Submitted: 13th of December 2015
Published: 18th of April 2016

Country of origin and use of social benefits: A pilot study of stereotype accuracy in Denmark

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

Open Differential Psychology

Abstract

We asked a small, broad online sample of Danes (N=60; N=48 after quality control) to estimate the use of social benefits for persons grouped by country of origin. The median personal stereotype accuracy correlation was .55 [CI95: .46 to .58]. The aggregate stereotype accuracy was .70 [$N_{\text{countries}}$=71, CI95: . 56 to .80]. The study was underpowered to detect relationships between the accuracy of beliefs and many predictors, but some plausible predictors were found including being male d = .86 [CI95: .17 to 1.56], being older r=.56 [CI95: .33 to .73], nationalism r=.34 [CI95: .07 to .57], personal liberalism, r=.32 [CI95: .04 to .55] and cognitive ability (r=.23 [CI95: -.06 to .48]). The study was preregistered.

Keywords: stereotypes, stereotype accuracy, Denmark, immigrants, social benefits, group differences, Muslims

1 Introduction

Much research in social psychology concerns group dynamics and stereotypes[1] (426000 results on Google Scholar as of writing for “stereotype”). It is therefore surprising that relatively few studies have examined the accuracy or inaccuracy of stereotypes (1220 results for “stereotype accuracy”). However, interest in stereotype accuracy seems to be increasing and there is growing agreement that stereotypes are generally fairly accurate (Jussim, 2012; Jussim et al., 2015).

To our knowledge, not much (or any) research has been done on stereotype accuracy in Denmark. Because we had access to a large dataset of register-based information about immigrant groups in Denmark from a previous study (Kirkegaard & Fuerst, 2014), it seemed in order to take a first stab at examining the question in Denmark.

2 Methods

The data and R source code are publicly available in the supplementary materials.

2.1 Participants

We wrote a proposal for a study in late 2014 and preregistered it in December (the proposal is found in the supplementary materials). We then designed a questionnaire using Google Forms (a free survey tool). We wanted to avoid a narrow student sample (see discussion in Henrich et al. (2010)) and thus decided to post the survey to multiple places on the Internet. We posted the survey on the following sites:

- A paid ad on Reddit.
- The Facebook group for members of the Danish Mensa.
- The Facebook group for the physics department at the University of Aarhus.
- The Facebook group for the physical education department at University of Aalborg.
- Emil’s Facebook wall.
- Julius’ Facebook wall.
- 180grader.dk, a news-aggregator mainly used by free-market proponents, nationalists and conservatives.
- On an intranet for a gymnasie (Danish secondary school) with the help of a friend who was a student at the time.

Furthermore, with the help of a teacher, we gave the survey to a class of students at a gymnasie in Viborg, central Jutland. In doing the above we expected to get a sample size of 50-200. We got a total sample of 60.

* University of Aarhus. E-mail: emil@emilkirkegaard.dk
† University of Aalborg. E-mail: juliusdb.science@gmail.com

[1] By *stereotype* we mean a belief about a group. This is a neutral definition allowing for both accurate and inaccurate stereotypes. This is in contrast to some other definitions of the word which assume that they must be inaccurate beliefs. However, as Lee Jussim has argued at length, such a definition is untenable (Jussim 2012, p. 269ff; Jussim et al. 2009, p. 200ff).

2.2 The survey

The Danish-language questionnaire can be found in the supplementary materials. An English translation of it can be also found in the supplementary materials.

Because of the sensitive nature of the matter, we expected a non-trivial number of unserious or trolling responses. For this reason, we added two initial control questions to make sure that participants understood the concept and were not just clicking thru. The questions concern the average height of men vs. women, and that of Europeans vs. East Asians. These were chosen for their lack of controversy and obviousness to the casual observer.

The structure of the survey was as follows:

- Page 1: Introduction and control questions.
  - Brief introduction.
  - Control question 1 (men vs. women; height).
  - Control question 2 (Europeans vs. Asians; height).
- Page 2: Stereotypes
  - Brief instructions.
  - 71 numeric fields for 71 countries of origin. The list of countries can be found in the appendix. Only numbers between 0 and 100 were acceptable answers.
- Page 3: Cognitive test
  - Brief instructions.
  - 3 CRT items.
  - 16 ICAR sample test items.
- Page 4: Political opinions and experience with ghettos.
  - Conservatism (1-7 likert).
  - Nationalism (1-7 likert).
  - Economic liberalism (1-7 likert).
  - Personal freedom (personal liberalism) (1-7 likert).
  - Which party they would vote for if they could vote (all large parties listed; randomized order).
  - Experience with ghettos (6 options). A *ghetto* was defined as being an area where:
    1. many immigrants live,
    2. there is a high unemployment rate and
    3. there is a high crime rate.

    This is a non-quantitative definition that follows the definition of *ghetto* as used by the former Ministry of Cities, Housing and Land districts.
- Page 5: Basic demographic information.
  - Gender (biological).
  - Age.
  - Email if they wanted more information about results (optional; not published).
  - Comments (optional; not published).

2.3 Cognitive test

We used the same items as used in Kirkegaard & Nordbjerg (2015). This is a Danish translation of the *International Cognitive Ability Resource* sample test and the *Cognitive Reflection Test* (Condon & Revelle, 2014; Toplak et al., 2011). The score was calculated as the sum of correct items.

2.4 Stereotypes

Participants were asked to estimate the percent of persons aged 30-39 who were receiving social benefits for each country of origin. This age group was chosen because members of it are old enough to be finished with education and thus would not receive the State Educational Grant that all students are eligible to in Denmark (see `http://www.su.dk/english/state-educational-grant-and-loan-scheme-su/`), and not old enough that many of them would be receiving retirement benefits.

We chose this measure of socioeconomic performance because it is on a ratio scale (has a true zero) and is simple to understand. Alternatives we could have used are crime rate, mean income, educational level and general socioeconomic factor scores (Kirkegaard & Fuerst, 2014). The criterion data were from 2012, so were only a little out of date.

We note that because the groups are small in some cases, there is probably some year to year variation which acts like sampling error. Unfortunately we did not have data for multiple years so that we could aggregate data or estimate the temporal stability. This probably has the effect of somewhat decreasing the observed accuracy.

3 Quality control and descriptive statistics

The total number of survey replies was 60.

3.1 Missing data

Because we used a survey that forced the user to answer the questions before continuing, missing data should not be a problem. However, despite this, one participant managed to have massive missing data. We excluded this case, leaving N=59.

3.2 Control questions

To avoid data contamination from persons who did not understand the stereotype questions or were not truthfully filling out the survey, we excluded participants who failed the control questions. This was done by filtering out anyone who did not answer that “most men are taller than most women” and “most Europeans are taller than most East Asians”. Other answer options included “Men and women are equally tall”, “All Europeans are taller than all East Asians”. This left N=50.

3.3 Unserious age

One participant stated that his age was 420. We excluded this person, N=49.

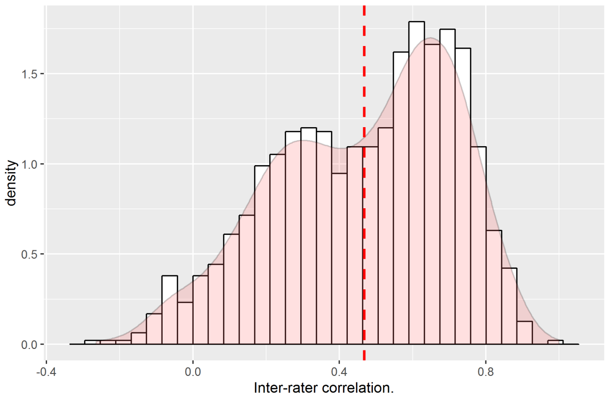

Figure 1: Distribution of inter-rater correlations.

3.4 *Unserious estimates*

2 participants gave the same estimate of social use for every country of origin (2 % and 6 %). It is possible that they really thought that the use of social benefits would be exactly identical for all groups. However, the use of a single digit suggests that they rather wanted to fill out the survey quickly. In any case, both participants failed one or more of the other controls above and so were already excluded.

3.5 *Descriptive statistics*

Table 1 shows descriptive statistics for numerical variables in the sample that passed quality control.

With regards to the non-numerical variables, they are less easy to summarize. The gender distribution was 73 % male (36 male, 13 female).

Ghetto experience: 6 % lived in a ghetto, 8 % were neighbors to one, 14 % had lived in one, 8 % has been a neighbor to one, 41 % had been in a ghetto and 22 % had no experience with ghettos.

Party support: Party voting desires tended towards nationalist, conservative and economically liberal parties: the Danish People's Party (Dansk Folkeparti) 24 %, don't know 16 %, Liberal Alliance 16 %, Venstre – Liberal Party 12 %, would not vote 8 % and the rest of the votes were split over the remaining parties.

Collection site: 180grader.dk 33 %, Gymnasie in class 20 %, Mensa 12 %, Gymnasie intranet 10 %, Emil's Facebook 10 %, Julius' Facebook 6 %, Reddit ad 4 % and physical education Facebook 4 %.

3.6 *Inter-rater consistency*

There are several measures of inter-rater consistency. Perhaps the simplest is to calculate all the correlations between raters' estimates. Figure 1 shows the distribution of rater intercorrelations.

The mean correlation is .47 while the median is .51. An alternative measure is the intraclass correlation (ICC). ICC is, unlike the correlation, sensitive to differences in means and standard deviations between raters; it is a type of absolute agreement (cf. Section 4.2). There are several variations on the measure depending on the goals and the data as implemented in the psych package for R[2]. In the case of this study, type 2 ICC is the correct choice. This type assumes that each target is rated by multiple raters, that these raters are considered a random sample from a population of raters, and that the goal is to estimate the true population ICC.

The ICC (type 2) correlation for this dataset is .31. This value is considered low by some guidelines (Hallgren, 2012). This reflects the fact that while raters showed substantial agreement in their ordering of countries and the relative distances between them, there were large inter-rater differences in the mean level of the estimates.

4 Measuring accuracy

Stereotype accuracy can be measured in multiple ways (Jussim, 2012) each of which has its advantages and disadvantages.

4.1 *Correlations*

One method is to correlate the estimates with the real values (Jussim, 2012, p. 205). This has the advantage of being on a familiar scale (-1 to 1) which can easily be compared with results from other areas of (social) science.

Correlations are based on deviation scores and so the mean estimate can be wrong while the correlation is 1.0. For instance, if the real values were 10, 20, 30, 40, 50 and the estimates were 20, 30, 40, 50, 60, the correlation would be 1.0 despite all estimates and the mean estimate being too high. This is called elevation bias or error (Jussim, 2012, p. 195). Pearson correlations are based on relative distances and so the absolute size of these do not matter. If the estimates were 0, 15, 30, 45, 60 the correlation would be 1, despite all the differences and the mean differences being too large. We might call this dispersion bias; the tendency to either overestimate or underestimate differences.

If only the order of groups matter, then the rank-order (Spearman) correlation can be used.

4.2 *Absolute delta (discrepancy) scores*

Another idea is to calculate the absolute difference between each pair of real value and estimate (Jussim, 2012, p. 317). This produces an error score and one can thus calculate the mean (or median etc.) error score. Because delta scores concern any deviation from perfect accuracy, they include both the elevation and dispersion bias components.

[2] See http://personality-project.org/r/html/ICC.html.

Table 1: Descriptive statistics for numerical variables.

Variable	mean	median	sd	mad	min	max	skew	kurtosis
Conservatism	3.9	4	1.7	1.5	1	7	0.03	−0.95
Nationalism	3.5	4	1.9	1.5	1	7	0.21	−1.05
Economic liberalism	4.8	5	1.7	1.5	1	7	−0.39	−0.88
Personal liberalism	6	6	1.3	1.5	1	7	−1.52	2.89
Age	31.4	25	14.4	11.9	16	67	0.63	−0.96
cognitive ability	12	13	4.2	5.9	5	19	0.01	−1.27

4.3 *Elevation bias*

This can be calculated by taking the mean (or another central tendency measure) of the estimates and subtracting from that the real values. A positive value indicates a tendency to overestimate values. An absolute value of this reflects the non-directional elevation error.

4.4 *Dispersion bias*

This can be calculated by taking the standard deviation (or another dispersion measure) of the estimates and subtracting that of the actual values. A positive value indicates a tendency to overestimate differences and negative the reverse. An absolute value of this reflects non-directional dispersion error.

4.5 *Which measure is the most important?*

It depends on what the goal is. In some cases, getting the order among the groups right is the most important task, in which case the rank-order (Spearman) correlation is a good measure. If relative distance matters, then Pearson's correlation is a good measure. If we are interested in whether participants tend to over- or underestimate in general, elevation error is a good measure. If we are interested in whether participants are prone to overestimating differences, then the dispersion error is a good measure. If all errors matter, then the mean absolute delta score is a good measure.

4.6 *Levels of analysis*

One can examine stereotype accuracy at two main levels (Jussim, 2012, p. 317):

1. individual-level (also called personal stereotypes)
2. aggregate-level (also called consensual stereotypes)

For instance, suppose we have two raters who each rate 5 groups on some trait. Their estimates are 10, 5, 7, 8, 5 and 11, 10, 8, 5, 8. Suppose further that the true values are 15, 4, 12, 5, 10. Using the Pearson correlation as the measure of accuracy, the raters' accuracy scores are .56 and .46, and their inter-correlation is .10. Their average accuracy is .51. If instead we aggregate the estimates first, they become 10.5, 7.5, 7.5, 6.5, 6.5. The accuracy of the aggregated estimates is .68.

5 Individual accuracy

We assessed accuracy with the following measures: Pearson correlation, rank-order correlation, mean absolute delta score, (absolute) elevation bias, and (absolute) dispersion bias. Of primary interest was the Pearson correlation. Figure 2 shows the distribution of Pearson correlations, each data point being a correlation between a participant's estimates of group values and the real group values.

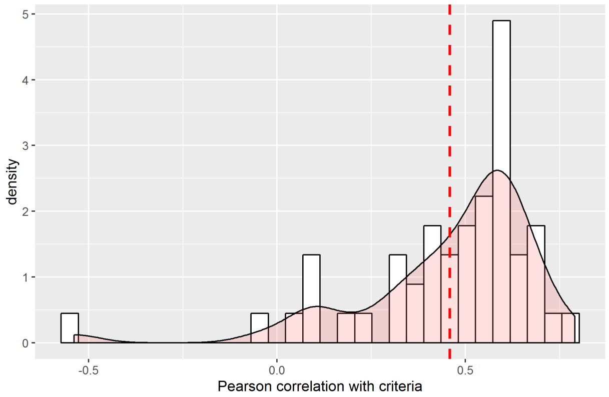

Figure 2: Distribution of Pearson correlations for individual stereotypes. The red line shows the mean.

The mean accuracy was .46 and the median was .54. It is clear that there is one extreme outlier, perhaps someone who misunderstood the survey question and filled it out reversely (i.e. estimating the percent of persons *not* receiving social benefits). We excluded this case from further analysis, yielding a mean of .48 and median of .55 [CI95 for the median: .46 to .58][3]. Using the cutoffs of .30 and .50 for levels of accuracy, as recommended in Jussim et al. (2015) 85 % and 58 % of personal stereotypes were accurate, respectively.

The Pearson correlations between measures of accuracy are shown in Table 2.

[3] The confidence interval is based on bootstrapping with 10000 runs. We used the accelerated and bias-corrected method (Efron, 1987) as implemented in the **boot** package for R (Canty & Ripley, 2015).

Table 2: Correlation matrix for individual accuracy measures. N=48. abs = absolute.

Measure	pearson r	rank r	mean abs delta	mean elevation error abs	dispersion error abs	mean elevation error	dispersion error
pearson r	1	0.96	-0.19	-0.17	-0.07	0.35	0.51
rank r	0.96	1	-0.17	-0.10	-0.11	0.34	0.47
mean abs delta	-0.19	-0.17	1	0.75	0.47	0.53	0.36
mean elevation error abs	-0.17	-0.10	0.75	1	0.35	0.10	-0.15
dispersion error abs	-0.07	-0.11	0.47	0.35	1	-0.02	0.29
mean elevation error	0.35	0.34	0.53	0.10	-0.02	1	0.74
dispersion error	0.51	0.47	0.36	-0.15	0.29	0.74	1

We see that the two correlation measures correlate near unity (.96), meaning that interval-level or rank-level accuracy was nearly the same across participants. As expected, there are negative correlations between the correlational measures and the absolute error measures (larger absolute errors means less correlational accuracy), but they are fairly weak.

It is interesting to note that both elevation and dispersion error had positive correlations with the correlation measures. In other words, participants that tended to overestimate the proportions in general and those who tended to exaggerate group differences were more accurate in identifying the relative size of group differences (r's .35 and .51).

5.1 *Correlates of accuracy*

In the survey we included a number of variables that might be related to accuracy. We focus on four accuracy measures: (Pearson) correlational accuracy, mean absolute error, elevation error (mean) and dispersion (sd) error.

5.1.1 *Continuous predictors*

The correlations between the continuous predictors and the accuracy measures are shown in Table 3. The cognitive ability score is based on a unit-weighted sum of correct items.[4]

Cognitive ability is consistently, but weakly related to better accuracy across all measures. However, the other predictors are not consistently related to accuracy. For instance, conservatism is related to *larger* absolute errors, but very weakly and *positively* to correlational accuracy. Nationalism, personal liberalism (preference for more personal freedoms) and especially age seem to be somewhat useful predictors of correlational accuracy. Interestingly, the same predictors were also useful predictors of tendencies to exaggerate group differences (dispersion error) and overestimate the proportion of persons receiving benefits.

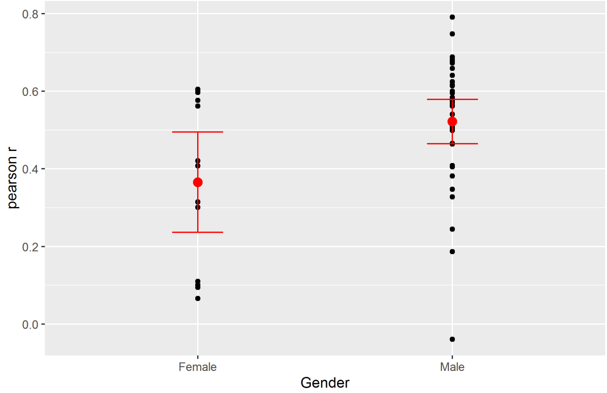

Figure 3: Correlational accuracy by gender. N=48. Red dots are the mean values, error bars are 95 % analytic confidence intervals.

5.1.2 *Categorical predictors*

The sample size was too small to examine ghetto experience, collection site and political party as predictors, but sufficiently powered to examine gender if a difference is large. Figure 3 shows the distribution of correlation accuracy by gender.

Men had higher accuracy: .52 vs. .37. This is actually a large standardized difference (d=.86 [CI95: 0.17 to 1.56]; using pooled sd). The female distribution of results is peculiar, so some caution is advised. We also note that the difference was much smaller but in the same direction using mean absolute errors (d=-.22 [CI95: -.89 to 0.45]). Recall, that lower values of mean absolute errors signal higher accuracy, so a d of -.22 is .22 in favor of male accuracy.

5.2 *Individual accuracy on average*

The mean and median accuracy scores across individuals are shown in Table 4.

Median scores were better than mean scores for most of the measures and trivially different for the two directional error measures (elevation and dispersion) indicating that a few outliers were weakening the results. The near-zero values of elevation and dispersion error indicate that our

[4] Factor analysis showed that all items load positively on a general factor as expected (Dalliard, 2013; Jensen, 1998). While we could have used a factor score, unweighted sums are more resistant to sampling error and were chosen instead. Furthermore, some participants complained about the length of the cognitive test (19 items) in the comments, so it is likely that the results are less reliable than would otherwise be expected.

Table 3: Correlations between continuous predictors and accuracy measures. N=48. The numbers in brackets show the 95 % analytic confidence intervals.

	pearson r	mean abs delta	mean elevation error	dispersion error
Conservatism	0.07 [-0.23 0.37]	0.32 [0.04 0.60]	0.40 [0.13 0.67]	0.34 [0.06 0.62]
Nationalism	0.34 [0.06 0.62]	0.08 [-0.22 0.37]	0.42 [0.15 0.69]	0.44 [0.18 0.71]
Economic liberalism	0.19 [-0.10 0.48]	0.11 [-0.18 0.41]	0.27 [-0.02 0.55]	0.37 [0.10 0.65]
Personal liberalism	0.32 [0.04 0.60]	0.00 [-0.30 0.29]	0.18 [-0.11 0.47]	0.27 [-0.01 0.56]
Age	0.56 [0.32 0.81]	0.11 [-0.19 0.40]	0.39 [0.12 0.66]	0.54 [0.29 0.79]
Cognitive ability	0.23 [-0.06 0.51]	-0.27 [-0.56 0.02]	-0.21 [-0.50 0.08]	-0.16 [-0.45 0.14]

Table 4: Mean and median accuracy scores across individuals.

Measure	Mean	Median
pearson r	0.48	0.55
rank r	0.47	0.53
mean abs delta	15.25	13.94
mean elevation error abs	9.00	8.89
dispersion error abs	7.53	6.95
mean elevation error	−0.38	−0.54
dispersion error	0.37	−0.50

participants on *average* got both the average level across groups and the differences between groups almost exactly right. In other words, they did not tend to exaggerate or minimize group differences, and neither did they tend to over or underestimate the number of persons receiving benefits.

6 Aggregate accuracy

Examining accuracy at the aggregate-level raises the question of how one should aggregate the estimates. We used the following methods: arithmetic mean, median and a 10 % trimmed mean. Results for selected accuracy measures are shown in Table 5.

The aggregate stereotypes were very accurate with a correlation of about .70. Interestingly, the arithmetic mean had the best performance, altho only slightly. This aggregate was used for further analysis. There was a slight tendency to underestimate group differences (negative dispersion error) but virtually no error in the mean level (~0 mean elevation error). Figure 4 shows a scatterplot of aggregate estimates and real values.

It is evident from the plot that some groups performed much better than estimated (e.g. Nigerians) while some performed much worse than estimated (e.g. Kuwaitis).

6.1 *Correlates of aggregate accuracy*

We did not ask participants what they based their estimates on, but it seems likely that they used a mixture of sources such as personal experience with members of the groups, media discussion of groups as well as widely known country of origin proxies (correlates) of performance such as GDP (per capita, which will be implicit in the remaining part of the paper). It is also possible to examine plausible candidates of *stereotype error* i.e. things that cause more inaccurate stereotypes (systematic error; (Jensen, 1980)).

6.1.1 *Size of the groups*

People are more likely to have personal experience with members of larger groups and based on this, the absolute errors should be smaller for these groups. Likewise, the media are probably likelier to discuss and report about members of the larger groups. Assuming that the media coverage has some validity, this should also increase accuracy for larger groups.

Population size data for the groups is available from a previous study and was merged with the dataset from a previous study (Kirkegaard & Tranberg, 2015). We used the population data from the same year as the economic data was from (2012) and we used the number for the entire population, not just the persons aged 30-39. Because the population of Denmark was an extreme outlier and because it is qualitatively different (it is not an immigrant group), we excluded it. We transformed the population data using the base-10 logarithm to make them more normal. Figure 5 shows the scatterplot of population ($\log_{10}$) and absolute error. As can be seen, there was only weak evidence for the influence of population size, at least in a correlation analysis.

6.1.2 *GDP as a proxy*

Our sample overestimated the performance of Kuwaitis and underestimated that of Nigerians, and at the same time, Kuwait and Nigeria lie at opposite ends of the GDP distribution. If participants relied on GDP of the origin country as a proxy for immigrant performance when they did not know anything else, then the estimates should be correlated with the origin countries' GDP. Furthermore and more critically, the estimate errors be correlated with GDP after GDP's relationship to the actual values has been statistically controlled.

To test the idea, we downloaded GDP (in US dollars) data for 2012 using the International Monetary Fund's *World Economic Outlook Database* (International Monetary Fund, 2015). Because GDP is not normally distributed, we used the $\log_{10}$ value.

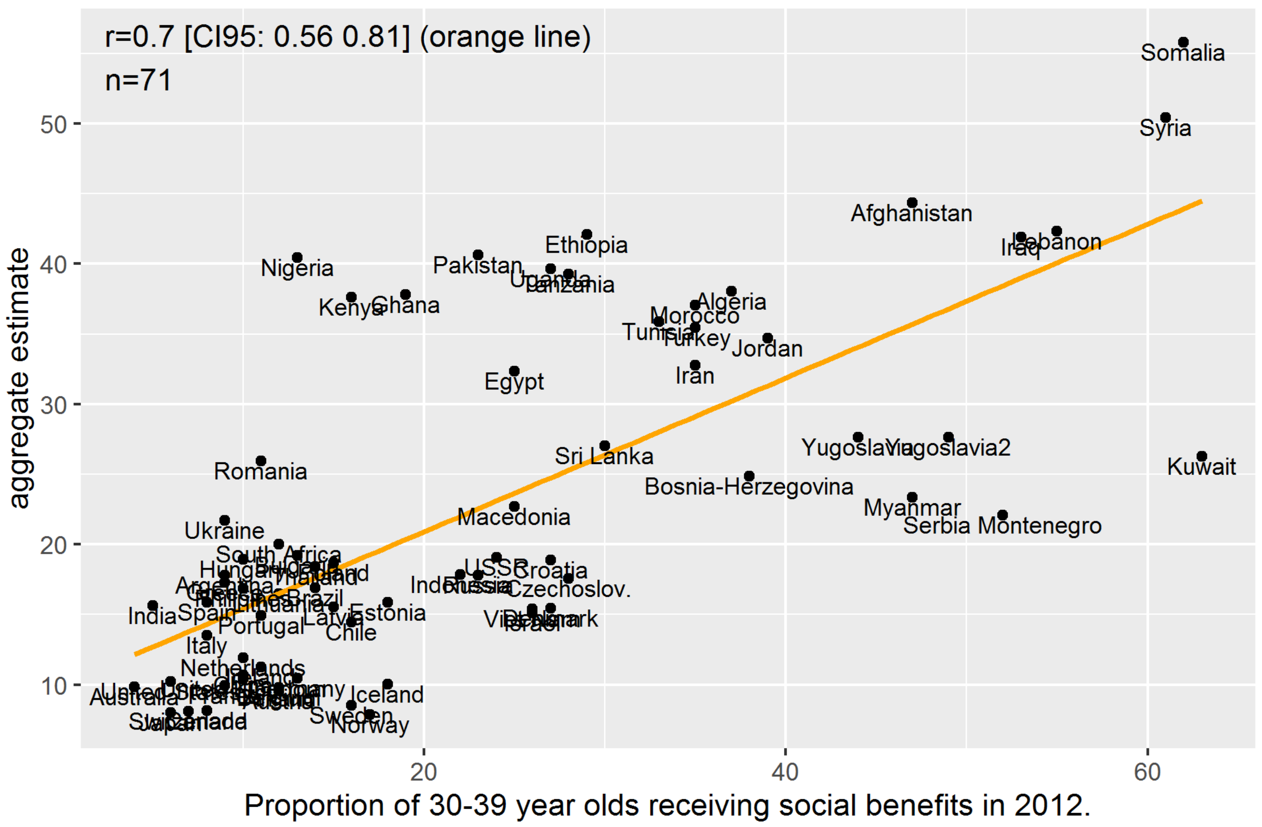

Figure 4: Scatterplot of aggregate estimates and real values. The two Yugoslavias refer to two different states existing in the same area at different times.

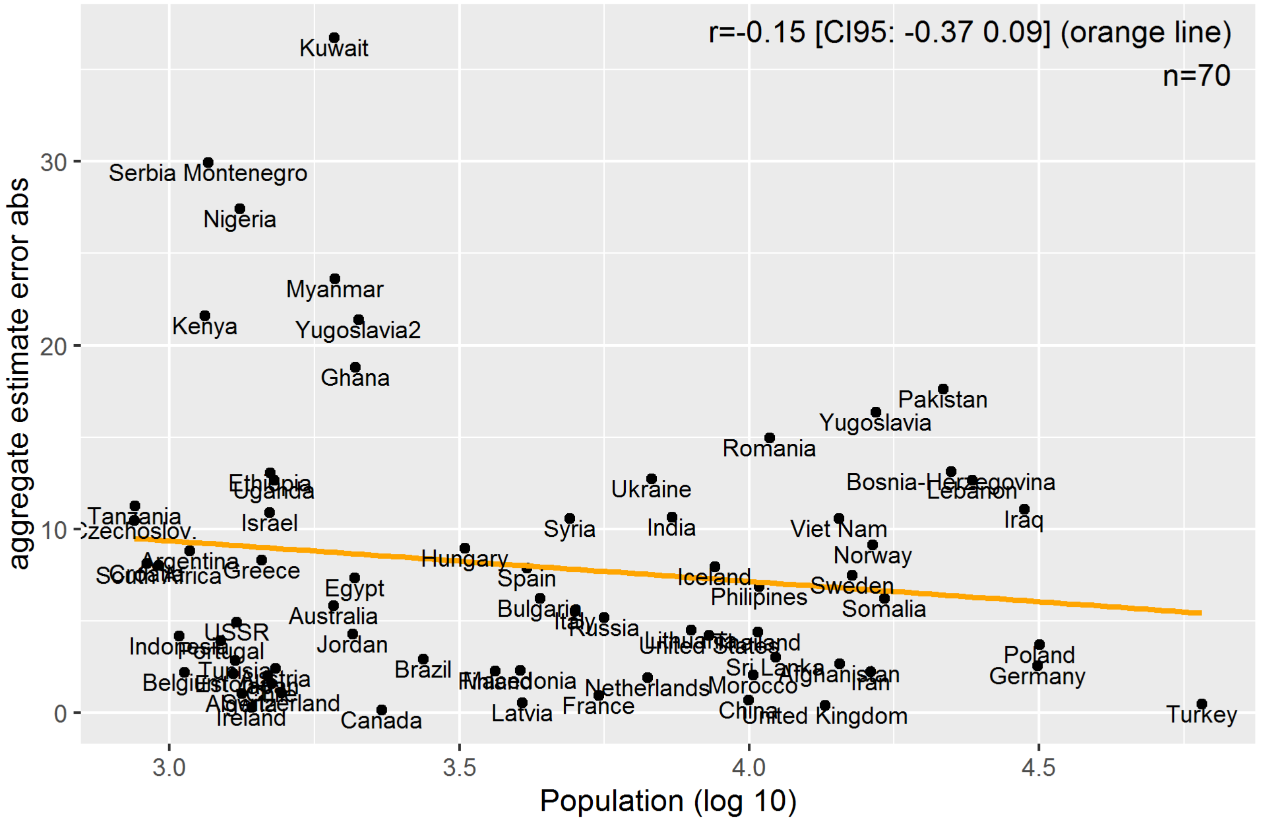

Figure 5: Population size ($\log_{10}$) and absolute aggregate stereotype error.

Table 5: Aggregation methods and their accuracy.

Method	pearson r	mean abs delta	sd	dispersion error	mean elevation	mean elevation error
mean	0.70	8.06	12.02	-3.43	22.47	-0.38
median	0.69	8.93	11.90	-3.56	18.38	-4.48
trimmed mean	0.70	8.11	12.36	-3.10	20.99	-1.87

Figure 6 shows the correlation between GDP ($\log_{10}$) and the aggregate estimates.

There is a strong negative relationship as expected. However, this could arise from GDP's association with the real values, shown in Figure 7. We see that immigrant performance is predictable from a country of origin variable, a pattern that has been replicated many times (Jones & Schneider, 2010; Fuerst & Kirkegaard, 2014; Kirkegaard & Fuerst, 2014; Kirkegaard, 2014, 2015a,b; Vinogradov & Kolvereid, 2010).

However, we also see that GDP's actual relationship to the real values is weaker (r=.39) than its relationship to the stereotypes (r=.76). But the best way to examine whether stereotypes are based on GDP, is to partial out GDP's relationship to the actual values and correlate the residuals with the estimate errors. Figure 8 shows a scatterplot of this relationship. We still see a strong correlation after removing the GDP x actual values relationship, in line with the GDP as proxy hypothesis. As expected, Kuwait and Nigeria lie in the corners.

6.1.3 Muslims as a source of stereotype bias

Public debate in Denmark about immigrants often concern the role of Islam or Muslims (Engelbreth Larsen, 2009; Fruensgaard, 2012; Holstein & Jenvall, 2014; Lassen, 2014) probably because immigrants from countries with more Muslims tend to fare poorly as measured by standard socioeconomic outcomes such as education, income, crime and use of social benefits. The correlation between percent Muslim in the origin country and a general index of socioeconomic outcomes is -.78 (N=58) (Kirkegaard & Fuerst, 2014, Table 12). For this reason, it is important to examine whether stereotypes are more or less accurate for these groups. To do this, we used percent Muslims in the origin countries from Pew Research's 2010 survey (Pew Research Center, 2011). This is probably a reasonable proxy for percent Muslims among these groups in Denmark, altho we are not aware of any studies of this. Figure 9 shows a scatterplot between percent Muslim in the origin country and aggregate estimate error, while Figure 10 shows that with *absolute* aggregate estimate error.

In neither case do we see any reliable pattern. If anything, there might be a slight tendency for more Muslim populations to fare worse than stereotypes about them.

6.1.4 Country of origin diversity and accuracy

In reviewing the paper (`http://openpsych.net/forum/showthread.php?tid=256&pid=3888#pid3888`), Peter Frost suggested that some countries of origin are more ethnically/racially diverse than others and that this might affect accuracy. To test this idea, we merged the present dataset with measures of ethnic, linguistic and religious diversity from Alesina et al. (2003) and from Fearon (2003). This yielded a total of 5 indicators. From these, we created a general measure using factor analysis (default settings of the *fa* function in package **psych**, (Revelle, 2015)).

Table 6 shows the correlations between the diversity measures and two measures of accuracy.

There were no strong linear relationships between estimate error (directional error) and diversity measures, but there were moderate to strong relationships to absolute error. Figures 11 and 12 show scatter plots of general diversity as well as estimate error and absolute estimate error, respectively. From both plots it can be seen that the relationship is not homoscedastic, meaning that there is more variation as for some values (higher diversity in this case) of the predictor than others. Still, there seems to be a genuine relationship such that stereotypes for more diverse countries are less precise.

7 Discussion and conclusion

Stereotypes are often mentioned in public discussion about immigration in Denmark. They have been proposed as a cause of group differences (Sareen, 2011) and unfair treatment such as housing discrimination (Ekberg, 2015; Hussein, 2014). Rarely is it considered that they may instead reflect group differences. We found that personal stereotypes about country of origin groups tend to be accurate (median correlation = .55), especially at the aggregate level (r = .70). Stereotypes did not generally exaggerate real differences and in case of the aggregate estimates they somewhat underestimated them (by -3.43 percent points).[5] These findings are similar to previous findings concerning demographic stereotypes (Jussim, 2012; Jussim et al., 2015).

We found some evidence that some variables are associated with stereotype accuracy. For instance, observed age had a correlation of 0.56 [CI95: 0.32 0.81] with correlational accuracy. If real, it is unclear whether this is an age or cohort effect. In general, we do not draw strong conclusions

[5] *The Good Judgment Project* also found that aggregation estimates tended to be too moderate, i.e. not extreme enough, which is the same as having negative dispersion error (Tetlock & Gardner, 2015; Ungar et al., 2012).

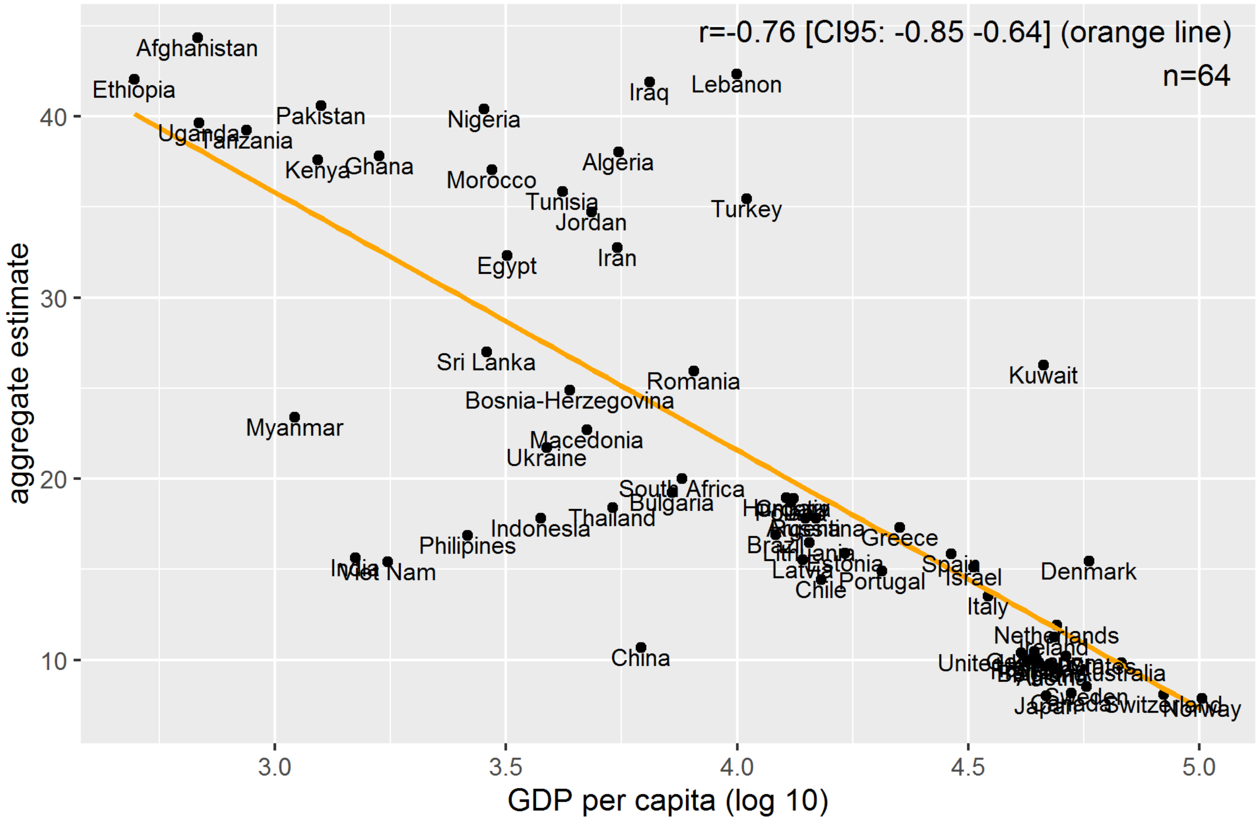

Figure 6: Scatterplot of GDP per capita ($\log_{10}$) and aggregate estimate.

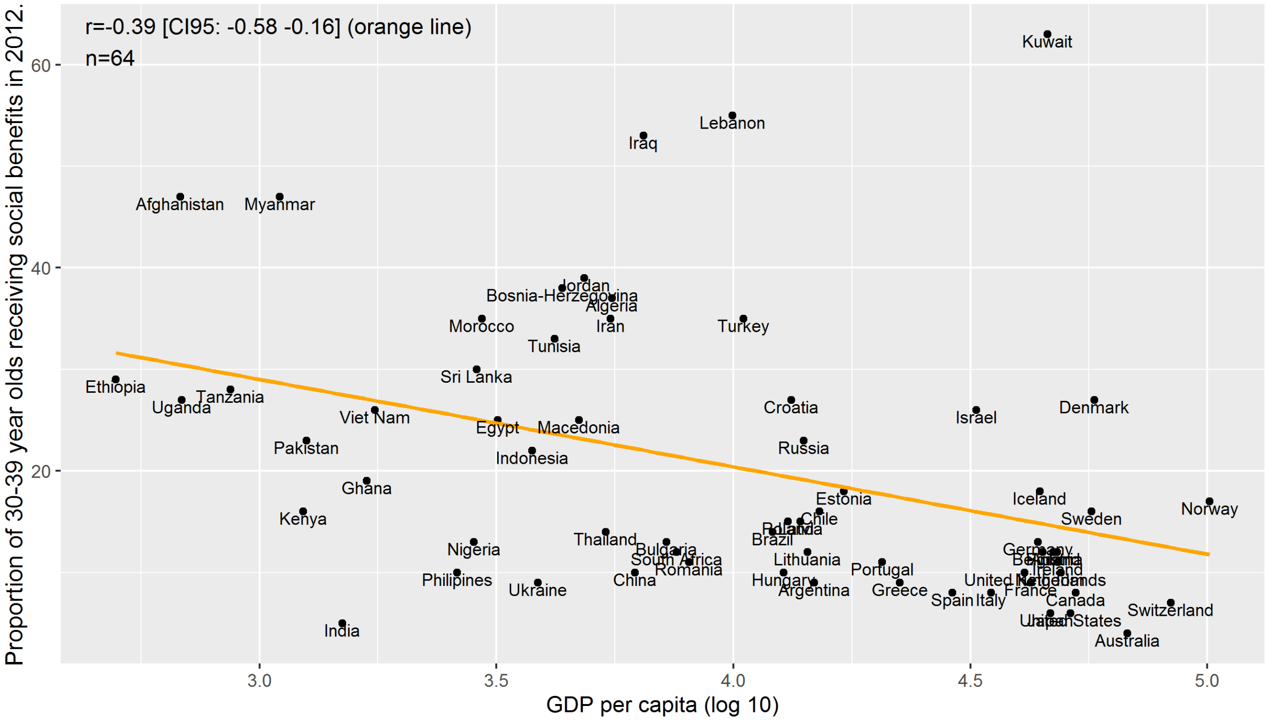

Figure 7: Scatterplot of GDP per capita ($\log_{10}$) and proportion of 30-39 year olds receiving social benefits in 2012.

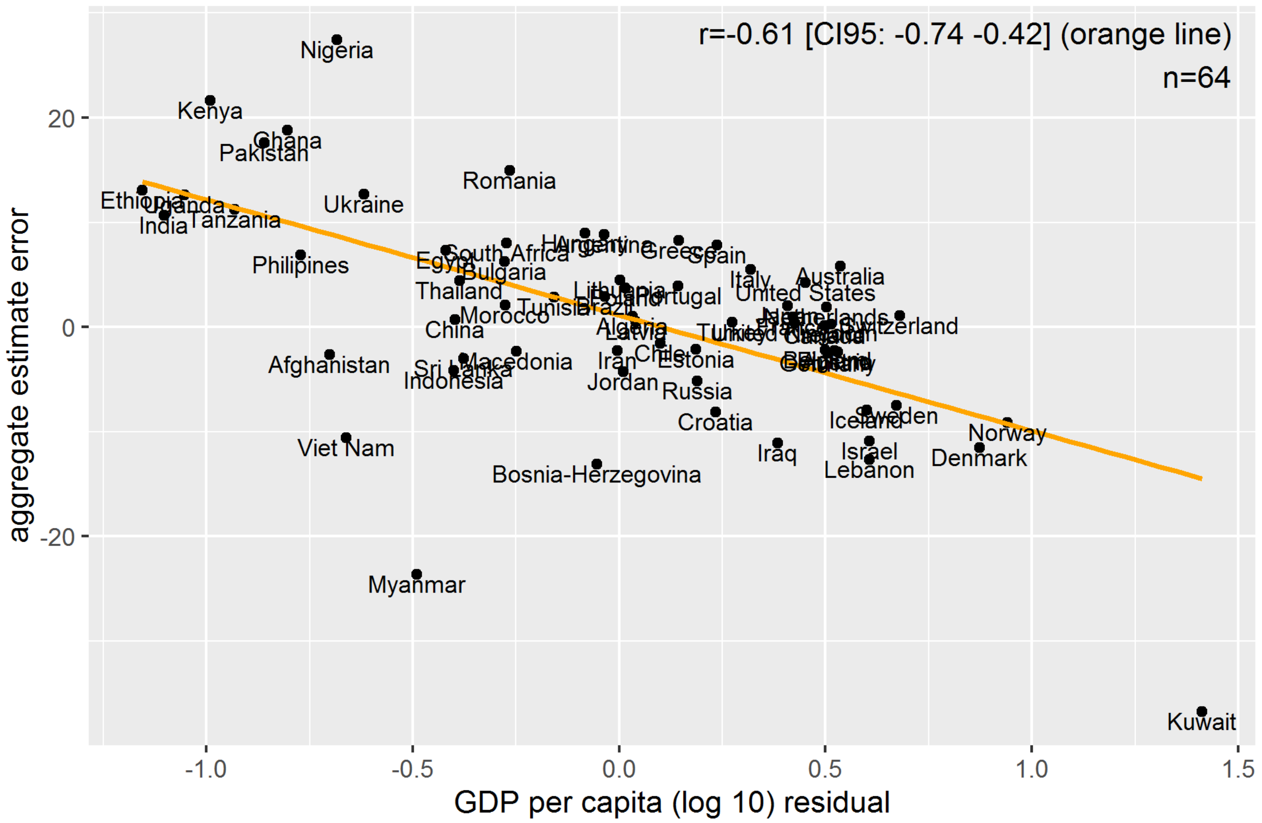

Figure 8: Scatterplot of GDP per capita ($\log_{10}$) residuals after removing actual values and the aggregate estimate errors.

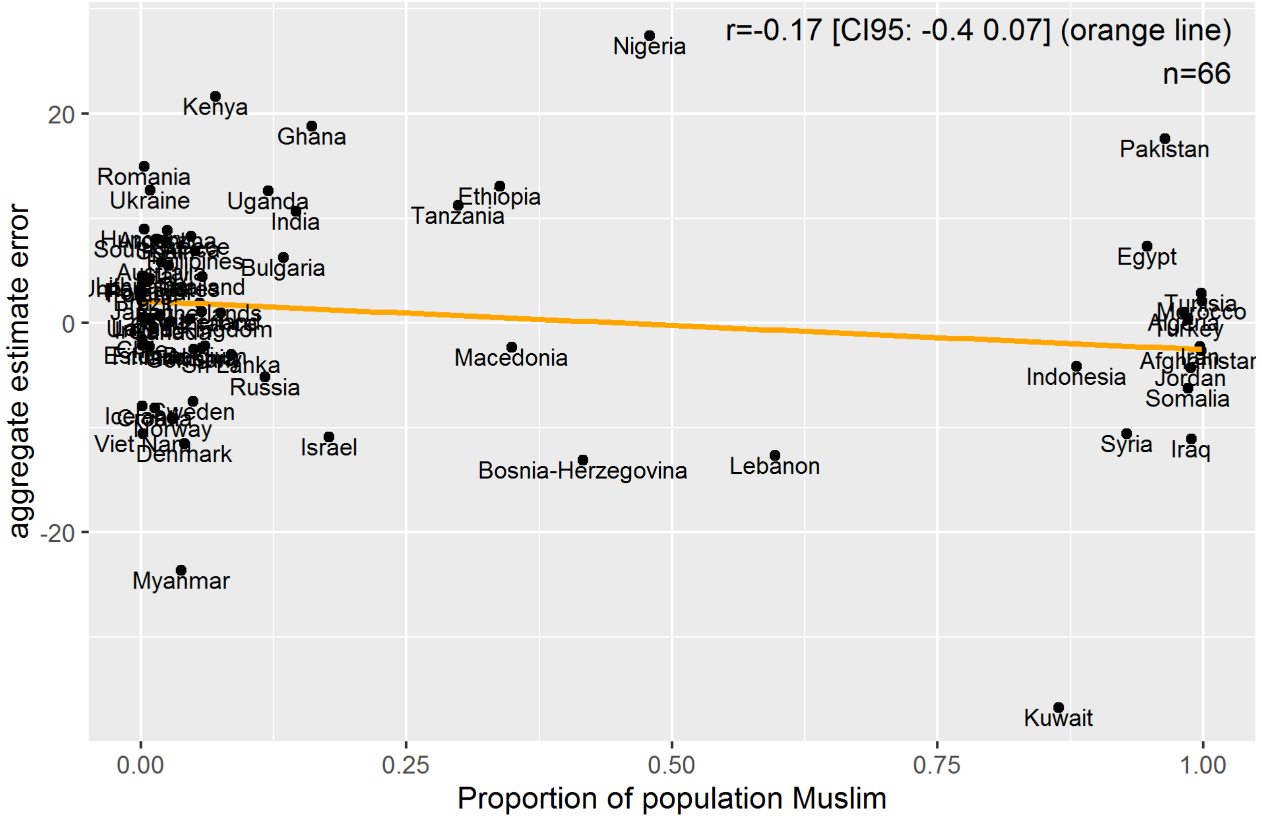

Figure 9: Proportion of population Muslim in country of origin and aggregate estimate error.

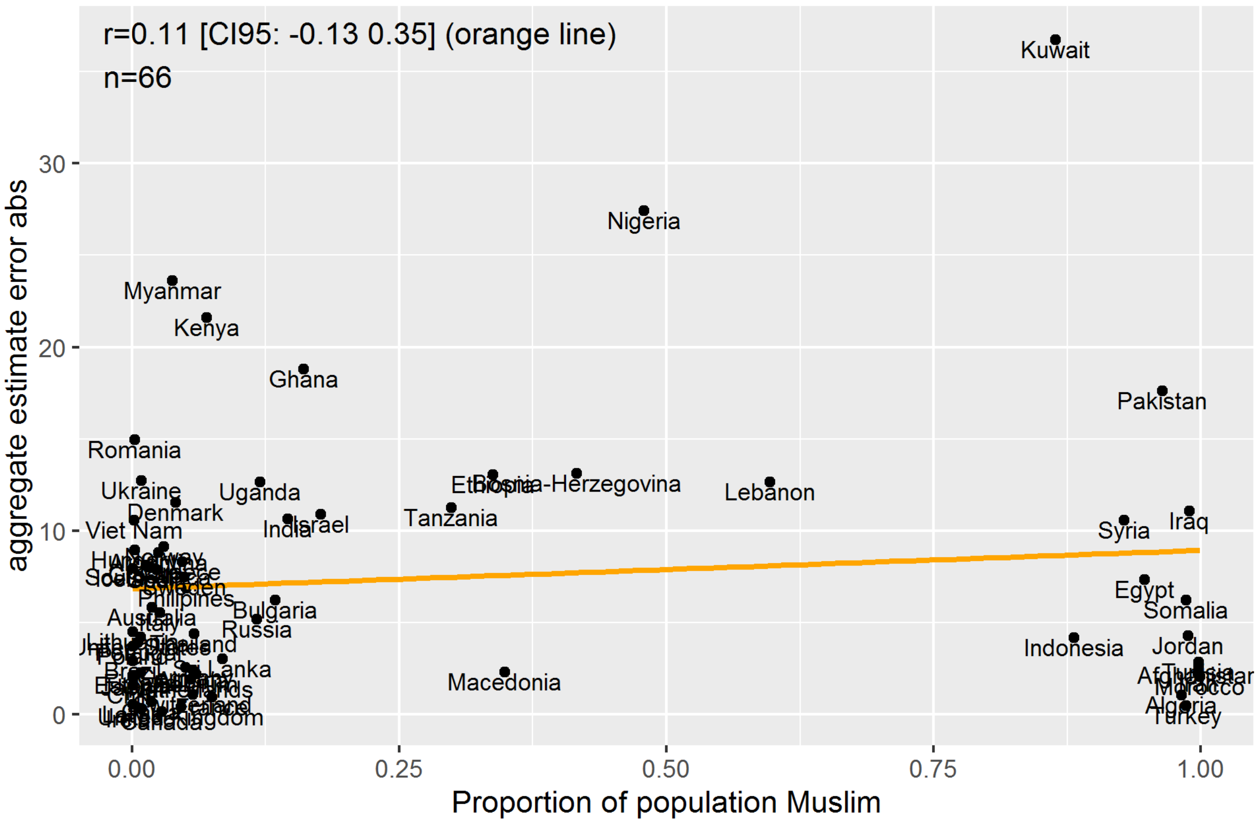

Figure 10: Proportion of population Muslim in country of origin and absolute aggregate estimate error.

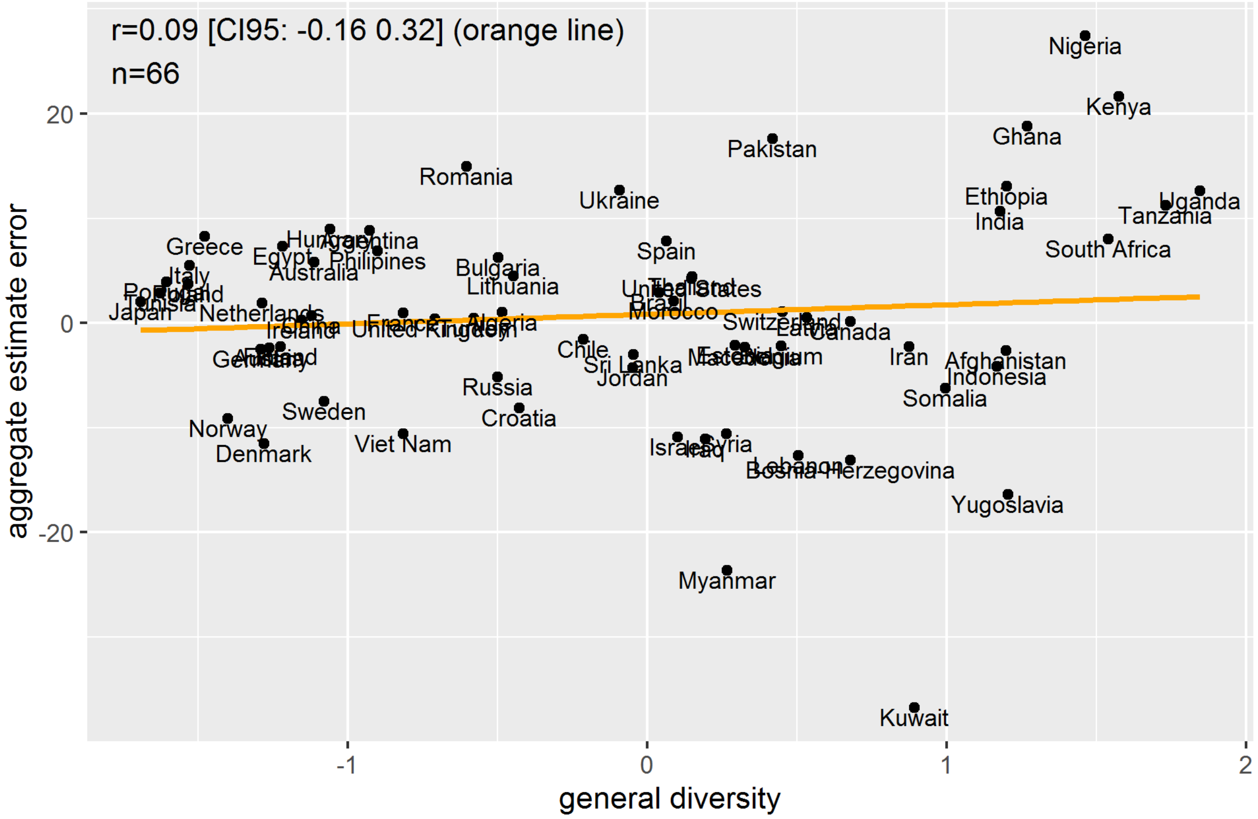

Figure 11: General diversity and aggregate estimate error.

Table 6: Measures of diversity and aggregate stereotype accuracy.

Predictor	aggregate estimate error	aggregate estimate error abs
EthnicFractionizationIndexFearon03	0.01	0.41
CulturalDiversityIndexFearon03	0.07	0.32
EthnicFractionalizationAlesina03	0.15	0.39
LinguisticFractionalizationAlesina03	0.29	0.31
ReligiousFractionalizationAlesina03	0.15	0.28
general diversity	0.09	0.42
aggregate estimate error	1	−0.22
aggregate estimate error abs	−0.22	1

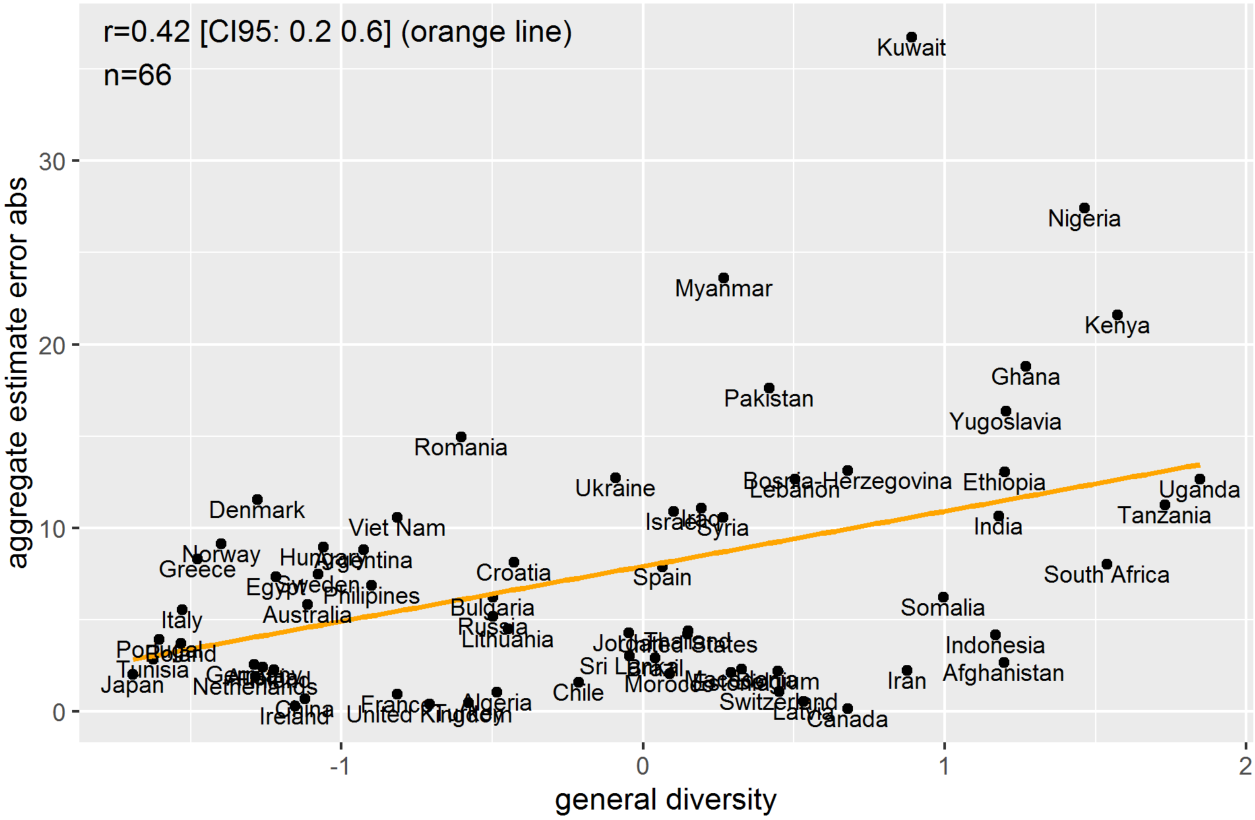

Figure 12: General diversity and aggregate absolute estimate error.

from the analyses of accuracy predictors because the sample was both small and unrepresentative.

We found evidence that stereotypes were derived from participants' estimates of the countries of origin's wealth (GDP per capita). We see two ways to further examine this question. First, one could ask participants to estimate the countries' GDP. If perceptions of GDP are used as proxies, the perceptions should be more closely related to the estimates than to the actual GDP values. Second, one can ask subjects what they base their estimates on. It is important to ask these questions *after* having the participants estimate the group performances in order not to influence the estimates.

Because we share all the data and R code, our results are open to re-analysis or re-use by other researchers.

7.1 Replication

We would like to replicate and expand the study with a larger and more representative sample. However, this requires more funding than the authors can afford out of pocket, so we would like to hear from persons interested in contributing the necessary funds.

7.2 Limitations

- Due to the non-random sampling and relationships between predictors, the findings should be seen as preliminary. Non-random sampling can produce or eliminate relationships between predictors which can result in spurious correlations or suppressing real correlations.

- The limited sample size makes conclusions uncertain, especially about correlates of (in)accuracy.
- Criterion data was only available for one year (2012) which was somewhat removed from the year of the survey (2015) and has unknown temporal stability. This probably somewhat decreases the observed accuracy scores.

Supplementary material and acknowledgements

Open peer review forum thread: http://openpsych.net/forum/showthread.php?tid=256.

Supplementary materials including the translated questionnaire, complete case-level data, R code and high quality figures can be found at the Open Science Framework repository: https://osf.io/ecpzj/.

We thank Gerhard Meisenberg, Peter Frost, Sean Stevens, Meng Hu, John Fuerst and especially Marc Dalliard for useful reviews.

References

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, *8*(2), 155–194. Retrieved from http://dx.doi.org/10.1023/A:1024471506938 doi: 10.1023/A:1024471506938

Canty, A., & Ripley, B. (2015). *Bootstrap functions (originally by angelo canty for s). cran.* Retrieved from http://CRAN.R-project.org/package=boot

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52 - 64. Retrieved from http://www.sciencedirect.com/science/article/pii/S0160289614000051 doi: http://dx.doi.org/10.1016/j.intell.2014.01.004

Dalliard. (2013). *Is psychometric g a myth? human varieties.* Retrieved from http://humanvarieties.org/2013/04/03/is-psychometric-g-a-myth/

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171-185. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478410 doi: 10.1080/01621459.1987.10478410

Ekberg, J. (2015, April 27). *Mustapha om boligmarked: Føler mig forulempet - mx.dk.* Retrieved from http://www.mx.dk/nyheder/danmark/story/21254390 (Retrieved December 10, 2015)

Engelbreth Larsen, R. (2009, April 3). *Muslimer, indvandrere og kriminalitet.* Retrieved from http://www.humanisme.dk/artikler/kronik075.php (Retrieved December 10, 2015)

Fearon, J. D. (2003). Ethnic and cultural diversity by country*. *Journal of Economic Growth*, *8*(2), 195–222. Retrieved from http://dx.doi.org/10.1023/A:1024419522867 doi: 10.1023/A:1024419522867

Fruensgaard, J. (2012, February 15). *Muslimer er stærkt overrepræsenterede i europas fængsler.* (URL http://www.kristeligt-dagblad.dk/udland/muslimer-er-starkt-overreprasenterede-i-europas-fangsler. Retrieved December 10, 2015)

Fuerst, J., & Kirkegaard, E. O. W. (2014). Do national iqs predict u.s. immigrant cognitive ability and outcomes? an analysis of the national longitudinal survey of freshman. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/04/do-national-iqs-predict-u-s-immigrant-cognitive-ability-and-outcomes-an-analysis-of-the-national-longitudinal-survey-of-freshman/

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23–24.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010, 6). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83. Retrieved from http://journals.cambridge.org/article_S0140525X0999152X doi: 10.1017/S0140525X0999152X

Holstein, E., & Jenvall, L. (2014, December 2). *Højere kriminalitetsrate hos muslimske grupper.* Retrieved from http://www.altinget.dk/artikel/hoejere-kriminalitetsrate-hos-muslimske-grupper (Retrieved December 10, 2015)

Hussein, T. Z. (2014, March 22). *Boligmarkedet er lukket land for en perker som mig.* Retrieved from http://politiken.dk/debat/ECE2241792/boligmarkedet-er-lukket-land-for-en-perker-som-mig/ (Retrieved December 10, 2015)

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

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, Connecticut: Praeger.

Jones, G., & Schneider, W. (2010). Iq in the production function: Evidence from immigrant earnings. *Economic Inquiry*, *48*(3), 743–755. Retrieved from http://doi.org/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., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. *Handbook of prejudice, stereotyping, and discrimination*, 608.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in)accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. Retrieved from http://doi.org/10.1177/0963721415605257

Kirkegaard, E. O. (2014). Crime, income, educational attainment and employment among immigrant groups in norway and finland. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/10/crime-income-educational-attainment-and-employment-among-immigrant-groups-in-norway-and-finland/

Kirkegaard, E. O. (2015a). Crime among dutch immigrant groups is predictable from country-level variables. *Open Differential Psychology*.

Kirkegaard, E. O. (2015b). Immigrant gpa in danish primary school is predictable from country-level variables. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2015/06/immigrant-gpa-in-danish-primary-school-is-predictable-from-country-level-variables/

Kirkegaard, E. O., & Fuerst, J. (2014). Educational attainment, income, use of social benefits, crime rate and the general socioeconomic factor among 70 immigrant groups in denmark. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/05/educational-attainment-income-use-of-social-benefits-crime-rate-and-the-general-socioeconomic-factor-among-71-immmigrant-groups-in-denmark/

Kirkegaard, E. O., & Nordbjerg, O. (2015). Validating a danish translation of the international cognitive ability resource sample test and cognitive reflection test in a student sample. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2015/07/validating-a-danish-translation-of-the-international-cognitive-ability-resource-sample-test-and-cognitive-reflection-test-in-a-student-sample/

Kirkegaard, E. O., & Tranberg, B. (2015). Increasing inequality in general intelligence and socioeconomic status as a result of immigration in denmark 1980-2014. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2015/03/increasing-inequality-in-general-intelligence-and-socioeconomic-status-as-a-result-of-immigration-in-denmark-1980-2014/

Lassen, H. (2014, December 29). *Årsrapporten om indvandrere i danmark 2014 fejltolket igen - denfri.dk.* (URL https://www.denfri.dk/2014/12danske-medier-forstaar-ikke-statistik-igen/. Retrieved December 10, 2015)

Pew Research Center. (2011, January 27). *Table: Muslim population by country.* Retrieved from http://www.pewforum.org/2011/01/27/table-muslim-population-by-country/ (Retrieved December 10, 2015)

Revelle, W. (2015). *psych: Procedures for psychological, psychometric, and personality research (version 1.5.4).* Retrieved from http://cran.r-project.org/web/packages/psych/index.html

Sareen, M. (2011, November 30). *Moderne ligestilling er også for mænd.* Retrieved from http://www.information.dk/286459 (Retrieved December 10, 2015)

Tetlock, P., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Random House.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289.

Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012, June). *The good judgment project: A large scale test of different methods of combining expert predictions* (Tech. Rep.). AAAI Technical Report,(FS-12-06). Retrieved from https://www.aaai.org/ocs/index.php/FSS/FSS12/paper/viewFile/5570/5871

Vinogradov, E., & Kolvereid, L. (2010). Home country national intelligence and self-employment rates among immigrants in norway. *Intelligence*, *38*(1), 151–159. Retrieved from http://doi.org/10.1016/j.intell.2009.09.004