Submitted: 8th of May 2016
Published: 3rd of November 2016

The OKCupid dataset: A very large public dataset of dating site users

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

Open Differential Psychology

Abstract

A very large dataset (N=68,371, 2,620 variables) from the dating site OKCupid is presented and made publicly available for use by others. As an example of the analyses one can do with the dataset, a cognitive ability test is constructed from 14 suitable items. To validate the dataset and the test, the relationship of cognitive ability to religious beliefs and political interest/participation is examined. Cognitive ability is found to be negatively related to all measures of religious belief (latent correlations -.26 to -.35), and found to be positively related to all measures of political interest and participation (latent correlations .19 to .32). To further validate the dataset, we examined the relationship between Zodiac sign and every other variable. We found very scant evidence of any influence (the distribution of p-values from chi square tests was flat). Limitations of the dataset are discussed.

Keywords: open data, big data, OKCupid, dating site, cognitive ability, IQ, intelligence, scale construction, religiosity, politics, astrology, Zodiac sign

1 Introduction

Despite many years of advocacy of proponents, it is still uncommon for social scientists to publicly share their datasets and even sharing data on request is rare (Krawczyk & Reuben, 2012; Savage & Vickers, 2009; Tenopir et al., 2011; Wicherts et al., 2011). Worse, there is some evidence which indicates that those who refuse to share data upon request make more statistical errors than those who share data (Wicherts et al., 2011). This is doubly problematic because the mistakes cannot be corrected by other researchers. Furthermore, a given dataset may have many uses not all of which are known to those who collected it. Even if they are known, the collectors may not be interested in them (or even interested in hiding the results (Duarte et al., 2015)), or they may simply not have enough time. Whichever it is, if the analyses are not done by the collectors and the data are unavailable to anyone else, the data is not used to its full extent. Because science is usually funded by the public, this wastes an incredible amount of public funds.

The lack of data sharing probably slows down the progress of science immensely because other researchers would use the data if they could. Supporting this, it can be observed that known public datasets enjoy widespread use. For instance, Project TALENT is a very large (N≈440,000) dataset of American high school students collected in 1960. The dataset is still used by researchers examining questions never conceived of when it was collected. According to Google Scholar (https://scholar.google.com), there were 255, 148 and 116 papers published in 2013, 2014 and 2015 that had "project talent" in their text[1], indicating that it can still be useful despite being about 56 years old. For instance, (Dunkel, 2014) used the data to examine the cognitive ability of children by the language spoken by their parents, while Major et al. (2014) examined the linearity of relationships between cognitive ability and personality traits, and Damian & Roberts (2015) examined whether birth order showed relationships to cognitive ability and personality.

The NLSY79, CNLSY and NLSY97[2] are nationally representative (using probability sampling) longitudinal

* Ulster Institute for Social Research, United Kingdom. E-mail: emil@emilkirkegaard.dk
† E-mail: juliusdb.science@gmail.com

[1] The seeming decline in uses over time is perhaps an effect of the way Google Scholar finds articles. Many newer articles have probably not been fully indexed yet, hence lowering the numbers.

[2] *NLSY* stands for *National Longitudinal Survey of Youth*, the

datasets about US citizens. As with Project TALENT, they enjoy widespread use by researchers. According to Google Scholar, the number of papers using the term "NLSY" were 991, 945 and 994 in 2013-2015. For instance, Britt et al. (2013) examined the relationship between locus of control and financial behavior, Hernandez & Pressler (2014) examined the relationships between childhood obesity, young adult obesity and demographic categories, and Rodgers et al. (2015) linked biological relatives so as to enable one to do a behavioral genetic analysis of the relationship between behavioral problems and the timing of menarche.

For reasons that will be given in the next section, OKCupid is an attractive site to gather data from, and in light of the above considerations regarding the use of open datasets, we decided to gather data and make them publicly available. The purpose of this article is to describe the data collection process including sampling procedures and present some example analyses done using the dataset to showcase its usefulness for psychological research. Our hope is that others will use the dataset for their own purposes and do so in a transparent way that allows for large-sample, reproducible research.

2 Why gather data from OKCupid?

To see why OKCupid is attractive for data gathering, the simplest way is to create a user oneself on the website and have a look around. Like other dating sites, OKCupid has a matching algorithm that matches users based on their common questions answered as well as demographic information (age, gender, sexual orientation, location etc.). More concretely, users select their answers to hundreds if not thousands of questions, which answers they would accept in a potential match, and the importance of each question. The algorithm then finds persons whose answers match the desired answers and weighs them by the given importance scores. The exact algorithm is secret, but an overall explanation of it can be found at https://www.okcupid.com/help/match-percentages.

When users answer questions, they can opt to do so privately (others cannot see their answers to specific questions), but the default setting is to do it publicly and the vast majority of users answer questions publicly. This means that every other user can see each answer they have given to each question, thus making it possible to automatically gather the data. Unfortunately, the importance scores and the answers they accept from potential matches are not immediately publicly available, so cannot easily be gathered.[3]

The questions on the site are mostly user-generated, but some initial questions were created by the staff. This means that unlike questions asked by scientists, the questions concern many odd domains not normally considered by scientists. Likewise, the questions do not include many that scientists would have included such as items from standard personality inventories. By default, the questions are presented to users in the order of the questions having the most answers already. This minimizes the number of questions that a new user has to answer out before having many in common with other users, but has the cost of starkly decreasing the diversity of questions answered. Still, because users often answer hundreds if not thousands of questions, a great amount of data is available. It is also possible to answer any question by going directly to it, or finding it in some other user's profile.

Some may object to the ethics of gathering and releasing this data. However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form.

3 The data collection method

We convinced a friend to write a scraper (a program that automatically gathers data from websites) in Python based on Scrapy (http://scrapy.org/). Due to a mostly cosmetic change in the site at a later point, the scraper no longer works, but is made publicly available in case someone wants to modify it and collect more data (https://github.com/Deleetdk/OKCubot).

Initially, the scraper used a decidedly non-random approach to find users to scrape because it selected users that were suggested to the profile the bot was using. This resulted in an oversampling of females located near Denmark because a male Danish profile was used. On inspection of the data collected, we saw the problem and switched to gathering data completely at random using a feature on the site that no longer exists. After collecting some data this way, we examined it and found that the most users had answered relatively few questions, thus making it inefficient to gather data from them. For this reason, we decided to only gather data from users who had answered at

number indicates the year it was begun (1979 and 1997) and the *C* in the second indicates that it is the children of the women in the first dataset. The datasets can be found at http://www.bls.gov/nls/.

[3] It is possible to gather them if one employes a team of scrapers. This is because when one has selected an answer that another user has marked as unacceptable, it is given a red color. Thus, one would have to make a team of 4 accounts which have chosen all the possible answers so that one can see which are acceptable to a given user.

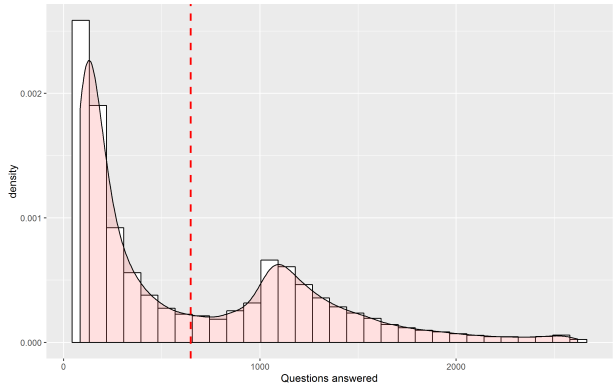

Figure 1: Density-histogram of the numbers of questions answered by users. The vertical line shows the mean.

least 1,000 questions but who were otherwise sampled at random. Figure 1 shows density-histogram of the number of questions answered by the users in the dataset.

It can clearly be seen that the distribution of questions answered is bimodal. The leftmost peak concerns the data that were gathered without filtering the users for number of answered questions, while the rightmost peak belongs to users who were specifically sampled because they had answered at least 1000 questions.

The data were gathered in the period November 2014 to March 2015.

3.1 Which data was collected

We did not gather all the possible data about the users. Specifically, we gathered the following datapoints:

- Profile information: username, age, gender(s), location, religion-related opinions, astrology-related opinions, interested in, number of photos, etc. (36 variables)
- OKCupid personality scales (50 variables)
  - These are personality dimensions that OKCupid calculates automatically. No information is given about how they are calculated as far as we know.
- Answers given to the top 2600 questions on the site.

Data we did not gather include:

- The profile text.
- The profile photos.
- Explanations given to chosen answers.
- Profile height.

Gathering the photos would have taken up a lot of hard drive space but could be done in a future scraping. Be advised that scraping and releasing users' photos may be illegal due to copyright or privacy laws. The other data were not collected because we forgot to include them in the scraper.

After collecting the data, they were processed in R to create one large datafile. Either during the data collection or the processing, some mistakes were made that left some variables corrupted. This left of total of 2541 questions, 50 personality scales and 29 variables related to each profile with uncorrupted data. The corrupted variables were the variables that had the lowest response rates (N<100), so they were nearly useless for analysis anyway.

Due to privacy concerns (Hackett, 2016), the username and city variables were removed from the published version of the dataset.

4 Descriptive analyses of the data

There are 68,371 unique users (by username[4]) in the dataset and 2620 variables. Almost no users have complete data and the data are not missing at random, but in systematic fashion due to the way the questions are presented by default (questions with most answers are presented first), cf. Figure 1.

4.1 Location

Table 1 gives an overview of the countries represented in the data.

As can be seen in the table, due to the very large number of users from the US and Canada, OKCupid treats their first order administrative divisions (states and provinces) as countries on their own. The full list of sample sizes can be found in the supplementary materials (sheet *Countries*).

Dating sites in general have more males than females, reflecting the mating behavior seen offline (more males being on the lookout). OKCupid features a very broad selection of possible genders. One must choose at least one category and up to five categories

[4] It is possible that some cases are the same person under multiple usernames. This can happen if a person creates more than one user (as we did for the purpose of getting the data), or if they change their usernames during the data collection phase (so the same data would appear under multiple names). However, this is probably a small problem. Only paying users can change their usernames.

Table 1: Sample size by country/state/province. Two letter names are US states except for UK which is the United Kingdom. Ontario, British Columbia, Alberta, Quebec, and Manitoba are Canadian provinces.

Rank	Name	N	Rank	Name	N	Rank	Name	N
1	UK	8438	31	MO	535	61	ME	154
2	NY	7989	32	TN	493	62	China	147
3	CA	7145	33	Ireland	462	63	Spain	138
4	TX	2768	34	Israel	457	64	ID	132
5	FL	1760	35	British Columbia	452	65	MS	131
6	Australia	1678	36	CT	437	66	Belgium	120
7	NJ	1662	37	Sweden	419	67	Mexico	114
8	IL	1560	38	India	391	68	Japan	112
9	PA	1533	39	OK	351	69	Indonesia	108
10	WA	1391	40	NV	327	70	Malaysia	108
11	MA	1345	41	SC	324	71	VT	108
12	LA	1337	42	DC	310	72	WV	101
13	Ontario	1059	43	KY	309	73	Russia	100
14	OH	1048	44	Philippines	293	74	Austria	94
15	Germany	1021	45	UT	279	75	Switzerland	91
16	VA	967	46	France	271	76	Norway	89
17	MI	962	47	KS	267	77	New Zealand	88
18	Netherlands	904	48	Alberta	261	78	MT	85
19	GA	895	49	Quebec	258	79	South Africa	85
20	OR	877	50	IA	249	80	Hong Kong	81
21	Denmark	868	51	AL	233	81	United Arab Emirates	79
22	NC	835	52	Brazil	216	82	DE	76
23	CO	779	53	Finland	186	83	AK	75
24	MD	779	54	NM	183	84	South Korea	73
25	MN	713	55	NH	174	85	Greece	71
26	AZ	670	56	Turkey	169	86	Taiwan	71
27	Italy	656	57	NE	165	87	Thailand	66
28	WI	581	58	RI	165	88	Romania	65
29	IN	579	59	AR	163	89	Portugal	58
30	Singapore	547	60	HI	154	90	Manitoba	56

of which the possible options are: Man, Woman, Agender, Androgynous, Bigender, Cis Man, Cis Woman, Genderfluid, Genderqueer, Gender Nonconforming, Hijra, Intersex, Non-binary, Other, Pangender, Transfeminine, Transgender, Transmasculine, Transsexual, Trans Man, Trans Women and Two Spirit. Nevertheless, almost everybody chooses one of the first two (39.1 % Women, 60.6 % Men, binary total = 99.7 %)[5]. The full count by type can be found in the supplementary materials (sheet *Genders*).

[5] The next two largest categories were Cis Woman and Cis Man, which is the same as Woman and Man. Note that users with non-binary genders or sexual orientations can choose to hide their profiles from regular users. Since the scraper user was a heterosexual male, these users would not be included in the dataset. It is unknown how many users hide their profiles, so the 99.7 % figure should be cautiously interpreted.

4.2 Age

Dating sites generally feature a younger audience. The distribution of the data by gender is shown in Figure 2.

As expected, the sample is somewhat younger than the general population in the sampled countries (mostly Western European countries and the countries they settled).

4.3 Variable types

The variables are not coded by type on the website. However, from inspecting them, it is easy to see that many of them are ordinal-level, while some others are nominal. A few are ratio scale (e.g. age). JDB undertook the large, manual task of reading every question

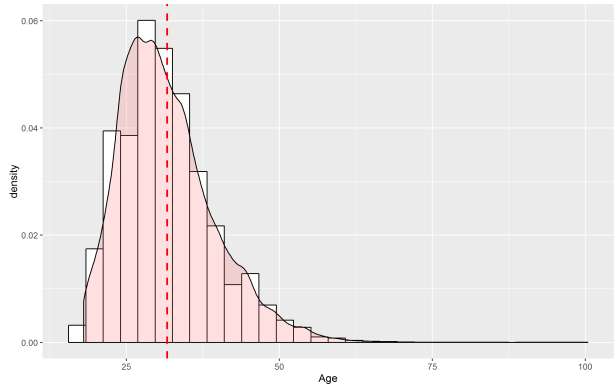

Figure 2: Density-histogram of age. The vertical line shows the mean (31.7).

and noting its level of measurement. Furthermore, some questions would have been ordinal-level if not for one item. For these ("Mixed"), it was noted which item was the odd one out, so that one can remove it and use the remaining data as an ordinal variable if desired. Table 2 shows the distribution of the variables by level of measurement.

Table 2: The distribution of levels of measurement.

Type	Count	Percent
Mixed	149	5.87 %
Nominal	319	12.56 %
Ordinal	2071	81.57 %

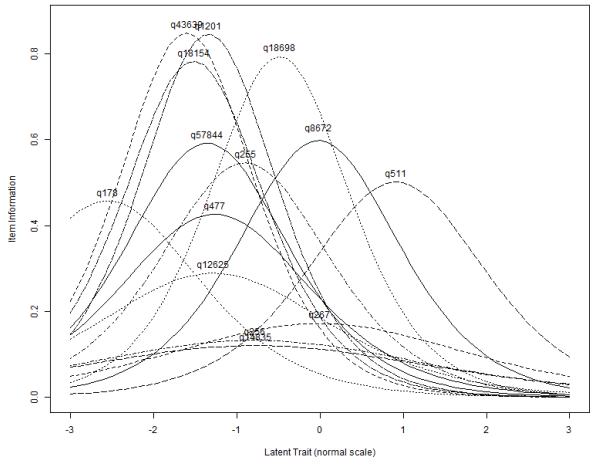

Figure 3: Item information of the 14-item test.

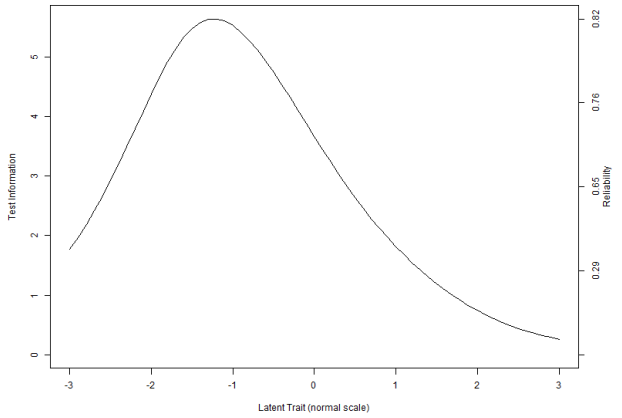

Figure 4: Test information of the 14-item test.

5 Example analyses

The following analyses are meant to be examples of the kind of research questions that one can examine with the data. It is our hope that other researchers will use the dataset for their own purposes.

5.1 *Can a measure of general cognitive ability be constructed using the questions?*

To construct a general cognitive ability test, one needs one or more variables that are loaded on the general factor of cognitive ability (Jensen, 1998). 14 suitable items were located among the variables. The items can be found in the supplementary materials (sheet *test_items*).

To verify that they loaded on a common factor, the latent correlations between the items were estimated. These are estimates of what the Pearson correlations would have been if the data had a continuous distribution instead of a discrete. These methods are also called *tetrachoric/polychoric correlations*, see Uebersax (2015). All intercorrelations between the items were positive (91/91) with a mean of .37 (range .07 to .63). The correlation matrix can be found in the supplementary materials (sheet *item_correlations*). The data were then factor analyzed using item response theory (2PN) based factor analysis (*irt.fa* in the **psych** package (Revelle, 2015)). Figures 3 and 4 show the item and test level information.

As can be seen, the test items were generally too easy; only one item (q511[6]) had a difficulty below 50 %. Thus, the test is not very good at discriminating between persons with above average (for this sample) cognitive ability.

If we use a test composed of all 14 items, the sample with complete data will be very much reduced (about .7 % of the sample had complete data for the 14 items). To examine the effect of using only a smaller number of items to increase the sample with complete data,

6 This item is: "If you flipped three pennies, what would be the odds that they all came out the same?" Answer options are: "I admit, I don't know!", "1 in 3", "1 in 4" and "1 in 8".

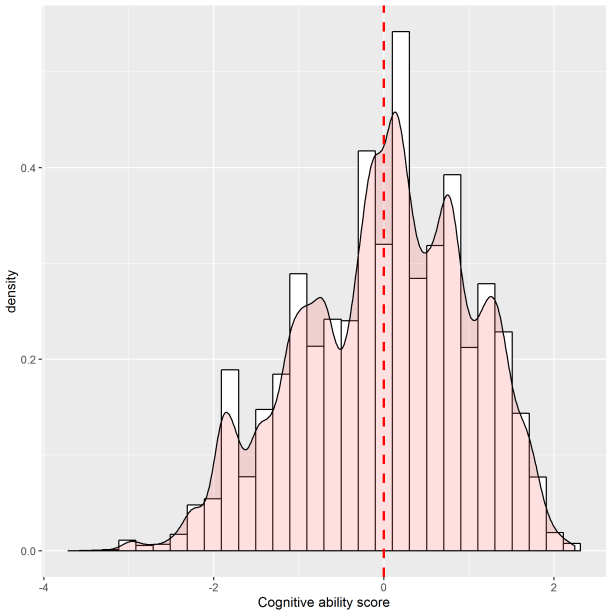

Figure 5: Cognitive ability scores derived from the 14-item test. The vertical line shows the mean.

we also created tests with 2-13 items. These were composed of the N items with the most data.[7] Scores for the tests were calculated using *score.irt* from the **psych** package. Table 3 shows the intercorrelations between the tests.

It can be seen that there is strong stability of estimates across different test compositions, with the smallest test (2 items) and largest (14 items) correlating at .68. The scores from the 14-item version were saved for further analysis. Figure 5 shows a density-histogram of cognitive ability.

The distribution of scores was approximately normal as expected, but with a long left tail (skew = -.33). This may reflect users that did not try their best to answer the questions, as has also been seen in other large datasets.

5.2 *Does cognitive ability relate to other variables in familiar ways?*

To validate the data and the measure of cognitive ability, it is useful to replicate well-known findings. To do this, we plotted the mean level of cognitive ability for each level of every variable with at most 8 levels (i.e. categorical or ordinal data) giving a total of 2551 plots. These can be found in the supplementary materials (*cognitive_ability_by_var.7z*).

[7] I.e., test 2 was composed of the two items with the most data, test 3 with the three items with the most data, etc.

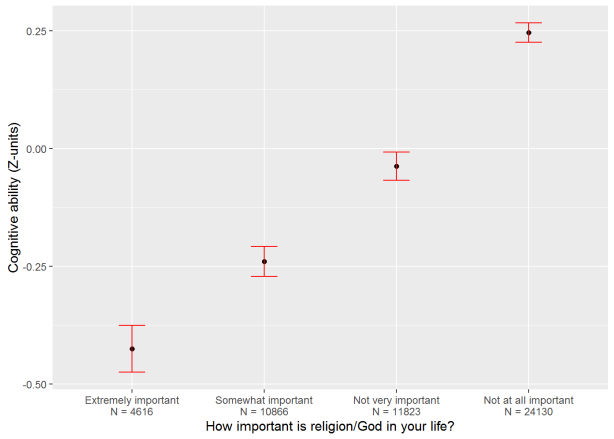

Figure 6: Mean cognitive ability by stated level importance of religion in life (q41). Error bars are 99.9 % confidence intervals.

5.2.1 *Religiosity*

There are many questions that concern religious matters and it is also possible to choose one's religion in the profile description. This is not surprising because studies find that matching on religion is very important, which is to say that assortative mating is very strong (r ≈ .70) for that trait (Hur, 2003; Watson et al., 2004).

A large body of evidence shows small to moderate negative relationships between measures of cognitive/scholastic ability and religious beliefs and practices (Dutton, 2014b; Zuckerman et al., 2013). Figure 6 shows the relationship between rated importance of religion/God in life and cognitive ability.

We see a linear negative relationship between the rated importance of religion/God in life and cognitive ability. The difference between the most and least religious groups is -.67 d (using the total sample standard deviation) and the latent correlation for the variables is -.26. Question 42 is very similar "Is your duty to religion/God the most important thing in your life?" (Yes/No), and shows a latent correlation of -.35. Question 210 is a binary question concerning the mere belief in God and shows a latent correlation of -.33. Finally, the profile page has a field where one can select one's religion from a list. Figure 7 shows the mean cognitive ability by chosen religion.

As seen in a previous study (Nyborg, 2009), we see that atheists and agnostics are in the top. Judaism has a reasonably high mean score as well. We could not find any question that concern specific subdivisions within religions, so it was not possible to test Nyborg's hypothesis that the more fundamentalist subgroups have lower cognitive ability (see also Dutton 2014a).

Table 3: Intercorrelations between tests. Tests are named after how many items they employed. N = cases with complete data for that test.

	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10	Test11	Test12	Test13	Test14
Test2	1.00	0.83	0.80	0.70	0.69	0.71	0.70	0.69	0.69	0.69	0.69	0.69	0.68
Test3	0.83	1.00	0.96	0.88	0.86	0.86	0.86	0.85	0.83	0.83	0.82	0.82	0.82
Test4	0.80	0.96	1.00	0.92	0.90	0.90	0.89	0.88	0.86	0.86	0.85	0.85	0.85
Test5	0.70	0.88	0.92	1.00	0.99	0.96	0.95	0.94	0.93	0.92	0.92	0.91	0.91
Test6	0.69	0.86	0.90	0.99	1.00	0.98	0.97	0.96	0.95	0.94	0.94	0.93	0.93
Test7	0.71	0.86	0.90	0.96	0.98	1.00	0.99	0.98	0.96	0.96	0.95	0.95	0.95
Test8	0.70	0.86	0.89	0.95	0.97	0.99	1.00	0.99	0.97	0.97	0.96	0.96	0.96
Test9	0.69	0.85	0.88	0.94	0.96	0.98	0.99	1.00	0.98	0.97	0.97	0.96	0.96
Test10	0.69	0.83	0.86	0.93	0.95	0.96	0.97	0.98	1.00	0.99	0.99	0.98	0.98
Test11	0.69	0.83	0.86	0.92	0.94	0.96	0.97	0.97	0.99	1.00	0.99	0.99	0.99
Test12	0.69	0.82	0.85	0.92	0.94	0.95	0.96	0.97	0.99	0.99	1.00	1.00	1.00
Test13	0.69	0.82	0.85	0.91	0.93	0.95	0.96	0.96	0.98	0.99	1.00	1.00	1.00
Test14	0.68	0.82	0.85	0.91	0.93	0.95	0.96	0.96	0.98	0.99	1.00	1.00	1.00
N	40085	36385	31884	29740	19958	17073	13968	7417	4761	3141	2107	1248	479

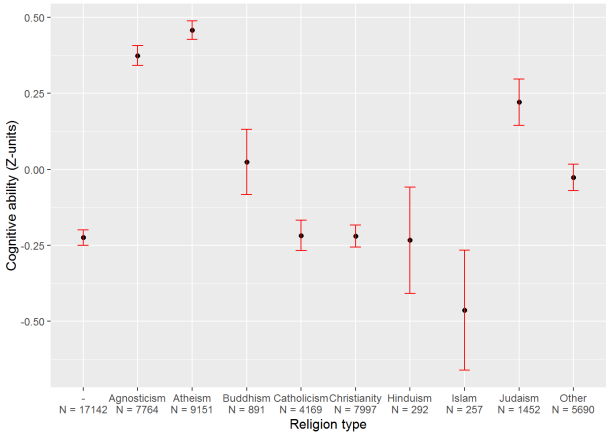

Figure 7: Mean cognitive ability by self-declared religion. Dash means no religion was declared. Error bars are 99.9 % confidence intervals.

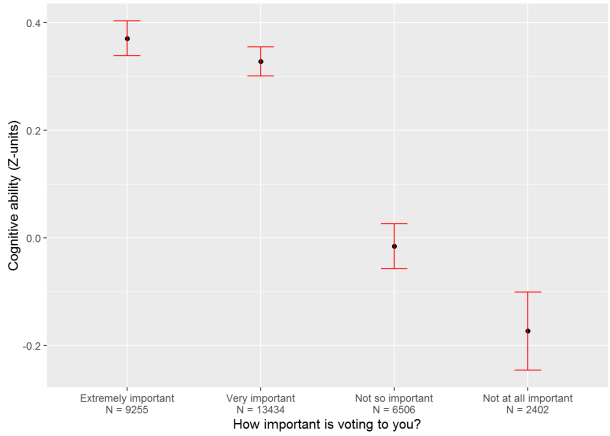

Figure 8: Mean cognitive ability by stated importance of voting (q15752). Error bars are 99.9 % confidence intervals.

5.2.2 Politics interest/participation

Many studies have reported positive associations between political interest/participation and cognitive ability. For instance, Rindermann et al. (2012) examined a Brazilian dataset and found that those who were without preference on a left-right axis had lower IQs than those who had a preference, and Deary et al. (2008) found that childhood IQ predicted adult voter turnout, voting preferences (towards center-parties), and political involvement (taking part in rallies, demonstrations and signing petitions) in the UK. Figure 8 shows the relationship between stated importance of voting and cognitive ability.

There is a monotonic relationship between belief in the importance of voting and cognitive ability, altho it may not be entirely linear (this depends on assumptions of equal intervals between the groups). The latent correlation is .19. Question 170 concerns past voting behavior in presidential (or equivalent) elections and shows a latent correlation of .29. Figure 9 shows the relationship between opinions regarding helping out a favored politician and cognitive ability.

As expected, we see that people willing to help out more had higher cognitive ability. It's not possible to calculate the latent correlation because the rank ordering of the answers is not clear: is time or money the greater sacrifice? Question 403 simply asks whether one enjoys discussing politics (yes/no) and shows a latent correlation of .32 with cognitive ability in favor of those answering yes, and question 6765 asks whether one has ever attended a demonstration or convention (yes/no), which shows a latent correlation of .27 with cognitive ability. Finally, Figure 10 shows the relationship between the opinion about the importance of one's own political beliefs and cognitive ability.

There is a moderate and fairly linear relationship between how important one's own political beliefs are to oneself and cognitive ability. The latent correlation is .31.

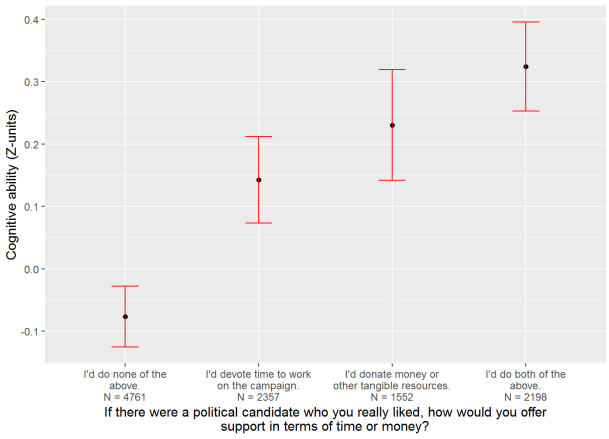

Figure 9: Mean cognitive ability by stated willingness to help a favored politician (q85974). Error bars are 99.9 % confidence intervals.

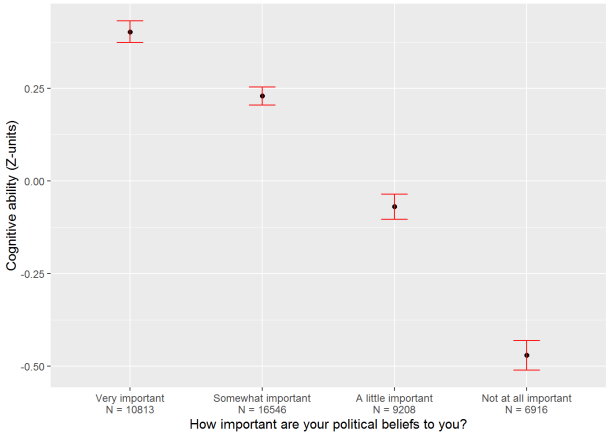

Figure 10: Mean cognitive ability by stated importance of own political beliefs (q212814). Error bars are 99.9 % confidence intervals.

5.3 *Is Zodiac sign related to answers to questions?*

Astrology is generally regarded by scientists as pseudoscientific. Hartmann et al. (2006) examined whether Zodiac sign was related to cognitive ability and multiple measures of personality in two large samples (N≈11,000 and N≈4,000). They found no noteworthy relationships.

It is possible to do a large-scale test of astrology using the OKCupid dataset by examining whether Zodiac sign is related to every question in the dataset. Zodiac sign is arguably a nominal variable and the questions are either ordinal (possibly interval-like) or nominal. Thus, to use all the questions, a test that can handle nominal x nominal variables was needed. We settled on using the standard chi square test because the goal was to look for any signal at all, not estimate effect sizes. This is a strong test because it is possible that there are effects of time of birth within a given

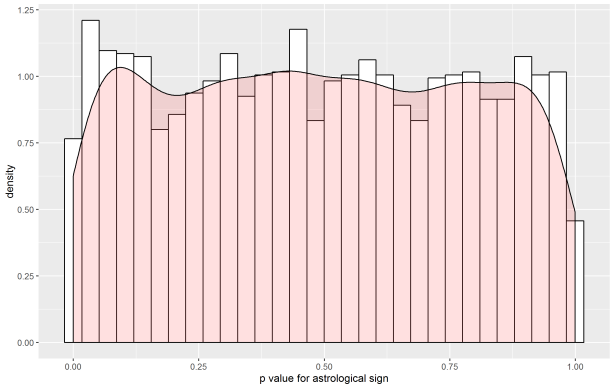

Figure 11: Distribution of p-values from chi square testing of Zodiac signs' relationship to all questions in the dataset. N=2,541.

year (e.g. spring vs. summer) which are unrelated to Zodiac sign. For instance, being born in summer may be related to which kind of activities one takes part in at age 3 due to limitations of the weather, and the experiences from these activities may have a causal impact on one's later personality (for a possible example of something of this sort, see Gobet & Chassy (2008)).

To clarify, the null hypothesis tested by the chi square test here is that the answers have the same frequency for all the 12 Zodiac populations. Figure 11 shows a density-histogram of the p-values.

The overall distribution is nearly uniform. Given the null hypothesis, this is what is expected (Simonsohn et al., 2014). The null hypothesis (given the assumptions of the test) also predicts that 5 % of p-values will be below .05. In fact, 6.7 % of the results were below .05. However, some of these were trivial. The question with the lowest p-value ($5.4e^{-24}$) was "Which is your favorite season?". In other words, the finding was that persons tend to favor the seasons that their own birthdays fall into. Overall, the results indicate that Zodiac sign is worthless as a predictor of just about anything.

6 Discussion and conclusion

Despite being based on data collected from public answers on a dating site, it was possible to construct a cognitive ability test that was seemingly well-functioning and which was related to known correlates (religiousness, political interest/participation) in expected ways (negative, positive). Zodiac sign was not generally related to the other variables except in trivial cases. Thus, this serves as a null test. From these three investigations, it seems that the dataset is useful for psychological research.

6.1 Limitations

It is worth emphasizing the limitations of the dataset. The sample is not representative of any national population, rather it is a self-elected convenience sample that consists mostly of young to middle-aged adults from the US, Canada and the UK. Furthermore, due to the way we sampled the data from the site, it is not even representative of the users on the site, because users who answered more questions are overrepresented.

The variables in the dataset were not created by psychologists, but either by the site staff or by users themselves. As such, they often contain imperfections that hinder interpretation of observed associations.

A limitation of the site is that questions can only have 2 to 4 answer options, which makes it problematic to treat the data as continuous (see e.g. http://emilkirkegaard.dk/understanding_statistics/?app=discretization). One will have to use statistical methods designed for ordinal or nominal data in most cases.

The data are exclusively self-reported and because the answers are given with a specific purpose (finding a partner) and are public, they may be incorrect. It is likely that users seeking a partner, skew their answers towards what that they think would be more acceptable to potential partners. On the other hand, users are encouraged on the site to answer honestly. This is also in their self-interest because the matching algorithm matches them with similar people and if they give incorrect answers, they do not get useful results.

The cognitive ability data is limited to about 14 items with sufficient amount of data. This necessarily limits the reliability of the measurement. Furthermore, as far as we know, these items have not been validated against known test batteries or used in any other studies.

Supplementary material and acknowledgments

The scraper code can be found at https://github.com/Deleetdk/OKCubot. Note that it is not functional at this time because of a change in the OKCupid website.

The R source code, high quality figures and full data can be found at https://mega.nz/#F!QIpXkL4Q!b3QXepE6tgyZ3zDhWbv1eg.

Thanks to a number of anonymous friends who helped gather the data by running the scraper script on their computers.

The peer review thread can be found at http://openpsych.net/forum/showthread.php?tid=279. Thanks to Davide Piffer, Gerhard Meisenberg and Bob Williams for reviewing.

References

Britt, S., Cumbie, J., & Bell, M. (2013). The influence of locus of control on student financial behavior. *College Student Journal*, *47*(1), 178–184.

Damian, R. I., & Roberts, B. W. (2015). The associations of birth order with personality and intelligence in a representative sample of us high school students. *Journal of Research in Personality*, *58*, 96–105. doi: http://doi.org/10.1016/j.jrp.2015.05.005

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

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences*, *38*, e130. doi: http://doi.org/10.1017/S0140525X14000430

Dunkel, C. S. (2014). Reassessment of jewish cognitive ability: Within group analyses based on parental fluency in hebrew or yiddish. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/05/reassessment-of-jewish-cognitive-ability-within-group-analyses-based-on-parental-fluency-in-hebrew-or-yiddish/

Dutton, E. (2014a). Nyborg's 'the intelligence-religiosity nexus' and the benefits of consilience. *Open Differential Psychology*. Retrieved from http://openpsych.net/ODP/2014/07/nyborgs-the-intelligence-religiosity-nexus-and-the-benefits-of-consilience/

Dutton, E. (2014b). *Religion and intelligence: An evolutionary analysis*. London: Ulster Institute for Social Research. Retrieved from http://www.ulsterinstitute.org/religionintelligence.html

Gobet, F., & Chassy, P. (2008). Season of birth and chess expertise. *Journal of biosocial science*, *40*(2), 313–316. doi: http://doi.org/10.1017/S0021932007002222

Hackett, R. (2016, May 18). *Researchers caused an uproar by publishing data from 70,000 okcupid*

users. Retrieved from http://fortune.com/2016/05/18/okcupid-data-research/

Hartmann, P., Reuter, M., & Nyborg, H. (2006). The relationship between date of birth and individual differences in personality and general intelligence: A large-scale study. *Personality and Individual Differences*, *40*(7), 1349–1362. doi: http://doi.org/10.1016/j.paid.2005.11.017

Hernandez, D. C., & Pressler, E. (2014). Accumulation of childhood poverty on young adult overweight or obese status: race/ethnicity and gender disparities. *Journal of epidemiology and community health*, *68*(5), 478–484. doi: http://doi.org/10.1136/jech-2013-203062

Hur, Y.-M. (2003). Assortative mating for personality traits, educational level, religious affiliation, height, weight, and body mass index in parents of a korean twin sample. *Twin Research*, *6*(6), 467–470. doi: http://doi.org/10.1375/twin.6.6.467

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

Krawczyk, M., & Reuben, E. (2012). (un) available upon request: Field experiment on researchers' willingness to share supplementary materials. *Accountability in research*, *19*(3), 175–186. doi: http://doi.org/10.1080/08989621.2012.678688

Major, J. T., Johnson, W., & Deary, I. J. (2014). Linear and nonlinear associations between general intelligence and personality in Project TALENT. *Journal of personality and social psychology*, *106*(4), 638–654. doi: http://doi.org/10.1037/a0035815

Nyborg, H. (2009). The intelligence–religiosity nexus: A representative study of white adolescent americans. *Intelligence*, *37*(1), 81–93. doi: http://doi.org/10.1016/j.intell.2008.08.003

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

Rindermann, H., Flores-Mendoza, C., & Woodley, M. A. (2012). Political orientations, intelligence and education. *Intelligence*, *40*(2), 217–225. doi: http://doi.org/10.1016/j.intell.2011.11.005

Rodgers, J. L., Van Hulle, C., D'Onofrio, B., Rathouz, P., Beasley, W., Johnson, A., ... Lahey, B. B. (2015). Behavior problems and timing of menarche: a developmental longitudinal biometrical analysis using the nlsy-children data. *Behavior genetics*, *45*(1), 51–70. doi: http://doi.org/10.1007/s10519-014-9676-4

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PloS ONE*, *4*(9), e7078. doi: http://doi.org/10.1371/journal.pone.0007078

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. doi: http://doi.org/10.1177/1745691614553988

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS ONE*, *6*(6), e21101. doi: http://doi.org/10.1371/journal.pone.0021101

Uebersax, J. S. (2015, September 8). *Introduction to the tetrachoric and polychoric correlation coefficients*. Retrieved from http://www.john-uebersax.com/stat/tetra.htm

Watson, D., Klohnen, E. C., Casillas, A., Nus Simms, E., Haig, J., & Berry, D. S. (2004). Match makers and deal breakers: Analyses of assortative mating in newlywed couples. *Journal of personality*, *72*(5), 1029–1068. doi: http://doi.org/10.1111/j.0022-3506.2004.00289.x

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS ONE*, *6*(11), e26828. doi: http://doi.org/10.1371/journal.pone.0026828

Zuckerman, M., Silberman, J., & Hall, J. A. (2013). The relation between intelligence and religiosity a meta-analysis and some proposed explanations. *Personality and Social Psychology Review*, *17*(4), 325–354. doi: http://doi.org/10.1177/1088868313497266