Submitted: 4th of July 2016
Published: 11th of November 2016

Country of origin and use of social benefits: A large, preregistered study of stereotype accuracy in Denmark

Emil O. W. Kirkegaard[*] Julius D. Bjerrekær[†]

Open Differential Psychology

Abstract

A nationally representative Danish sample was asked to estimate the percentage of persons aged 30-39 living in Denmark receiving social benefits for 70 countries of origin (N = 766). After extensive quality control procedures, a sample of 484 persons were available for analysis. Stereotypes were scored by accuracy by comparing the estimated values to values obtained from an official source. Individual stereotypes were found to be fairly accurate (median/mean correlation with criterion values = .48/.43), while the aggregate stereotype was found to be very accurate (r = .70). Both individual and aggregate-level stereotypes tended to underestimate the percentages of persons receiving social benefits and underestimate real group differences.

In bivariate analysis, stereotype correlational accuracy was found to be predicted by a variety of predictors at above chance levels, including conservatism (r = .13), nationalism (r = .11), some immigration critical beliefs/preferences, agreement with a few political parties, educational attainment (r = .20), being male (d = .19) and cognitive ability (r = .22). Agreement with most political parties, experience with ghettos, age, and policy positions on immigrant questions had little or no predictive validity.

In multivariate predictive analysis using LASSO regression, correlational accuracy was found to be predicted only by cognitive ability and educational attainment with even moderate level of reliability. In general, stereotype accuracy was not easy to predict, even using 24 predictors (k-fold cross-validated R^2 = 4 %).

We examined whether stereotype accuracy was related to the proportion of Muslims in the groups. Stereotypes were found to be less accurate for the groups with higher proportions of Muslims in that participants underestimated the percentages of persons receiving social benefits (mean estimation error for Muslim groups relative to overall elevation error = -8.09 %-points).

The study was preregistered with most analyses being specified before data collection began.

Keywords: stereotypes, stereotype accuracy, Denmark, immigrants, social benefits, group differences, Muslims, replication, preregistered, open data

1 Introduction

Stereotypes, that is, people's beliefs about groups[1], are often assumed to be exaggerated and inaccurate (Jussim, 2012). However, whether this is so is rarely examined. The existing body of research reveals that stereotypes are usually fairly accurate and rarely exaggerate real differences (Jussim, 2012; Jussim et al., 2015). Demographic stereotypes tend to be among the more accurate. As far as we know, only one prior (pilot) study has examined stereotype accuracy in Denmark (Kirkegaard & Bjerrekær, 2016a). The study was small (N = 48 after quality control), had a strongly unrepresentative sample but was preregistered. It found that stereotypes were fairly accurate (median correlational accuracy score = .51), but the results are hard to generalize to the overall population. The present study is a replication and expansion of the prior study using a large, nationally representative sample.

2 Methods

This study and most analyses were preregistered at `https://osf.io/wxqma/`. The setup closely fol-

[*] Ulster Institute for Social Research. Corresponding author, E-mail: emil@kirkegaard.dk

[†] E-mail: juliusdb.science@gmail.com

[1] We follow some researchers in using stereotypes to refer to people's beliefs about groups. In our opinion defining stereotypes as being inaccurate or exaggerated beliefs about groups results in a number of problems in talking about the essential issues. For a longer discussion of this, see Chapter 15 in Jussim (2012).

lows that of the pilot study (Kirkegaard & Bjerrekær, 2016a).

2.1 Participants

Participants were recruited with the help of a pollster (http://www.survee.dk/). A sample size of 500 was planned because this was judged to give sufficient power to estimate small effects. For instance, a correlation of .20 would have a standard error of about .044.[2]

2.2 The questionnaire

We designed a questionnaire to measure both stereotypes and other variables that may be related to stereotypes. The questionnaire as well as an English translation are available in the supplementary materials (*questionnaire.pdf* and *questionnaire_en.pdf*).

The structure of the questionnaire was as follows:

- Page 1 – Introduction:
  This included telling the participants about the purpose of the study, namely to find out how accurate stereotypes about immigrants are.

- Page 2 – Stereotype control questions:
  Two questions about easy and uncontroversial stereotypes (height for male/female, European/East Asian). These serve as control questions that the participant has understood the task and is not just clicking thru at random.

- Page 3 – Stereotypes:
  Detailed instructions, then followed by a list of 70 countries of origin. The countries that no longer exist have brief explanatory notes about this (e.g. Yugoslavia). The order of the countries is randomized for each participant to prevent order effects. The exact formulation (in the last wave) was: “People who live in Denmark originate from many countries. We would like you to estimate how many people among the 30-39 year olds that you think are on public assistance, from each country of origin.” (English translation, see project files for the original Danish.)

- Page 4 – Cognitive test:
  The 5-item ICAR test. This test was previously validated in a sample of children (Kirkegaard & Bjerrekær, 2016b).

- Page 5 – Political ideologies:
  Participants were asked to rate themselves (0-100) on four scales: conservatism, nationalism, economic liberalism and personal liberalism.[3]

- Page 6 – Political parties:
  Participants were asked to rate their agreement (0-100) with each of the main political parties in Denmark (15 parties), including parties outside parliament.[4] They had the option of stating that they had no preference or no knowledge about a particular party. This was because the list included relatively unknown parties outside parliament.

- Page 6 – Political opinions related to immigrants and experience with ghettos:
  Participants were asked to rate their agreement (0-100) with statements related to immigrants and immigration policy:

  1. whether to expel criminal immigrants,
  2. whether to expel immigrants that cannot take care of themselves,
  3. whether to limit social welfare to immigrants who have lived in Denmark for at least 5 years,
  4. whether Muslims constitute a special problem,
  5. whether immigrants will fare significantly better in Denmark in 20 years, and
  6. whether Denmark is for Danes.

- Page 7 – Optional contact info:
  Optionally, the participants could give their emails if they wish to receive follow-up information about this survey. They could also give comments in a comment field.

Since many readers will not be familiar with the Danish political parties, Table 9 in the appendix gives an empirical description of each party based on the political ideology and political opinion questions in this survey.

2.3 The criterion data

To score stereotypes for accuracy, one must know the real values (Jussim, 2012, Chapter 11). An earlier study examined immigrant groups in Denmark and

[2] The standard error of a correlation is $\sqrt{\frac{1-r^2}{n-2}}$ (Cohen & Cohen, 2003, p. 42).

[3] Economic liberalism is free-market orientation, while personal liberalism concerns the liberty to do actions that do not (seem to) harm others but which they may consider unethical, e.g. smoking cannabis or having sex with prostitutes or same-gendered persons.

[4] The list was copied from a list of active parties that the first author maintains (http://emilkirkegaard.dk/partier_paa_sociale_medier/). However, by accident a few major parties outside parliament were omitted.

had for that purpose bought data from the official Danish statistics bureau (Kirkegaard & Fuerst, 2014). This data was used for the present study. As with the pilot study, we used percentage of the population receiving social benefits among 30-39 year olds as our variable. This variable was chosen because it is relatively easy to understand and had a simple scale of possible values (0-100 %). The middle-aged group was chosen to avoid interpretation problems related to the educational grant that all Danish students are eligible for. Very few persons in their 30s are under formal education.

The comparison data concerns the year 2012, which is about 4 years prior to the collection of the stereotype data. There are probably small changes in the proportions of persons receiving social benefits since 2012. Furthermore, the samples were not always large so there may be some 'sampling error'.[5] For these reasons, the stereotype accuracy scores in this article are likely to be small underestimates of the real values.

3 Quality control and data collection

The total number of survey replies was 766. The data was analyzed as it become available to spot problems early. There could be no stopping-related p-hacking because the sample size was planned and because the study did not use null significance testing (Simmons et al., 2011).

3.1 *Missing data*

For most items, it was not possible to skip them. However, the items concerning agreement with the political parties were skipable and some chose to do so. The reason for allowing this was that it is meaningless to give an agreement rating with a party one is not familiar with. Additionally, for one user the gender and age information was missing from the pollster.

3.2 *Control questions*

The participants were paid a small sum to take part in the survey by the pollster. However, some participants filled out the questionnaire with nonsense information. To guard against this, we utilized two initial control questions related to uncontroversial stereotypes (height of men vs. women, or Europeans vs. East Asians). These also guard against people who genuinely do not understand the instructions.

Table 1: Participants by collection wave.

Wave	Freq	Proportion
1	85	0.11
2	32	0.04
3	169	0.22
4	51	0.07
5	112	0.15
6	317	0.41

3.3 *Data collection*

The data were collected in April-June, 2016.

We collected data in waves and the data were analyzed after each wave. This was done to make sure that the data quality was sufficiently high and to fix problems under the collection, rather than having to deal with them only after having collected all the data. Table 1 shows an overview of the number of responses collected in each wave.

3.3.1 *Wave 2*

After collecting wave 1, we noticed that some persons had given very small and seemingly random estimates to all groups. The first wave used a slider where participants could simply click the values they wanted to use. This also made it easier to quickly fill out random values that were not easily identifiable as such. The reason sliders make it difficult to spot fake values is that users clicking the sliders while scrolling down the page do not select exactly the same values every time due to the imprecision of the slider.[6] For this reason, we changed the sliders to text-boxes, so as to make it easier to spot unserious estimates. Text boxes make this easier because users who wish to quickly fill out all the boxes usually put the same (single digit) value in all boxes. We also noted that some users' estimates were opposite of reality, indicating that they did not understand the task or that they were purposefully answering dishonestly.

Furthermore, we added three control questions among the list of countries. These were very easy arithmetic questions such as “What is two plus two?”. Users who fail these questions could then easily be filtered afterwards. Additionally, we changed the order of the cognitive items to be random so that the order of presentation could not have a systematic effect on the responses.

3.3.2 *Wave 3*

After collecting wave 2, one comment stated that the s/he was unsure whether retirement counted as a

[5] Technically, since these data concern the entire population of relevant persons, there cannot be sampling error. However, there can be year to year fluctuations akin to sampling error.

[6] E.g. users intending to pick a low number would fluctuate around values of 5.

social benefit or not. It is in fact counted as a social benefit. However, because the age group is 30-39, it is legally impossible to be retired. However, to avoid this possible misinterpretation, we added more mentions of the age group in the instructions so that participants would not overlook this information.

3.3.3 Wave 4

After collecting wave 3, we noticed that two participants had left comments signaling that they were in doubt whether the estimates were supposed to be about immigrants living in Denmark grouped by country of origin, or persons living in the countries. While this interpretation was inconsistent with what we had written, we decided to rephrase the instructions to explicitly clarify this. We also added a fictive example.

3.3.4 Wave 5

After collecting wave 4, we saw that some users were still filling out the estimates in a way that had a reverse relationship to reality and most of the other participants. To try to find out why they did this, we created a second, brief questionnaire that was only sent to users identified as having given reverse patterns (more on this below). This questionnaire had an example of a person giving estimates that are opposite of reality. Then it asked the participant three questions:

(1) Have you given answers that are the reverse of reality as in the example above? [yes, no, don't know]

(2) If you have given reversed answers, why have you done so?

- I had misunderstood the question. Please write below how you had misunderstood it.
- To combat negative stereotypes about some groups.
- To make research about stereotypes difficult.
- For fun.
- Other. Please explain in the comment field below.

(3) Please describe how you made your estimates. [comment field]

The questionnaire did not state that the person who received it was identified as having made reverse estimates because we did not want them to know we were singling them out for further analysis. The results from this questionnaire indicated that most people who gave reverse answers had misunderstood the task as being about persons in the home country, not immigrants from that country in Denmark. One person answered that they deliberately answered dishonestly to cancel out others' stereotypes.

Furthermore, we added three more questions at the end of the main questionnaire:

(1) How precise do you think your stereotypes are? [0-100 scale]

(2) Have you filled out the stereotypes honestly? [yes, no]

(3) If you have given other answers than your best estimates, why have you done so? [only given to those who answer no to (2)]

- To combat negative stereotypes about some groups.
- To make research about stereotypes difficult.
- For fun.
- Other. Please explain in the comment field below.

People who answered no to (2) were tagged as *dishonest*. Of course, a person trying to influence the results could lie to these extra questions and so they cannot be used to filter out all participants that answer dishonestly.

3.3.5 Wave 6

After collecting wave 5, we saw that some were still giving reverse estimates. To further test whether people were understanding the estimation task, we added a question before the estimation task asking participants whether they are going to rate the countries after:

(1) People living in Denmark that hail from a given country or whose parents do so.

(2) Persons living in that country.

Anyone who answered (2) was sent to an extra ad hoc page explaining them again that this is not the task. Whether a user had failed this extra control question was saved to the dataset *(failed to understand)*.

3.4 Exclusions

Four exclusion criteria were established prior to data collection:

1. Having substantial missing data.

2. Failing any of the two stereotype control questions (*height gender* and *height race*).

3. Having an unlikely or impossible age.

4. Giving the same estimate for every group (*sd*).

As mentioned earlier, we added three more control questions to guard against participants giving random or unserious estimates and one where we simply asked people if they gave honest answers. Two additional exclusion criteria were thought of:

1. whether participants gave 0 % as an estimate for any group (*has 0*), and

2. the same for 100 % (*has 100*).

These may represent participants that have misunderstood the question because values of 0 % or 100 % are almost nonexistent for any large group of persons. Furthermore, one might exclude participants who score below a given threshold level of accuracy on account of them either having misunderstood the task or deliberately trying to sabotage the research (see Section 3.4.1). We examined whether each participant should be excluded according to each criteria. The exclusion criteria intercorrelations are shown in Table 2. Time spent added for comparison purposes.

Positive correlations indicate that persons who failed one control also tended to fail the second. It can be seen that for the first 7, there are only positive intercorrelations and furthermore each of them correlate negatively with time spent (people who failed controls also spent less time on filling out the questionnaire). Finally, the group with reversed answers did not generally fail the control questions, so they were paying attention. They were, however, more likely to use the implausible values of 0 or 100, and they did fail the comprehension question somewhat more often (r's .22 to .26).

For our analyses, we excluded all participants who failed one of the first 7 controls or who gave reversed answers (see below).

With regards to the question about which group of persons to base estimates on, 23 % of the participants failed that question. Among the participants that were not excluded for any other reason, this number was 19 %. Furthermore, the group of people who failed to understand had lower correlational accuracy[7], even after having gotten the task explained an extra time (mean r's .45 vs. .34 [CI95: .03 to .19], d = .57). This may indicate that they still did not understand the task or that whatever traits dispose one to fail to understand the task in the first try also dispose one have poorer stereotype accuracy).

Table 2: Latent correlations among exclusion criteria (Uebersax, 2015). The logarithm of time spent used due to a long right tail.

	height gender	height race	arithmetic1	arithmetic2	arithmetic3	dishonest	sd	has 0	has 100	reversed	failed to understand	time spent
height gender	1.00	0.51	0.42	0.53	0.39	0.25	0.14	0.10	0.16	−0.06	−0.01	−0.33
height race	0.51	1.00	0.60	0.66	0.64	0.18	0.33	0.02	0.32	−0.19	0.26	−0.47
arithmetic1	0.42	0.60	1.00	0.98	0.98	0.45	0.76	−0.02	0.13	−0.73	0.11	−0.60
arithmetic2	0.53	0.66	0.98	1.00	0.96	0.40	0.70	0.07	0.18	−0.69	0.15	−0.55
arithmetic3	0.39	0.64	0.98	0.96	1.00	0.37	0.63	−0.06	0.31	−0.70	0.14	−0.56
dishonest	0.25	0.18	0.45	0.40	0.37	1.00	0.28	−0.01	0.06	−0.14	0.17	−0.19
sd	0.14	0.33	0.76	0.70	0.63	0.28	1.00	−0.06	−0.49	−0.71	−0.04	−0.59
has 0	0.10	0.02	−0.02	0.07	−0.06	−0.01	−0.06	1.00	0.24	0.26	0.07	−0.12
has 100	0.16	0.32	0.13	0.18	0.31	0.06	−0.49	0.24	1.00	0.23	0.14	−0.08
reversed	−0.06	−0.19	−0.73	−0.69	−0.70	−0.14	−0.71	0.26	0.23	1.00	0.22	0.07
failed to understand	−0.01	0.26	0.11	0.15	0.14	0.17	−0.04	0.07	0.14	0.22	1.00	−0.14
time spent	−0.33	−0.47	−0.60	−0.55	−0.56	−0.19	−0.59	−0.12	−0.08	0.07	−0.14	1.00

[7] This accuracy score is simply the Pearson correlation between a person's estimates and the criterion values. See Section 7.

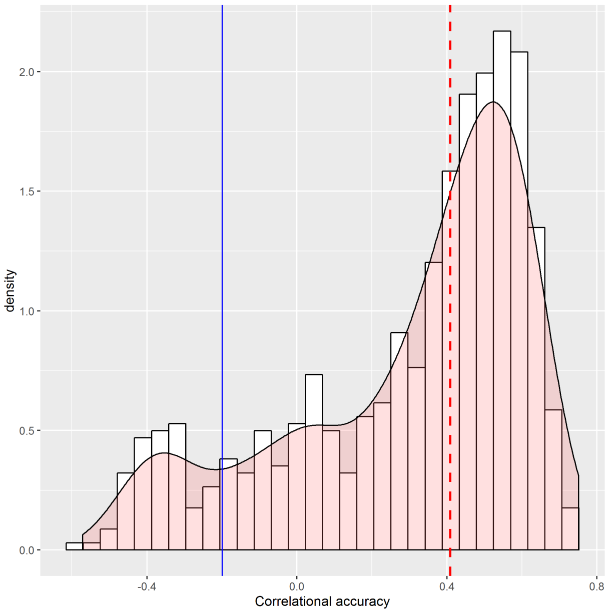

Figure 1: Distribution of correlational accuracy pre-exclusion. The red line is the median. The blue line is the cut-off for reversed answers.

3.4.1 *Reverse answers*

What constitutes a reverse answer? Consider the distribution of correlational accuracy scores in Figure 1.

One can obtain negative accuracy correlations by chance if one has no accuracy at all, but how likely is that? We took a simulation approach to deciding the cutoff to be used for inclusion in the reverse cluster. We simulated 1e5 sets of 70 random estimates based on the uniform distribution (numbers 0-100 are equally likely). Then we scored these for accuracy and calculated the centiles of the correlational accuracy distribution. We used the 5th centile of this distribution as our cutoff. If one answers at random, there is only 5 % chance of getting a more negative correlational accuracy score than this. This value turns out to be very close to -.20.

We recognize that some may not accept our exclusion method and may feel it is an overcorrection. We agree that a choice was made and provide the full data so that others may re-analyze the data using their own exclusion criteria.

4 Cognitive ability

Cognitive ability was measured with the 5-item ICAR5 test, which was previously validated on a sample of Danish students (Kirkegaard & Bjerrekær, 2016b). This test was chosen because it is in the public domain and can therefore be used for free for any purpose by anyone, already has an existing Danish translation that has been validated, and takes only a short time to complete. The validation study showed that the test was too difficult for younger children (below approximately age 14) but that is not a problem with this study because the youngest person in the current sample was 18.

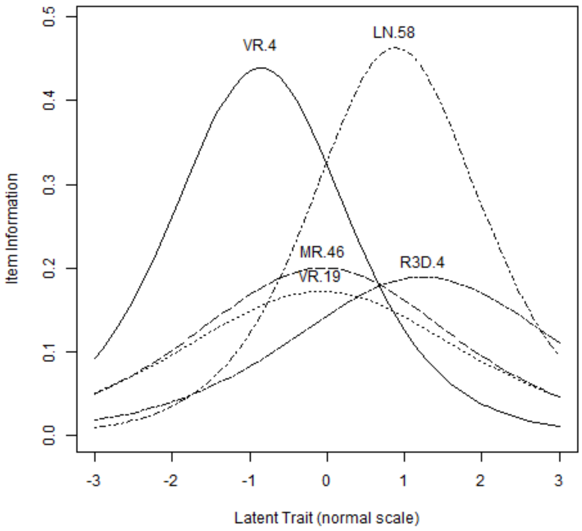

Figure 2: Item information plot.

The items were analyzed using item response theory-based (IRT) factor analysis using a 2 parameter model based on the normal distribution, 2PN; (Revelle, 2016)).[8] Figure 2 shows the item information plot.

All items had some discriminative ability (y-axis) and they were fairly spread out on the latent trait, meaning that the test did not have marked floor or ceiling effects for this sample. This can more clearly be seen in the test information plot, which is merely the summed version of the item information plot shown in Figure 3.

The test was scored using IRT. This was chosen over simple sums because these scores are theoretically better. Using simple sums means getting a harder item (those further on the right in Figure 3) right counts just as much as getting an easier item right. Using scoring based on the 2PN model, the item difficulty and discriminative ability (akin to a factor loading in classic factor analysis) is taken into account such that getting a harder item right counts for more than an easier item. After the scores were calculated, they were standardized (mean = 0, SD = 1).

5 Descriptive statistics

Table 3 shows basic descriptive statistics for the numerical variables. This includes ordinal variables

[8] The function is *irt.fa* in the **psych** package (Revelle, 2015).

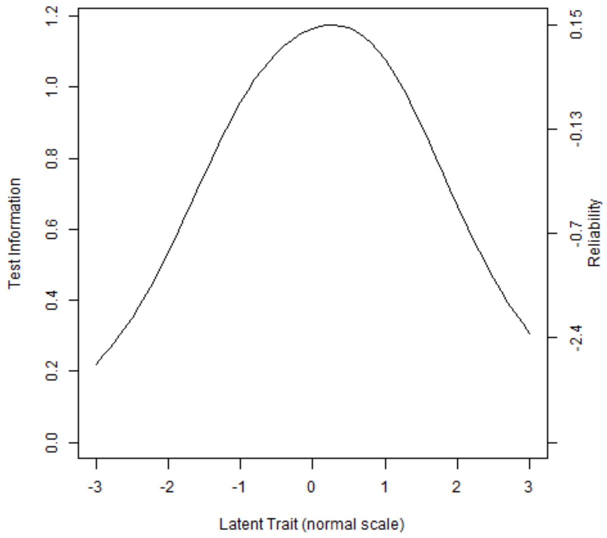

Figure 3: Test information plot.

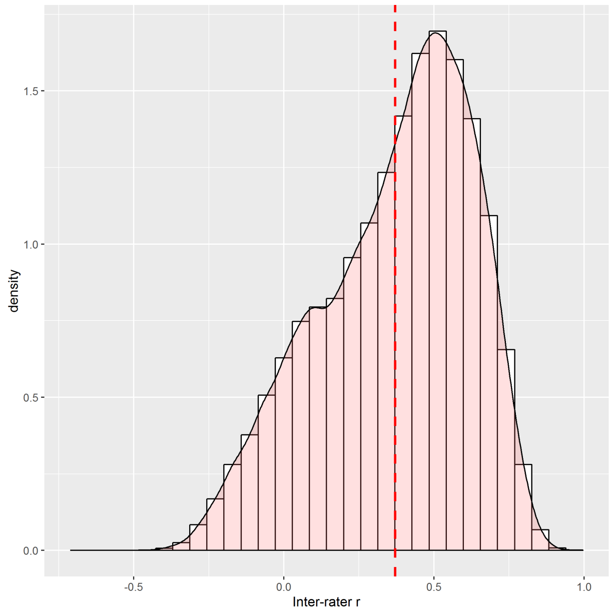

Figure 4: Distribution of inter-rater agreement. Vertical red line = mean.

that can be taken as a numerical variable without too much bias.

The abbreviations are agreement with each of the political parties, which are described in the appendix. With regards to representativeness, we only sampled persons aged 18-80. We downloaded data from the Danish statistics agency to calculate what the mean age is for this group, which was 47.1.[9] 53 % of the sample was female while for Denmark this number is 50.3 %. For educational attainment, 13.6 % of our sample had a university degree and for the general population this number is 11.7 %.[10] Thus, our sample was fairly representative, but somewhat younger and slightly better educated.

With regards to agreement with the statements about immigrants, there was broad agreement that criminals should be expelled (mean/median agreement = 82/92 %). On the other hand, Danes are not very ethnonationalist in that most disagree with the statement that Denmark is for Danes (mean/median agreement = 27/18 %). For the remaining statements, there was little agreement among Danes (mean agreement close to 50 %).

With regards to the political parties in parliament (N = 9), there was a strong positive relationship between stated agreement with each party and the parties' number of seats in parliament (r's = .79/.46 based on mean/median agreement, respectively). Our interpretation of this result is that it is due to our sample being fairly representative in terms of political opinion and the multi-party democracy in Denmark doing a fairly decent job of representing people's preferences.

6 Inter-rater agreement

Figure 4 show the distribution of inter-rater correlations.

The inter-rater correlation is moderate to large (mean/median .37/.41). The intraclass correlation (type 2) was .19/.99 at the individual/average level. This represents a fairly low level of inter-rater compared to the benchmarks values given by e.g. Cicchetti (1994). According to his guidelines, an ICC below .40 represents poor inter-rater agreement. The low inter-rater agreement is due to the fact that different raters have different elevations and dispersions. Correlations are scale independent so differences in dispersion or elevation between raters have no effect.

7 Individual accuracy

There are many ways in which one can score individuals for accuracy (Jussim, 2012, Chapter 12). We used the (Pearson) correlation, mean absolute difference (or delta), elevation error (based on the mean), absolute elevation error, dispersion error (based on the standard deviation) and absolute dispersion error as our measures of accuracy.

To understand each metric, it is useful to consider some examples. Suppose participants were asked to rate the mean trait level of 5 groups and that the true values for these groups were (10, 20, 30, 40, 50).

[9] Table FOLK2. http://www.statistikbanken.dk/
[10] Table HFUDD10.

Table 3: Descriptive statistics for numerical or pseudo-numerical variables. The one- and two-letter acronyms are agreement ratings with each of the main political parties. The appendix has descriptions of the parties.

Variable	Mean	Median	SD	MAD	Min	Max	Skew	Kurtosis
Conservatism	38.03	33.00	25.83	31.13	0.00	100.00	0.25	−1.06
Nationalism	48.93	53.00	27.09	32.62	0.00	100.00	−0.15	−0.91
Econ. lib.	46.29	50.00	26.31	29.65	0.00	100.00	−0.04	−0.93
Pers. lib.	70.84	74.00	22.54	22.24	0.00	100.00	−0.84	0.65
Expel criminals	81.92	92.00	23.57	11.86	0.00	100.00	−1.49	1.80
Expel passives	50.08	50.00	30.03	34.10	0.00	100.00	0.10	−0.99
Time before passive	49.80	46.00	32.29	38.55	0.00	100.00	0.15	−1.16
Muslim problem	59.49	61.00	32.44	41.51	0.00	100.00	−0.34	−1.05
Future perf. immi.	48.32	50.00	27.29	29.65	0.00	100.00	−0.15	−0.82
Denmark for Danes	26.62	17.50	28.94	25.95	0.00	100.00	0.90	−0.28
DF	44.94	45.00	31.64	42.25	0.00	100.00	0.13	−1.26
K	38.35	40.00	23.74	26.69	0.00	100.00	0.22	−0.67
S	51.88	54.00	23.09	25.20	0.00	100.00	−0.24	−0.56
RV	39.36	41.00	23.45	25.20	0.00	100.00	−0.04	−0.70
SF	43.57	47.50	24.00	27.43	0.00	100.00	−0.06	−0.88
V	42.98	42.50	27.27	33.36	0.00	100.00	0.06	−1.06
LA	39.43	40.00	26.77	32.62	0.00	100.00	0.11	−1.09
Ø	42.47	45.00	29.73	37.06	0.00	100.00	0.06	−1.22
Å	43.60	50.00	26.44	27.43	0.00	100.00	0.00	−0.79
RF	36.14	38.50	23.81	25.95	0.00	95.00	0.04	−0.88
PP	35.33	31.00	29.11	40.77	0.00	99.00	0.33	−1.12
TP	36.72	45.00	25.56	28.91	0.00	81.00	−0.24	−1.40
DP	26.55	12.00	29.40	17.79	0.00	100.00	0.71	−0.86
FD	32.95	37.00	23.79	26.69	0.00	99.00	0.07	−0.78
NP	25.48	20.00	25.38	29.65	0.00	93.00	0.59	−0.92
Ghetto experience	4.67	5.00	1.44	1.48	1.00	6.00	−0.92	−0.31
Educational attainment	3.40	3.00	1.86	1.48	1.00	8.00	0.48	−0.77
Age	39.32	38.00	14.68	19.27	18.00	78.00	0.29	−0.94
Cognitive ability	0.00	−0.13	1.00	1.19	−1.54	1.98	0.29	−0.80
Self-rated stereotype precision	54.46	57.00	21.20	25.20	0.00	100.00	−0.11	−0.61

Correlational accuracy (Pearson) is the correlation between the estimates and the true values. So if one person estimated that the values were (35, 15, 40, 30, 60) his correlational accuracy would be .63. *Absolute accuracy* (mean absolute difference) is the mean of the absolute differences between the true values and the estimated values. In the above case, the absolute differences would be (25, 5, 10, 10, 10) and the mean of these is 12. *Elevation error* (based on the mean) is the misestimation of the general level (or height) of the values. This is calculated by subtracting the true values from the estimates and then taking the mean. In the above case, the difference scores would be (25, -5, 10, -10, 10) the mean of which is 6. So, because the error was positive, the estimates were on average a bit too large. *Dispersion error* (based on the standard deviation) is the misestimation of the size of the differences between the groups. This is calculated by subtracting the standard deviation of the true values from the standard deviation of the estimates. In the above case, this would be 16.36 - 15.81 ≈ .54. So, because the value is positive, the estimated group differences were on average a bit too large.

Note that correlational accuracy only concerns the relative differences between the groups. This means that correlational accuracy can be perfect while either or all of the other metrics reveal non-zero errors. For instance, if the estimates were (20, 30, 40, 50, 60), the correlational accuracy would be 1.00, but the elevation error would be 10 – all values are 10 too large. Similarly, if the estimates were (0, 15, 30, 45, 60), the correlational accuracy would be 1.00, but the dispersion error would be 7.91. Figure 5 shows the distribution of correlational accuracy scores.

There was substantial accuracy but with a long left tail. The mean/median correlations were .43/.48. Per Jussim et al. (2015) cutoff levels of .30 and .50, 78 %

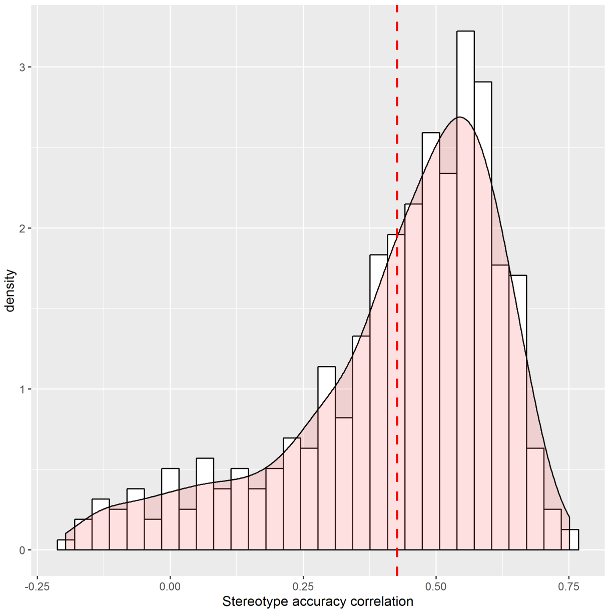

Figure 5: Distribution of correlational accuracy scores. Red line = mean.

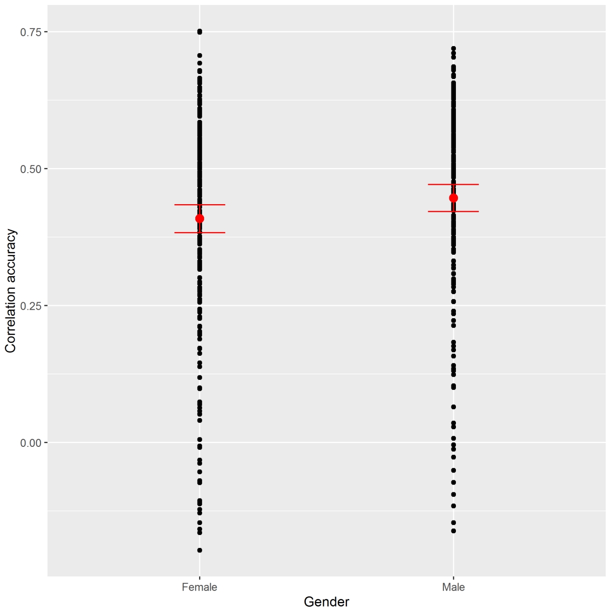

Figure 6: Correlational accuracy by gender.

and 45 % of stereotypes were accurate. Table 4 shows descriptive statistics about each measure of accuracy.

Though individuals did not always get the relative order and distance between groups right, they did on average get the elevation and dispersion about right. In fact, the dispersion bias is slightly negative on average (-1.00 %-points; -6 %) meaning that participants had a slight tendency to underestimate real differences between groups. The elevation error was slightly negative as well (-2.61 %-points; -12 %) meaning that participants tended to underestimate the proportion of persons receiving social benefits.

7.1 *Correlates of individual accuracy*

What predicts stereotype accuracy? Table 5 shows the correlations between predictors and accuracy measures.

In general, predictor correlations were not large. The mean absolute correlational accuracy is .10. Many of the predictors that predicted better correlational accuracy predicted worse absolute accuracy, a pattern also seen in the pilot study. For instance, agreement with the claim that criminal immigrants should be expelled was associated with better correlational accuracy (r = .14) but worse absolute mean delta accuracy (.10).[11] Cognitive ability, however, had the ’normal’ pattern (r = .22 and absolute delta = -.12). In contrast to the pilot study, age was not a particularly strong predictor as it was in the pilot study (pilot study r = .56, replication study r = .08).

Of note is that agreement with conservatism, nationalism and the immigrant critical statements predicted higher dispersion error. Agreement with the red coalition parties predicted lower dispersion error. Note, however, that interpretation is tricky because there was a slight general trend to underestimate dispersion (-1.00 %-points), so a slight positive correlation with dispersion errors is associated with higher accuracy.

It is of interest that the self-rated stereotype precision had about zero relationship to actual accuracy.

The pilot study found a particularly strong effect of gender on correlational accuracy (d = .86). Figure 6 shows the distribution of correlational accuracy scores by gender in the current study.

Males do have higher correlational accuracy, but the difference is small: d = .19 [95CI: .01 to .37]. There is an even smaller gender effect for absolute mean delta (d = -.12 [95CI: -.30 to .06]). The large difference seen in the pilot study did not replicate, but the direction was consistent.

7.1.1 *LASSO regression*

Because OLS tends to overfit models, we employed LASSO regression to locate useful predictors (James et al., 2013). We ran the LASSO regression 500 times with cross-validation. LASSO regression agreed that the only predictors that may be useful to predict correlational stereotype accuracy were cognitive ability (81 % of runs), educational attainment (74 % of runs), disagreement with SF party (43 % of runs) and agreement with the statement about expelling criminal immigrants (23 % of runs).

[11] Recall that higher values are worse for the absolute delta measure because this represents larger errors.

Table 4: Descriptive statistics about accuracy measures.

	mean	median	sd	mad	min	max	skew	kurtosis
pearson r	0.43	0.48	0.20	0.16	−0.20	0.75	−1.11	0.66
mean abs delta	16.89	15.60	6.15	5.01	8.72	45.77	1.77	4.20
dispersion error	−1.00	−2.21	8.79	9.29	−14.33	27.96	0.64	−0.20
dispersion error abs	7.34	6.90	4.92	5.40	0.05	27.96	0.75	0.53
elevation error	−2.61	−4.96	14.26	13.34	−22.05	45.69	0.95	0.59
elevation error abs	11.77	11.19	8.45	9.09	0.01	45.69	0.96	1.40

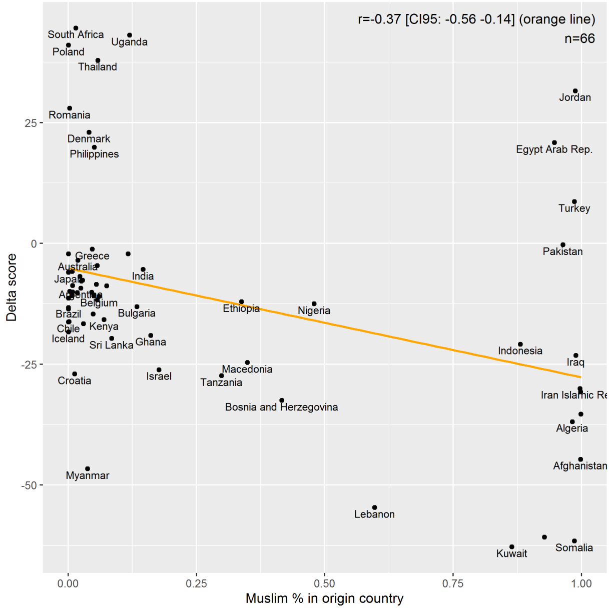

Figure 7: Scatterplot of one person's estimates and the Muslim % in the origin countries.

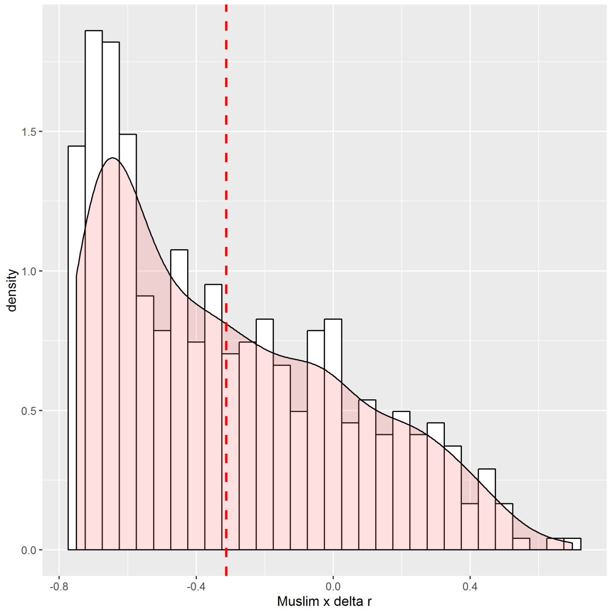

Figure 8: Distribution of Muslim bias correlations. Red line = mean.

7.2 Muslim bias

It is possible to examine whether an individual's estimates are biased for or against Muslims. This can be done by calculating the delta (difference) scores for each estimate that person has made and correlating these with the estimated proportion of each group that is Muslim.[12] Figure 7 shows the delta scores for one person and the proportion of Muslims in each country of origin.

It can be seen that the mean delta score (absolute error) for the non-Muslim countries is around 0, while for the Muslim countries, the mean error is about -25. Thus, this person tends to underestimate the proportion of Muslims receiving social benefits (r = -.37). The pattern seen in the scatterplot is typical of the dataset. Figure 8 shows the distribution of Muslim bias correlations.

The mean/median Muslim bias correlation is -.31/-.38.

We also operationalized Muslim bias in a second way by calculating the weighted mean delta using Muslim % as weights and using non-Muslim as weights and subtracting the first from the second. This is the difference in errors made for the Muslim vs. the non-Muslim countries. This, however, added little because this second measure correlated .96 with the first. Using the absolute variants, the two measures correlated at .95. Due to the very high correlations, these additional measures were not used.

7.2.1 Correlates of Muslim bias

It may be of interest to know what predicts Muslim bias (positive, negative or absolute). One can examine this by correlating the predictors with both Muslim bias and its absolute variant. The difference is that the first is directional and the second is not. Table 6 shows the results.

[12] The proportion of the population that is Muslim in the home countries. These values were copied from Pew Research's study (Pew Research Center, 2011). We used the 2010 values.

Table 5: Correlates of individual stereotype accuracy. Numbers in brackets are analytic 95 % confidence intervals. The one- and two-letter acronyms are agreement ratings with each of the main political parties. The appendix has descriptions of the parties.

Predictor	pearson r	mean abs delta	dispersion error	dispersion error abs	elevation error	elevation error abs
Conservatism	0.13 [0.04 0.21]	0.02 [-0.07 0.11]	0.15 [0.06 0.24]	-0.03 [-0.12 0.06]	0.09 [0.00 0.18]	-0.03 [-0.12 0.06]
Nationalism	0.11 [0.02 0.20]	0.16 [0.07 0.24]	0.32 [0.23 0.39]	0.07 [-0.02 0.15]	0.20 [0.11 0.28]	0.01 [-0.08 0.10]
Econ lib	0.07 [-0.02 0.16]	0.00 [-0.09 0.09]	0.09 [0.00 0.17]	-0.03 [-0.12 0.06]	0.05 [-0.04 0.14]	-0.03 [-0.12 0.06]
Pers lib	0.02 [-0.07 0.11]	0.02 [-0.07 0.11]	0.00 [-0.09 0.09]	0.08 [-0.01 0.17]	-0.06 [-0.15 0.03]	0.01 [-0.07 0.10]
Expel criminals	0.14 [0.06 0.23]	0.10 [0.01 0.19]	0.28 [0.20 0.36]	-0.02 [-0.11 0.07]	0.15 [0.06 0.24]	-0.01 [-0.10 0.08]
Expel passives	0.06 [-0.03 0.15]	0.18 [0.10 0.27]	0.29 [0.20 0.37]	0.04 [-0.05 0.12]	0.16 [0.07 0.24]	0.05 [-0.04 0.14]
Time before passive	0.06 [-0.03 0.15]	0.16 [0.07 0.25]	0.26 [0.18 0.34]	0.05 [-0.04 0.14]	0.17 [0.08 0.25]	0.05 [-0.04 0.14]
Muslim problem	0.08 [-0.01 0.17]	0.15 [0.07 0.24]	0.33 [0.24 0.40]	0.01 [-0.08 0.10]	0.21 [0.13 0.30]	0.00 [-0.09 0.09]
Future perf. immi.	-0.12 [-0.21 -0.03]	-0.17 [-0.26 -0.08]	-0.33 [-0.41 -0.25]	-0.12 [-0.21 -0.03]	-0.21 [-0.29 -0.12]	-0.02 [-0.11 0.06]
Denmark for Danes	-0.02 [-0.11 0.07]	0.22 [0.13 0.30]	0.21 [0.12 0.29]	0.04 [-0.05 0.13]	0.18 [0.10 0.27]	0.12 [0.03 0.20]
DF	0.04 [-0.05 0.13]	0.16 [0.07 0.24]	0.25 [0.16 0.33]	0.01 [-0.08 0.10]	0.17 [0.09 0.26]	0.03 [-0.06 0.12]
K	0.09 [0.00 0.18]	0.01 [-0.08 0.10]	0.09 [0.00 0.18]	-0.08 [-0.17 0.00]	0.10 [0.01 0.19]	-0.02 [-0.11 0.07]
S	-0.09 [-0.18 0.00]	0.01 [-0.07 0.10]	-0.01 [-0.10 0.08]	-0.04 [-0.13 0.05]	0.01 [-0.08 0.10]	-0.02 [-0.11 0.07]
RV	-0.01 [-0.09 0.08]	-0.16 [-0.25 -0.07]	-0.16 [-0.25 -0.07]	-0.16 [-0.25 -0.07]	-0.05 [-0.14 0.04]	-0.08 [-0.17 0.01]
SF	-0.17 [-0.26 -0.08]	0.00 [-0.09 0.09]	-0.13 [-0.21 -0.04]	-0.04 [-0.13 0.05]	-0.01 [-0.10 0.08]	0.00 [-0.09 0.09]
V	0.09 [0.00 0.17]	-0.02 [-0.11 0.07]	0.09 [0.00 0.18]	-0.07 [-0.16 0.02]	0.08 [-0.01 0.17]	-0.04 [-0.13 0.05]
LA	0.05 [-0.04 0.14]	-0.04 [-0.13 0.05]	0.02 [-0.07 0.10]	-0.09 [-0.18 0.00]	0.05 [-0.04 0.14]	-0.05 [-0.13 0.04]
Ø	-0.16 [-0.24 -0.07]	0.00 [-0.09 0.09]	-0.19 [-0.28 -0.11]	0.00 [-0.09 0.09]	-0.07 [-0.16 0.02]	0.06 [-0.03 0.15]
Å	-0.08 [-0.17 0.01]	-0.08 [-0.17 0.00]	-0.21 [-0.30 -0.13]	-0.10 [-0.19 -0.01]	-0.06 [-0.14 0.03]	-0.01 [-0.10 0.08]
RF	0.06 [-0.10 0.21]	-0.02 [-0.18 0.14]	0.02 [-0.14 0.18]	-0.01 [-0.17 0.15]	0.11 [-0.04 0.27]	0.02 [-0.14 0.18]
PP	-0.09 [-0.30 0.12]	-0.14 [-0.34 0.08]	-0.25 [-0.44 -0.04]	-0.06 [-0.27 0.16]	0.02 [-0.19 0.24]	-0.05 [-0.26 0.16]
TP	-0.12 [-0.36 0.14]	-0.07 [-0.32 0.18]	-0.37 [-0.57 -0.13]	0.03 [-0.23 0.28]	-0.17 [-0.41 0.08]	0.11 [-0.15 0.36]
DP	-0.11 [-0.25 0.04]	0.23 [0.09 0.36]	0.22 [0.08 0.35]	0.11 [-0.03 0.25]	0.29 [0.15 0.42]	0.10 [-0.04 0.24]
FD	-0.21 [-0.42 0.02]	-0.04 [-0.26 0.19]	-0.19 [-0.39 0.04]	-0.17 [-0.38 0.06]	0.10 [-0.13 0.32]	0.02 [-0.21 0.24]
NP	-0.17 [-0.30 -0.04]	-0.01 [-0.14 0.12]	-0.02 [-0.16 0.11]	-0.05 [-0.18 0.09]	0.12 [-0.02 0.25]	-0.05 [-0.19 0.08]
Ghetto exp	-0.03 [-0.12 0.06]	0.03 [-0.06 0.12]	0.00 [-0.09 0.09]	-0.01 [-0.10 0.08]	-0.01 [-0.10 0.08]	0.06 [-0.03 0.15]
Education	0.20 [0.11 0.28]	-0.05 [-0.14 0.04]	0.03 [-0.06 0.12]	0.06 [-0.03 0.15]	-0.01 [-0.10 0.08]	-0.04 [-0.13 0.05]
Age	0.08 [-0.01 0.17]	0.11 [0.02 0.20]	0.16 [0.08 0.25]	0.12 [0.03 0.20]	0.04 [-0.05 0.13]	0.04 [-0.05 0.13]
Cognitive ability	0.22 [0.13 0.30]	-0.12 [-0.21 -0.03]	-0.06 [-0.15 0.03]	0.04 [-0.05 0.12]	-0.12 [-0.20 -0.03]	-0.07 [-0.16 0.02]
Self-rated stereotype accuracy	-0.02 [-0.14 0.09]	0.10 [-0.01 0.22]	0.13 [0.01 0.24]	0.03 [-0.09 0.15]	0.11 [0.00 0.23]	0.06 [-0.06 0.17]

Table 6: Predictors of Muslim bias. Numbers in brackets are 95 % confidence intervals. The one- and two-letter acronyms are agreement ratings with each of the main political parties. The appendix has descriptions of the parties.

Predictor	Muslim bias r	Muslim bias abs r
Conservatism	0.18 [0.09 0.27]	-0.08 [-0.17 0.01]
Nationalism	0.30 [0.21 0.38]	-0.12 [-0.21 -0.03]
Econ lib	0.09 [0.00 0.18]	-0.04 [-0.13 0.05]
Pers lib	0.00 [-0.09 0.09]	0.11 [0.02 0.20]
Expel criminals	0.27 [0.19 0.35]	-0.09 [-0.18 0.00]
Expel passives	0.26 [0.18 0.35]	-0.08 [-0.17 0.01]
Time before passive	0.24 [0.16 0.33]	-0.11 [-0.19 -0.02]
Muslim problem	0.32 [0.24 0.40]	-0.14 [-0.23 -0.05]
Future perf. immi.	-0.31 [-0.39 -0.23]	0.13 [0.04 0.21]
Denmark for Danes	0.17 [0.09 0.26]	-0.11 [-0.20 -0.02]
DF	0.23 [0.14 0.31]	-0.09 [-0.18 0.00]
K	0.11 [0.02 0.20]	-0.09 [-0.17 0.00]
S	-0.03 [-0.12 0.06]	-0.06 [-0.15 0.03]
RV	-0.15 [-0.23 -0.06]	-0.03 [-0.12 0.06]
SF	-0.18 [-0.26 -0.09]	-0.05 [-0.14 0.04]
V	0.11 [0.02 0.19]	-0.04 [-0.13 0.05]
LA	0.04 [-0.05 0.13]	-0.05 [-0.14 0.04]
Ø	-0.23 [-0.31 -0.14]	0.03 [-0.06 0.12]
Å	-0.22 [-0.30 -0.13]	-0.02 [-0.11 0.07]
RF	0.04 [-0.12 0.19]	-0.12 [-0.27 0.04]
PP	-0.26 [-0.45 -0.04]	0.00 [-0.21 0.22]
TP	-0.33 [-0.54 -0.08]	0.27 [0.01 0.49]
DP	0.19 [0.04 0.32]	-0.08 [-0.22 0.06]
FD	-0.19 [-0.40 0.04]	-0.20 [-0.41 0.03]
NP	-0.05 [-0.19 0.08]	-0.05 [-0.18 0.09]
Ghetto exp.	0.00 [-0.09 0.09]	0.05 [-0.04 0.14]
Education	0.10 [0.01 0.19]	0.02 [-0.07 0.11]
Age	0.17 [0.08 0.25]	0.00 [-0.08 0.09]
Cognitive ability	0.02 [-0.07 0.11]	0.12 [0.03 0.21]

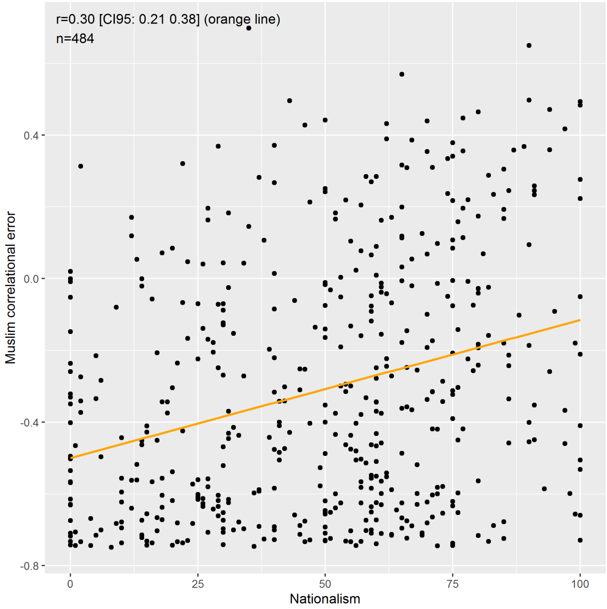

Figure 9: Self-rated nationalism and Muslim correlational bias.

In general, the patterns are what we have seen before, namely that nationalism, conservatism, agreement with immigrant critical statements and parties predicts higher values of Muslim correlational bias. Because most people have a negative Muslim correlational bias (tend to underestimate the proportions of Muslim groups that receive benefits relative to the non-Muslim groups), predictors that have positive associations with Muslim correlational bias correlations at first result in higher accuracy (as the bias goes towards 0), but also lower accuracy at a later point. This is best understood with the help of a scatterplot. Figure 9 shows the relationship between self-rated nationalism and Muslim correlational bias.

It can be seen that even the most extreme nationalists in this sample are still not biased against Muslim groups in their ratings because the regression line does not cross 0.[13]

Alternatively, one can examine the absolute Muslim correlational bias, which is an absolute measure of differential error for Muslim vs. non-Muslim countries. These scores are not well predicted by anything. There are a few predictors that have confidence intervals that do not cross 0. Cognitive ability and personal liberalism (r's .12 and .11) correlate positively such that smarter people and people who think personal freedom is important tend to be slightly more biased with regards to the groups with higher proportions of Muslims. On the other hand, nationalism and four of

[13] Recall that a positive Muslim correlational error means that one tends to overestimate the proportions of persons from the groups with more Muslims receiving social benefits relative to the groups with fewer Muslims.

Table 7: Example of individual and aggregate estimates.

	Person A	Person B	Person C	Mean estimates	True values
Estimates	15	30	40	28.3	10
	5	10	20	11.7	20
	40	20	30	30.0	30
	20	30	40	30.0	40
	25	30	40	31.7	50
Correlational accuracy	0.43	0.35	0.35	0.48	

the six immigrant-statement predictors had validities with confidence intervals that did not cross 0. However, they are all small by conventional standards (r's -.11 to -.14).

8 Aggregate accuracy

So far we have examined individual-level (in)accuracy and its correlates (also *called personal stereotypes*). However, one can also aggregate the estimates and then examine (in)accuracy and its correlates *(consensual stereotypes)* (Jussim, 2012).

To give an example. Suppose, as before, that the real values for 5 groups are (10, 20, 30, 40, 50). Now suppose we get 3 people to estimate these values and they give the estimates (5, 10, 10, 20, 23), (20, 10, 20, 30, 30), (40, 20, 30, 40, 50). Table 7 shows the individual estimates, the average (mean) estimates, the true values and the correlational accuracy.

Thus, we see that each person is fairly accurate (r's .35 to .43, mean of .38), but the set of aggregate estimates was more accurate than any individual set of estimates (r = .48). This is a common finding and is expected from basic statistics (cf. Spearman-Brown formula).

8.1 Aggregation methods

How should estimates be aggregated? In the pilot study we compared the mean, the 10 % trimmed mean and the median. We followed the same approach here. Table 8 shows the accuracy scores for each aggregated stereotype.

As can be seen, the simple mean was the best method, but the differences between the methods were very slight. This result was also found in the pilot study and in fact, the mean accuracy was almost exactly the same as well (.70 in the pilot study). Accuracy is fairly high for the aggregated estimates, the average distance to the real value is only about 8 %-points (mean abs. delta).

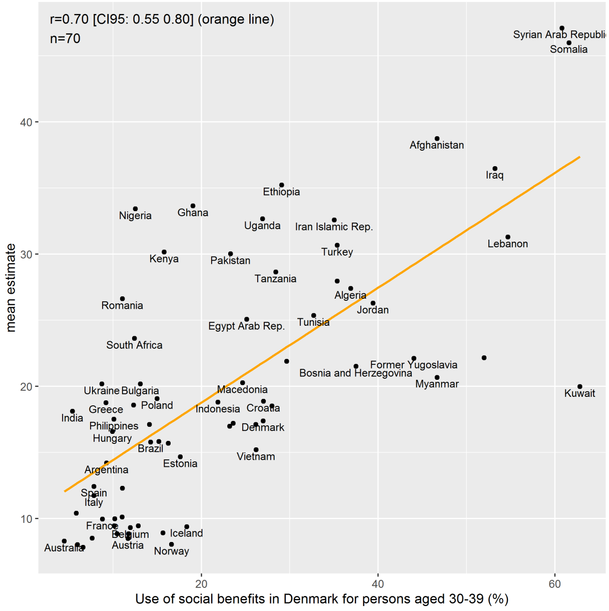

Figure 10: Mean stereotype and real values.

Overall, stereotypes had a strong tendency to underestimate real group differences (dispersion error = -5.7; -38 %) and underestimate the proportion of persons receiving social benefits (elevation error = -2.6; -12 %). Figure 10 shows the average stereotypes and the true values.

We see that Sub-Saharan African countries (Nigeria, Ghana, Kenya, Ethiopia and Uganda) are positive outliers (fewer persons from these countries receive social benefits than what stereotypes say) and Kuwait is a negative outlier (more persons receive social benefits than stereotypes say).

8.2 Aggregate accuracy by number of raters

It is known that when averaging data sources that have different biases and some signal, the error averages out and the signal becomes stronger (Remmers et al., 1927). This is seen in the present dataset in that the aggregate estimates are more accurate than the individual estimates. Still, one might wonder how many raters' estimates one needs to average in order to get a substantially superior estimate compare to the typical individual estimate. To investigate this, we sub-sampled the dataset repeatedly (1000 times) for 1, 2, ..., 100 raters and averaged their estimates (100k samples). Finally, these estimates were correlated with the true values. Figure 11 shows the results.

We see that the estimate quickly reaches the limit; the aggregate accuracy is about .65 with 10 raters and about .69 with 43 raters.

8.3 Predictors of accuracy

As with the individual-level analyses, one may want to try to predict which country of origin groups are more accurately stereotyped than others.

Table 8: Accuracy scores for aggregated estimates by method.

Method	pearson r	mean abs delta	dispersion	dispersion error	dispersion error abs	elevation	elevation error	elevation error abs
mean	0.696	8.160	9.476	-5.717	5.717	19.864	-2.615	2.615
median	0.668	10.199	9.208	-5.985	5.985	14.007	-8.471	8.471
trimmed mean	0.694	8.868	9.611	-5.582	5.582	17.050	-5.428	5.428

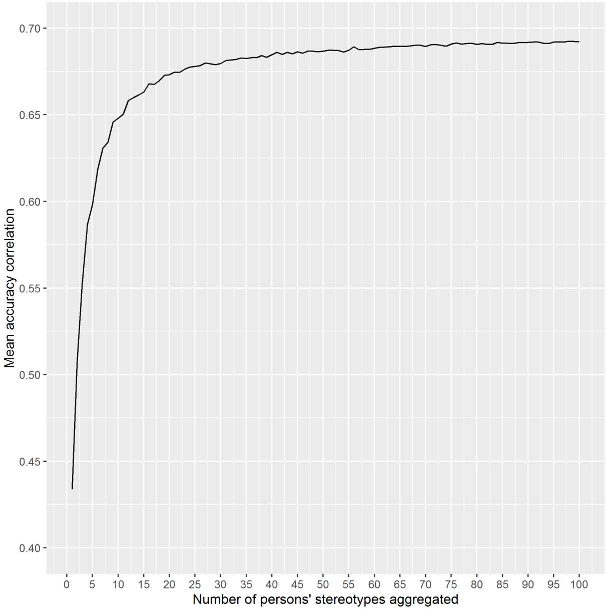

Figure 11: Aggregate accuracy as a function of the number of raters. Based on 1000 random samples of raters for each number of raters.

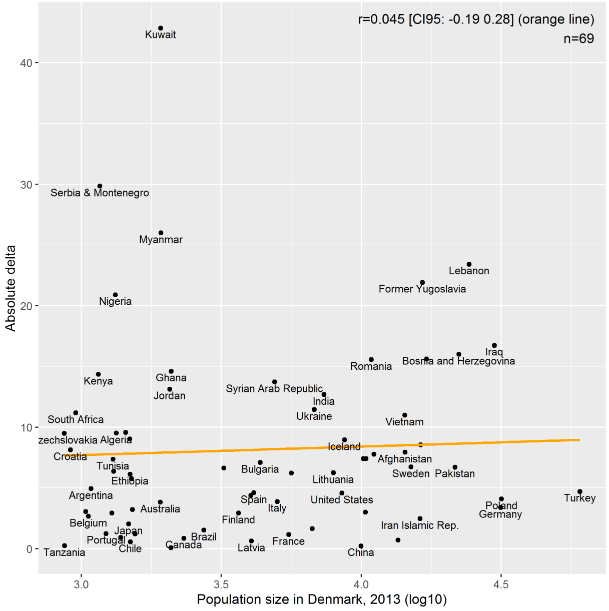

Figure 12: Absolute delta and population size.

8.3.1 *Population size*

If stereotypes are based in part on personal experience or media discussion of groups, and one is more likely to meet or read about persons from larger groups, one would expect that stereotypes are more accurate for the larger groups. One can investigate this by correlating the abs. delta values for each group with the population sizes in Denmark. We exclude Denmark itself from this analysis because it is qualitatively different (it is not an immigrant group) and because it has a population size far exceeding any of the others. Figure 12 shows the results.

We see that there is about no relationship (r = .05) between the absolute delta and population size; stereotypes are not more or less precise for the larger groups. Thus, we find contrary results to those predicted by a personal experience/media-exposure model.

8.3.2 *GDP per capita as a proxy*

Do raters rely upon GDP per capita[14] as a proxy for estimating persons on social benefits in Denmark? It seems that way because the positive outliers are very poor countries while the strongest negative outlier, Kuwait, is a very rich country (refer back to Figure 10). The reasoning here is that there is a negative correlation between how rich a country is and the proportion of persons receiving social benefits in Denmark originating from that country (r = -.39). If people rely upon GDP per capita to estimate immigrant performance, then their estimates will be highly correlated with these, which they are (r = -.79). Finally, this implies that the estimation errors (residuals) will be correlated between social benefits ~ GDP per capita and social benefits ~ estimates. Figure 13 shows a path model of this scenario.

The model here is that the GDP per capita in the home country and the use of social benefits in Denmark has a common cause (see the literature on the spatial transferability hypothesis (e.g. Jones & Schneider 2010; Kirkegaard 2014, 2015; Kirkegaard & Fuerst 2014)). People's stereotypes are then caused by (their estimates of) the origin countries' GDP per capita. If this is so, then one would expect the residuals for the two dotted lines to be correlated. To investigate this, we correlated the estimate errors (abs. delta) with the residuals of regressing the criterion values on GDP

[14] We used the International Monetary Fond's estimates of GDP per capita (PPP) (International Monetary Fund, 2015).

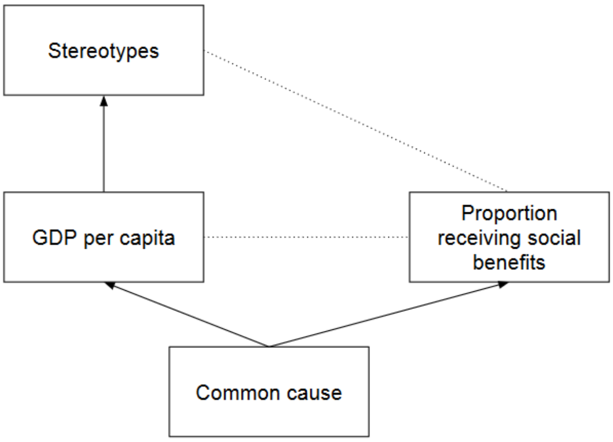

Figure 13: Path model of the relationship between GDP, stereotypes and use of social benefits.

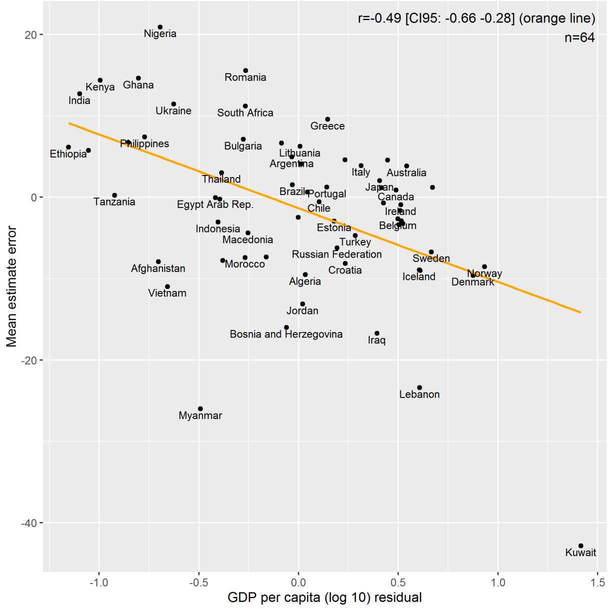

Figure 14: GDP per capita ($\log_{10}$) residual and the mean estimate error.

per capita ($\log_{10}$).[15] Figure 14 shows the results.

This model was supported for the present dataset, although not as strongly as it was in the pilot study (r = -.61).

8.3.3 *Muslim bias*

Just as we explored the role of Muslims in stereotype (in)accuracy at the individual level, so one can do at the aggregate level. Figure 15 shows the correlations between the estimation errors and the proportion of Muslims in the origin countries.

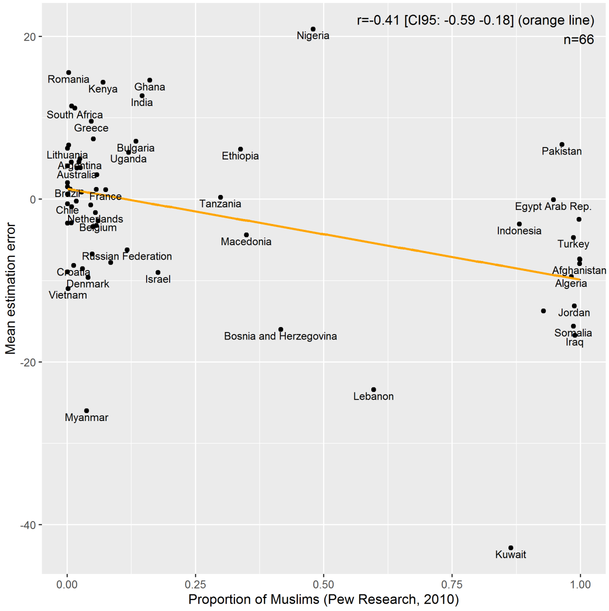

Figure 15: Mean estimation errors and the proportion of Muslims in the origin countries.

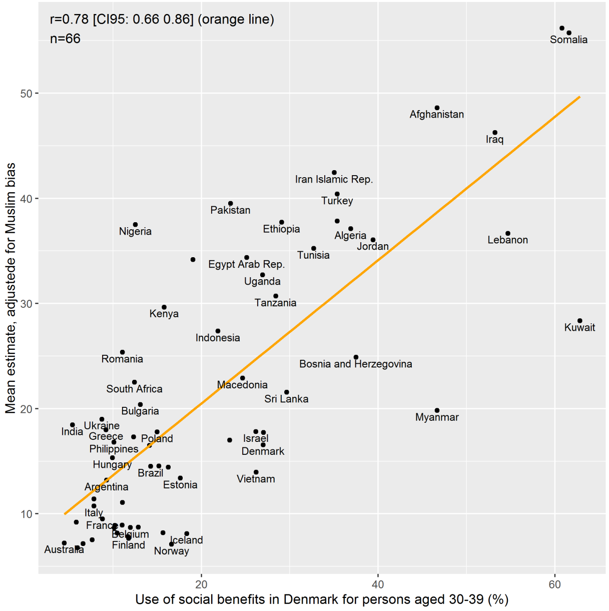

Figure 16: Aggregate stereotypes adjusted for pro-Muslim bias and criterion values.

Just as we saw at the individual level, participants tended to underestimate the proportion of Muslim groups that received social benefits. Because the non-Muslim countries' errors clustered around 0, this resulted in the absolute estimation errors being larger for the Muslim countries (r = .31).

It is possible to adjust for this bias in the estimates. This was done by using a linear model with delta estimate as the outcome and the Muslim proportion as the predictor. The predicted values for each group were then subtracted from the mean estimated to produce adjusted estimates, which are shown in Figure 16.

[15] GDP per capita values are very skewed. We used the base 10 logarithm to yield more normal and still interpretable values.

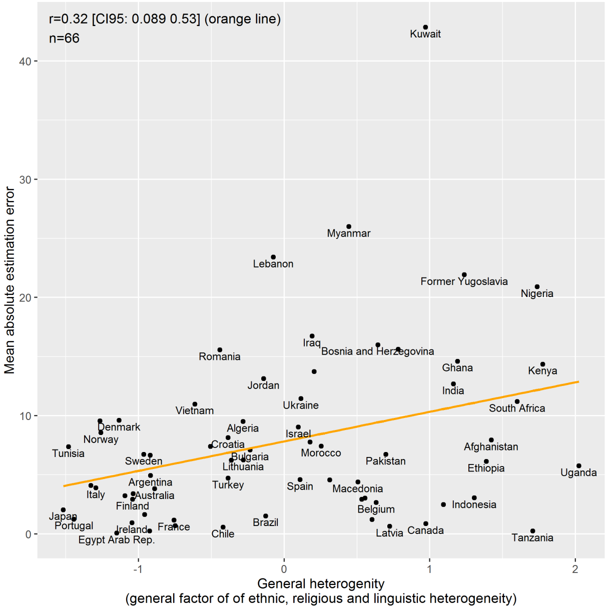

Figure 17: General heterogeneity and mean absolute estimation error.

As expected, this adjustment increased the accuracy (any bias causes a decrease in accuracy), but only slightly: from .70 to .78.

8.3.4 *Ethnic heterogeneity*

A hypothesis suggested by Peter Frost is that if stereotypes rely on visual cues such as racial appearance or cultural practices, then countries of origin that are more ethnically/racially/culturally heterogeneous should be harder to stereotype. To investigate this, we correlated absolute estimation errors with a general measure of heterogeneity.[16] The results are shown in Figure 17.

As suggested, errors were larger (in any direction) for the more heterogeneous countries of origin.

9 Discussion and conclusion

We observed relatively high levels of accuracy. The accuracy for aggregate stereotypes was much higher (r = .70) than the median individual accuracy (r = .48) as expected based on the Spearman-Brown formula. In thinking about stereotypes, the aggregate stereotypes are usually the important ones to focus on. This is because these represent the typical or average expectations of the population. The beliefs and any resultant actions of single persons average out with each other.

In general, the present results are similar to those found in the pilot study. The only findings that did not replicate were the strong predictive validities of age and gender observed in the pilot study.

The findings fit well with the general literature on stereotype accuracy (Jussim, 2012; Jussim et al., 2015). The average correlation in social psychology has been estimated to be around .20 (Richard et al., 2003),[17] while we found that 78 % of participants had accuracy correlations above .30 and 45 % had scores above .50. Previous studies of racial/ethnic stereotypes reported average accuracies between .36 to .69 and .53 to .93 for individual and aggregate-level stereotypes, respectively (Jussim, 2012, p. 327).

There are many reasons why the present study is a strong contribution to the literature:

1. It uses a large (N≈500), nationally representative sample. The studies of racial/ethnic stereotypes listed by Jussim (Jussim, 2012, p. 327) have sample sizes from 62 to 94, all of which were non-representative convenience samples.
2. The study was pre-registered and most of the analyses were designated in advance making it difficult to use QRPs to force particular results (Simmons et al., 2011).
3. The analysis materials, analysis code, and data are public so that other researchers can examine the code and re-use the data for any purpose without having to get approval.
4. The number of stereotyped groups is large (70) compared to most previous studies.
5. The study includes a large number of covariates that can be related to (in)accuracy measures and to each other.

There are a number of limitations with the study:

1. Many participants did not comply with instructions and filled out the questions seemingly at random. We attempted to filter these out by inclusion of more control variables, but this probably made the sample less representative and we missed some of the random responses.
2. The group of participants who gave estimates that were opposite of reality. Of those we surveyed in the follow-up questionnaire, 12/15 stated that they had misunderstood the question,

[16] This was a general factor extracted from 5 measures of heterogeneity. See the pilot study for details (Kirkegaard & Bjerrekær, 2016a).

[17] This value is very likely to be too large. The estimate is based on a large number of meta-analyses which mostly did not correct for the endemic publication bias in this field (Open Science Collaboration, 2015).

in that they were estimating the proportion of persons receiving benefits in the home country, not in Denmark, however they may have been lying. Future studies of stereotype accuracy should take extra care in filtering out non-compliant participants, ensuring that participants understand the task and figuring out if some people give reverse answers on purpose and if so, why.

3. We openly stated the purpose of the study in the introduction of the survey. This was done in hope of increasing participant compliance and understanding. In hindsight, this may have been a mistake and may have biased results in an unknown direction. Future studies should experiment with both stating and not stating the purpose of the study to see how this affects participants' estimates.

Supplementary material and acknowledgments

The data, R analysis code and high quality figures can be found at https://osf.io/wxqma/.

Thanks to Jesper Lindhardt of Survee (http://www.survee.dk/) who provided valuable help with the design of the survey as well as with discovering the cause of the reversed answers.

This research was supported by two anonymous research contributions. Neither contributor had anything to do with the data collection, analyses or reporting.

Thanks to Bob Williams, Gerhard Meisenberg and Heiner Rindermann for reviewing: https://openpsych.net/forum/showthread.php?tid=287

References

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi: 10.1037/1040-3590.6.4.284

Cohen, J., & Cohen, J. (Eds.). (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed ed.). Mahwah, N.J: L. Erlbaum Associates.

International Monetary Fund. (2015). *World Economic Outlook Database.* Retrieved from https://www.imf.org/external/pubs/ft/weo/2015/01/weodata/index.aspx

James, G., Witten, D., Hastie, T., & Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: with applications in R* (No. 103). New York: Springer.

Jones, G., & Schneider, W. J. (2010, July). IQ in the Production Function: Evidence from Immigrant Earnings. *Economic Inquiry*, *48*(3), 743–755. doi: 10.1111/j.1465-7295.2008.00206.x

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

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science*. doi: 10.1177/0963721415605257

Kirkegaard, E. O. W. (2014, October). Crime, income, educational attainment and employment among immigrant groups in Norway and Finland. *Open Differential Psychology*. Retrieved 2014-10-13, from https://openpsych.net/paper/29

Kirkegaard, E. O. W. (2015). Crime among Dutch immigrant groups is predictable from country-level variables. *Open Differential Psychology*. Retrieved from https://openpsych.net/paper/16

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016a, apr). Country of origin and use of social benefits: A pilot study of stereotype accuracy in Denmark. *Open Differential Psychology*. Retrieved 2016-07-04, from https://openpsych.net/paper/11

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016b, July). ICAR5: a 5-item public domain cognitive test. *Open Differential Psychology*. Retrieved from https://openpsych.net/paper/5

Kirkegaard, E. O. W., & Fuerst, J. (2014, May). Educational attainment, income, use of social benefits, crime rate and the general socioeconomic factor among 71 immigrant groups in Denmark. *Open Differential Psychology*. Retrieved 2014-10-13, from https://openpsych.net/paper/21

Open Science Collaboration. (2015, August). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. Retrieved 2016-09-27, from http://science.sciencemag.org/content/349/6251/aac4716 doi: 10.1126/science.aac4716

Pew Research Center. (2011, January). *Table: Muslim Population by Country.* Pew Research Center. Retrieved 2015-08-07, from http://www.pewforum.org/2011/01/27/table-muslim-population-by-country/

Remmers, H. H., Shock, N. W., & Kelly, E. L. (1927, March). An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue rating scale. *Journal of Educational Psychology*, *18*(3), 187–195.

Revelle, W. (2015, April). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Retrieved 2015-04-29, from http://cran.r-project.org/web/packages/psych/index.html

Revelle, W. (2016). *An introduction to psychometric theory with applications in R.* Retrieved from http://www.personality-project.org/r/book/

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. doi: 10.1037/1089-2680.7.4.331

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011, November). False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Uebersax, J. S. (2015, September). *Introduction to the Tetrachoric and Polychoric Correlation Coefficients.* Retrieved from http://john-uebersax.com/stat/tetra.htm (latent correlation)

Appendix

Political parties in Denmark

Denmark has a multi-party system based on proportional representation (D'Hondt method). Currently, there are 9 parties in parliament. To avoid possibly disputable short description of the parties, we instead give an empirical description of each party based on what their supporters answered in our survey. Specifically, for each party we select the party supporters by finding the persons that agreed at least 80 % with the party and then calculate their mean agreement with the political ideologies and the political statements. The list below has the included parties, a literal English translation of their names and sometimes an alternative English translation of the name when the literal translation is not very sensible.

- DF – Dansk Folkeparti / Danish People's Party
- K – Det Konservative Folkeparti / The Conservative People's Party
- S – Socialdemokraterne / The Social Democrats
- RV – Det Radikale Venstre / The Radical Left [Danish Social Liberal Party]
- SF – Socialistisk Folkeparti / Socialist's People's Party
- V – Venstre / Left[18] [Denmark's Liberal Party]
- LA – Liberal Alliance / Liberal Alliance
- Ø – Enhedslisten / Unitlist [Red-Green Alliance]
- Å – Alternativet / The Alternative
- RF – Retsforbundet / The Justice Band [Justice Party of Denmark]
- PP – Piratpartiet / The Pirate Party
- TP – Teknologipartiet / The Technology Party
- DP – Danskernes Parti / The Danes' Party
- FD – Forenede Demokrater / United Democrats
- NP – Nationalpartiet / The National Party

Danish politics generally relies upon two coalitions, called the blue and the red coalitions. The blue coalition consists of V, K, DF, LA while the red consists of S, RV, SF, Ø and Å.

Table 9 shows the results.

Some of the parties outside parliament had very small numbers of supporters in our survey, so the results for them should be treated with skepticism. The parties DF thru Å are represented in parliament. Among them, some patterns can be seen. K is unsurprisingly the most conservative party followed by LA and V. Surprisingly, DF is not among the conservative top three. Our guess is that while the party's politicians themselves are fairly conservative, the voters are not necessarily so. DF is the main anti-immigration party in Denmark (the only in parliament) and can thus be expected to attract many voters for that reason alone, voters that may not agree with the party's other policies. The blue coalition parties are more conservative than the red coalition parties (means: 63 vs. 23). DF and K are the most nationalist parties with RV, Ø and Å being the least. Perhaps surprisingly, S and V, usually the most powerful parties in each coalition, are similar in nationalism (49 vs. 60). The blue coalition is more nationalist than the red coalition (63 vs. 35). On the economic liberalism axis, K and LA are the most liberal whereas Ø (the communist party) is the least liberal.

There are no marked differences with respect to personal liberalism. This is possibly due to the way the question was framed. If one had asked participants

[18] This party is not 'left-wing', the name comes from its old sitting position in parliament where the main opponent was the conservative party (then called *Højre/Right*). https://en.wikipedia.org/wiki/Venstre_%28Denmark%29#History

about their support for different personal freedom related policies (e.g. gay marriage, cannabis/drug legalization, incest sex, e-cigarettes), one would probably have seen larger differences.

Table 9: Empirical description of the political parties in Denmark.

Party	Conservatism	Nationalism	Econ lib	Pers lib	Expel criminals	Expel passives	Time before passive	Muslim problem	Future perf immi	Denmark for Danes	N supporters
DF	50	70	57	73	97	76	79	90	32	56	92
K	78	64	71	74	93	72	72	80	26	36	20
S	27	49	34	70	82	45	43	60	52	18	57
RV	30	31	41	64	63	23	26	25	73	2	20
SF	20	40	26	71	71	30	27	41	62	16	26
V	62	60	69	75	92	64	67	74	40	39	50
LA	62	56	78	79	89	59	61	63	46	23	39
Ø	20	33	25	75	66	27	27	34	63	13	61
Å	20	25	33	72	63	26	31	35	62	10	48
RF	57	53	74	85	71	54	52	63	37	29	7
PP	19	11	39	85	34	33	33	19	74	1	8
TP	15	72	76	51	100	100	100	80	14	81	1
DP	67	88	65	85	100	95	83	100	18	80	9
FD	91	70	78	81	83	89	34	71	32	81	1
NP	26	40	54	65	68	30	40	43	50	38	4