose, we simulated four datasets under the following conditions:

1. Random normal errors, mean=0, sd=1
2. Half signal + random normal errors, mean=0, sd=1
3. Half signal + random normal errors, mean=0, sd=varying between 0 and 2 per uniform distribution

4. Half signal + random normal errors, mean=varying per normal distribution, sd=varying between 0 and 2 per uniform distribution

Each simulation was based on n=1000 raters each rating 20 groups. The criterion data were a vector of random normal values (mean=0, sd=1) which has mean=0.19, sd=0.91. Thus, in the first case, all the data are completely random, and there is no signal even if data are aggregated. Any relations between accuracy metrics thus come only from their relatedness and coincidence. The simplest way to understand the data concisely, is to inspect the pairwise scatter and distribution plots, shown in Figure SX below. These are made using the **GGally** package. Table SX shows descriptive statistics.



SX. Pairwise scatter and distribution plot for simulation 1.

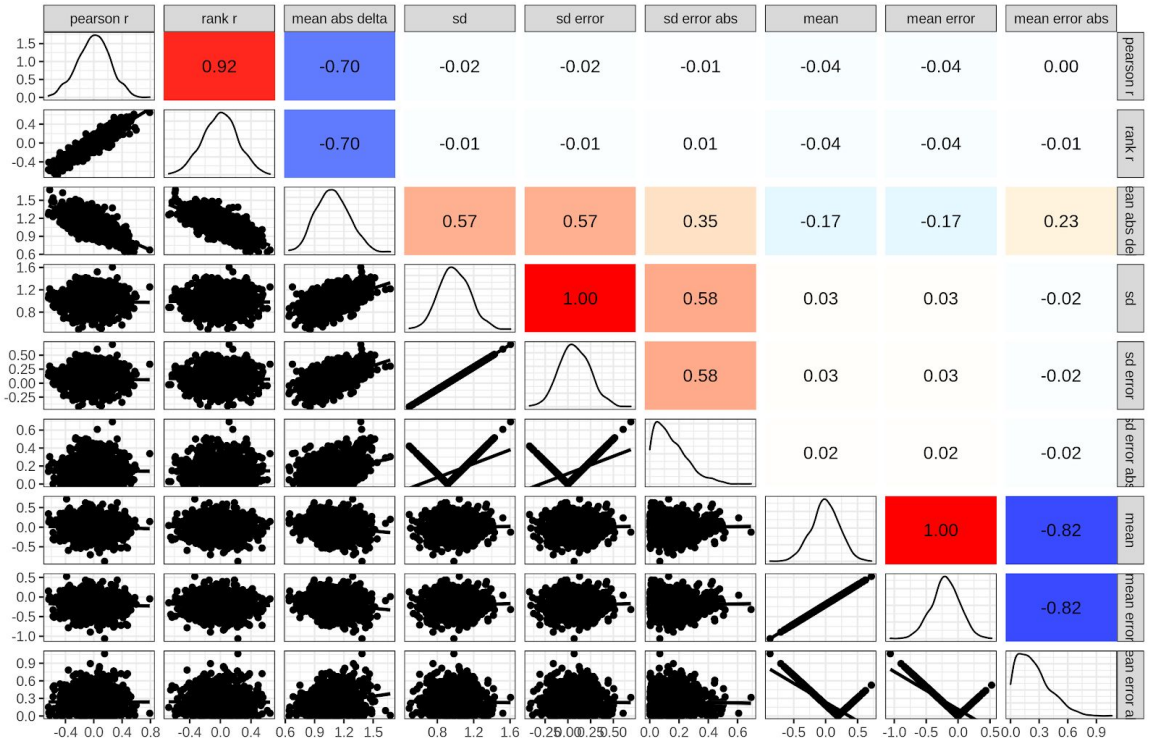
Variable	Mean	SD	Median	Mad	Min	Max	Skew	Kurtosis	Aggregate
pearson r	0.00	0.23	0.00	0.23	-0.64	0.79	-0.09	-0.13	-0.09
rank r	0.00	0.23	0.00	0.23	-0.67	0.65	-0.06	-0.18	-0.02
mean abs error	1.08	0.17	1.08	0.17	0.64	1.68	0.16	-0.20	0.75
sd	0.99	0.17	0.98	0.17	0.49	1.61	0.12	-0.02	0.03
sd error	0.08	0.17	0.07	0.17	-0.42	0.69	0.12	-0.02	-0.89
sd error abs	0.14	0.11	0.12	0.11	0.00	0.69	1.03	1.00	0.89
mean	-0.01	0.22	0.00	0.21	-0.87	0.72	-0.10	0.14	-0.01
mean error	-0.20	0.22	-0.19	0.21	-1.06	0.53	-0.10	0.14	-0.20
mean	0.24	0.17	0.21	0.17	0.00	1.06	0.89	0.58	0.20

error abs									

SX. Descriptive statistics for simulation 1 variables. Aggregate = the aggregated estimate's metrics.

With regards to the distributions, all the non-absolute values are approximately normally distributed, as reflected in their skew and kurtosis near 0. The variants with absolute values of course cannot have negative values, and thus they follow something close to a half normal distribution (which can be seen also in their skew and kurtosis). The plot also contains the (Pearson) correlations between each variable in the upper triangle. Thus we see that Pearson and Spearman (rank) correlational accuracy are highly correlated (r = .92) as might be expected. These are also strongly *negatively* related to the MAD (mean abs error), which is not surprising since having estimates further from their true values on average means the correlation will also be more negative. The relations to the mean and SD errors are approximately null, as these scale differences do not affect correlations. It is also worth noting that the measure based on means and SDs have some curious relations. The two pairs, mean+mean error, sd+sd error, show clean relationships, in fact, these are correlated at 1.00 since mean error = estimate mean - true mean (same for SD). The absolute version shows the characteristic V shape pattern reflecting the fact that over- and underestimating the mean/SD by 1 or -1 is the same amount of error in absolute terms. As there was no signal at all, the aggregated estimates do not show more favorable statistics either, having a correlation accuracy near 0 as well. It is worth noting that the SD error of the aggregated estimates is much smaller (0.03), since the random errors cancel out leaving only the signal, of which there is none.

In simulation 2, we add some true signal. Specifically, we add the criterion values times 0.5 (i.e., half signal) to the estimates along with the same random errors from before. Results are shown in Figure SX and Table SX.



SX. Pairwise scatter and distribution plot for simulation 2.

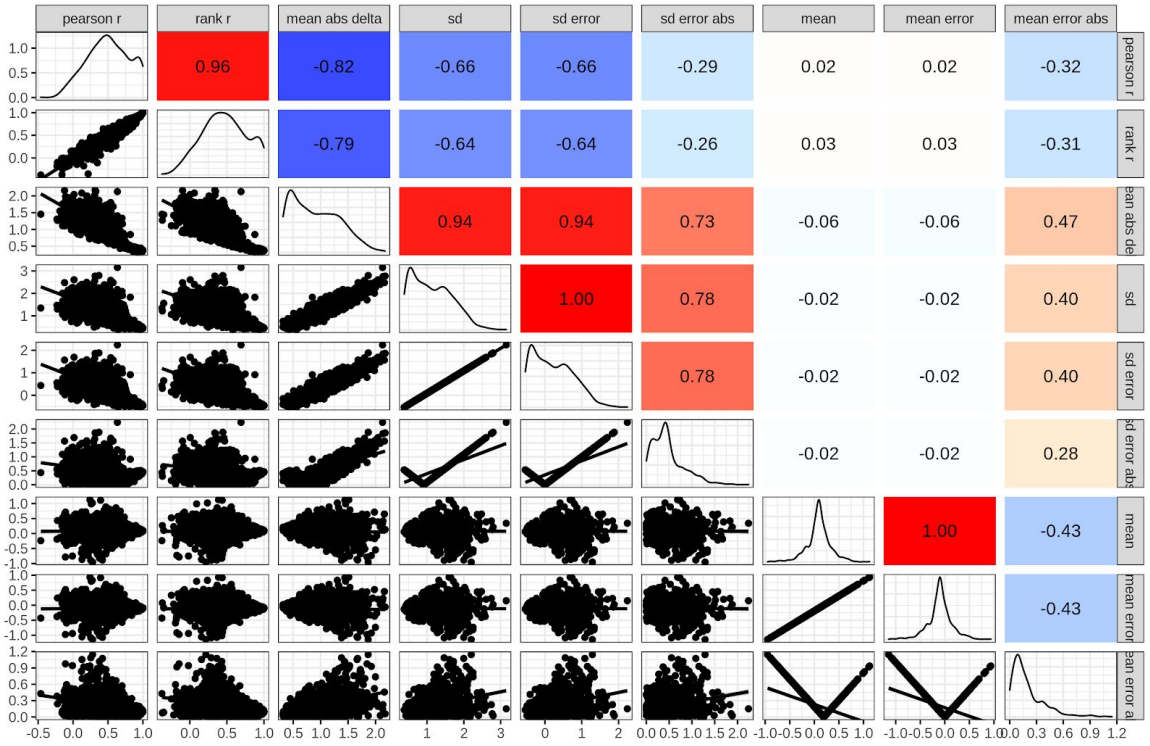
Variable	Mean	SD	Median	Mad	Min	Max	Skew	Kurtosis	Aggregate
pearson r	0.41	0.19	0.43	0.18	-0.30	0.88	-0.63	0.37	1.00
rank r	0.38	0.20	0.40	0.19	-0.32	0.80	-0.47	0.03	0.99
mean abs error	0.88	0.15	0.87	0.16	0.46	1.37	0.15	-0.23	0.38
sd	1.09	0.18	1.08	0.18	0.59	1.78	0.20	-0.09	0.45
sd error	0.17	0.18	0.16	0.18	-0.32	0.87	0.20	-0.09	-0.46
sd error abs	0.20	0.15	0.17	0.15	0.00	0.87	0.79	0.25	0.46
mean	0.09	0.22	0.09	0.21	-0.78	0.81	-0.11	0.14	0.09
mean error	-0.10	0.22	-0.10	0.21	-0.97	0.62	-0.11	0.14	-0.10
mean error abs	0.19	0.15	0.16	0.13	0.00	0.97	1.10	1.23	0.10

SX. Descriptive statistics for simulation 2 variables.

The new part about simulation 2 is mainly that we now see some average level of accuracy present, e.g. mean Pearson r is now .41 compared to .00 before. We now also see some paradoxical findings. For instance, correlational accuracy is now positively related to SD error abs, though it should in some sense be negative. The aggregate estimates are now essentially perfectly accurate in terms of correlations (.99 to 1.00), however they still suffer from downwards SD error. In fact, the SD of the aggregate estimates are half as large as they should be, and that is of course because we only

used half the signal strength in this simulation. The errors are otherwise random, so they canceled out leaving only the hal signal SD remaining.

In simulation 3, we add further realism, and allow for individual variation accuracy. This is done by varying the strength of the random error from SD=0 to 2 (uniform). Thus, individuals who have higher SDs have less relative signal in their values. Figure SX and Table SX show the results.



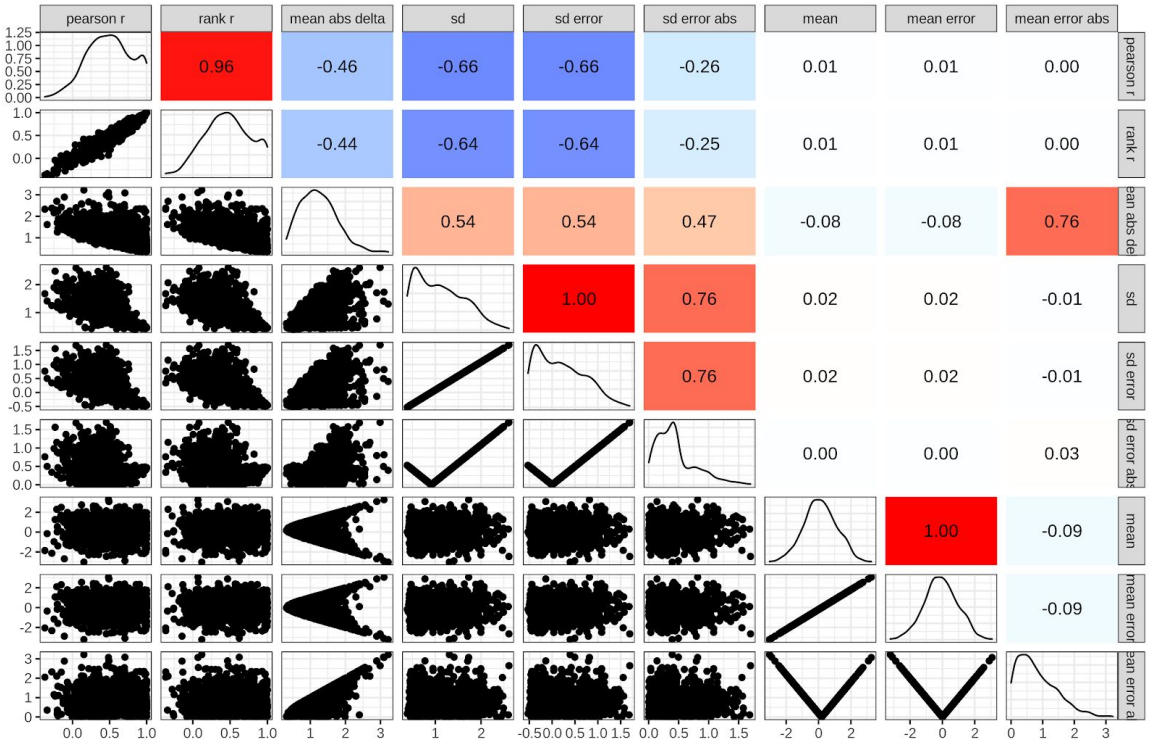
SX. Pairwise scatter and distribution plot for simulation 3.

Variable	Mean	SD	Median	Mad	Min	Max	Skew	Kurtosis	Aggregate
pearson r	0.50	0.30	0.50	0.33	-0.46	1.00	-0.14	-0.71	1.00
rank r	0.47	0.30	0.46	0.33	-0.38	1.00	-0.07	-0.69	1.00
mean abs error	0.92	0.43	0.86	0.53	0.31	2.17	0.47	-0.73	0.36
sd	1.14	0.54	1.07	0.64	0.38	3.15	0.50	-0.53	0.47
sd error	0.23	0.54	0.16	0.64	-0.53	2.24	0.50	-0.53	-0.45
sd error abs	0.47	0.35	0.41	0.31	0.00	2.24	1.19	1.62	0.45
mean	0.09	0.26	0.10	0.17	-0.95	1.13	-0.21	2.31	0.09
mean error	-0.10	0.26	-0.10	0.17	-1.14	0.93	-0.21	2.31	-0.10
mean error abs	0.20	0.19	0.14	0.12	0.00	1.14	1.82	4.01	0.10

SX. Descriptive statistics for simulation 3 variables.

The addition of the real variation in estimating skill across subjects made a large difference in that now there are appropriate negative correlations between the correlational metrics and SD error abs. Like before, the aggregate estimates are nearly perfect in terms of correlations, but suffer from the same SD error as before. The fact that some subjects have a larger error spread than others does not alter the fact that these cancel out across subjects.

Finally, in simulation 4, we add elevation errors to subjects, so that they both vary in their ability to get the elevation right and the dispersion right. The situation is now approaching reality. Figure SX and Table SX show the results.



SX. Pairwise scatter and distribution plot for simulation 4.

Variable	Mean	SD	Median	Mad	Min	Max	Skew	Kurtosis	Aggregate
pearson r	0.50	0.30	0.50	0.32	-0.37	1.00	-0.12	-0.64	1.00
rank r	0.47	0.30	0.47	0.32	-0.36	1.00	-0.06	-0.66	1.00
mean abs error	1.23	0.51	1.19	0.53	0.32	3.22	0.56	0.24	0.37
sd	1.12	0.52	1.06	0.62	0.38	2.61	0.51	-0.68	0.46
sd error	0.20	0.52	0.14	0.62	-0.53	1.70	0.51	-0.68	-0.46
sd error abs	0.45	0.33	0.38	0.28	0.00	1.70	1.13	0.98	0.46
mean	0.11	1.01	0.10	0.98	-3.03	3.26	0.00	-0.15	0.11
mean error	-0.08	1.01	-0.09	0.98	-3.22	3.07	0.00	-0.15	-0.08

mean error abs	0.81	0.61	0.68	0.61	0.00	3.22	0.90	0.44	0.08

SX. Descriptive statistics for simulation 4 variables.

We now see the V curve for the mean variables as well, showing the variation in mean estimate across persons and how this relates to the mean of the criterion values. Despite this added realism, the aggregate estimate is still perfect in terms of correlational accuracy. As a matter of fact, this never happens in real datasets because the true estimates do not simply consist of the criterion values and random errors. The real life errors are systematic, varied, and probably interrelated. People do not have a source of perfect knowledge about group differences in most cases, but rely on various proxies (shortcuts). For instance, it has previously been found that when people are asked to estimate immigrant groups' economic contributions, they seem to rely upon knowledge of the origin countries' wealth in terms of GDP per capita. The evidence for this comes from correlated errors between the estimates people produce and those produced from predicting from GDP per capita. See X for details.

Case representativeness method

To pick a representative (central, typical) case, we devised a simple method. In this method, the variables are first standardized, then the central tendency is subtracted (if it's not the mean), absolute values are taken, and finally a mean is taken. Thus, the value is how far on average the case differs from the central tendency across all the variables. By standardizing the variables, they are given equal weight, and by taking the absolute value, the variables are not allowed to offset each other (otherwise, negative distance to central tendency on one variable would cancel out with positive distance on another). To illustrate the method, we computed the principal components on the mpg (car) dataset in R. Figure SX shows the results with the most central cases marked.

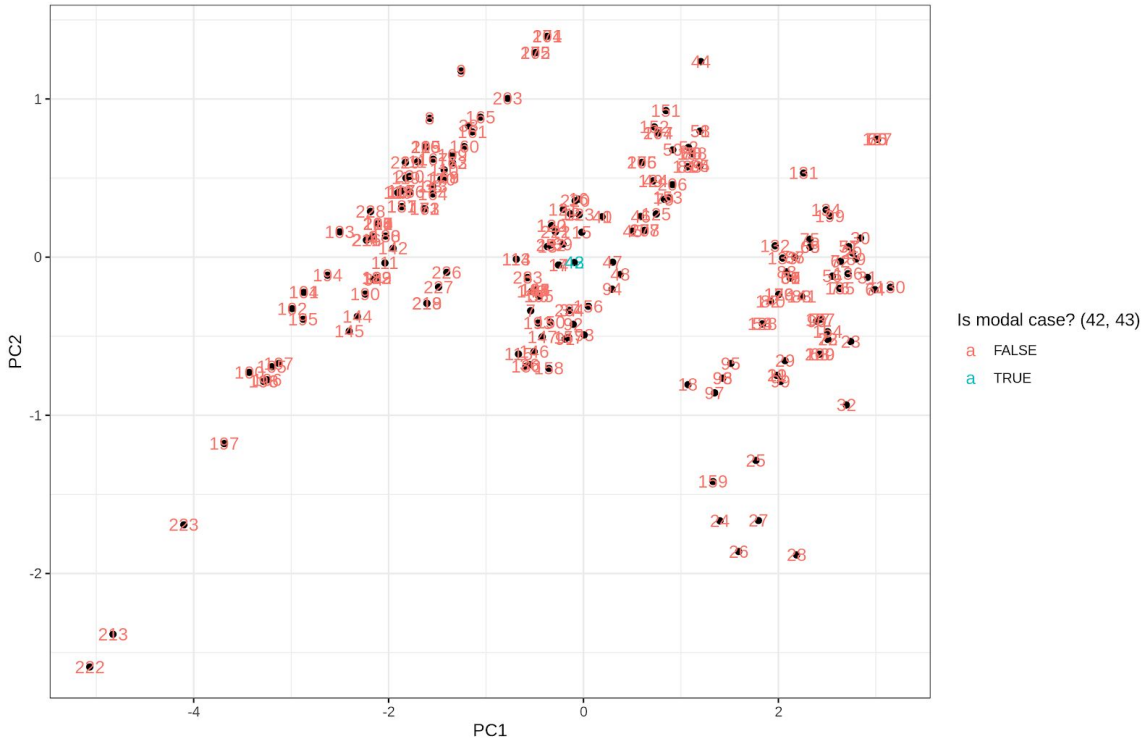


Figure SX. Central cases in mpg dataset. Scatterplot shows 1st and 2nd principal components.

As expected, the most central two cases (equally central) are roughly in the middle of the plot.