Submitted: 12th of February 2016
Published: 11th of July 2016

ICAR5: design and validation of a 5-item public domain cognitive ability test

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

Open Differential Psychology

Abstract

A 5-item abbreviation of the ICAR (*International Cognitive Ability Resource*) 16-item sample test was created thru exhaustive search. The 5-item version (ICAR5) was optimized for correlation with the 16- item version and for administration time based on estimated item administration times. To validate the test, it was given to students in 6th to 10th grade in two Danish schools (N=236). Age was used as a criterion variable and showed the expected positive relationship (r=.43). Results furthermore showed that the abbreviated test was too difficult for the younger students (6th and 7th grades), but not for the older students. One item was found not to be very discriminative, so it may need to be replaced in an updated version.

Keywords: ICAR, international cognitive ability resource, cognitive ability, intelligence, IQ, abbreviation, age, Danish, scale construction

1 Introduction

Currently, the most used cognitive ability tests are commercially owned. This has at least two significant downsides. First, the tests cost a small fortune to acquire which severely limits their use. For instance, WAIS-4 (*Wechsler Adult Intelligence Scale*, version 4) costs about 1200 USD for the basic kit (http://www.pearsonclinical.com/psychology/products/100000392/wechsler-adult-intelligence-scalefourth-edition-wais-iv.html#tab-pricing). With a price like that, many researchers and practitioners are not be able to afford the testing kit.

Second, because the tests are under copyright, researchers who do not have permission from the copyright holders cannot legally modify the test for their own purposes, such as translating them to another language or creating abbreviated versions.

To overcome the problems, we wanted to build and improve upon public domain (not owned by anyone) cognitive assessment tools. The ICAR (*International Cognitive Ability Resource*; https://icar-project.com/) project has a similar aim. Researchers working on that project have created a 60-item test and validated it extensively Condon & Revelle (2014) on college students. In two earlier papers, we showed that a Danish translation of the 16-item sample test (ICAR16) had good psychometric properties (Kirkegaard & Bjerrekær, 2016; Kirkegaard & Nordbjerg, 2015). However, from using the 16-item test in those projects, we found that some participants thought that taking the test took too long. This is a problem because it reduces the number of persons who are willing to spend their time participating in studies. Thus, to overcome the problem, we wanted to create a version that had satisfactory validity but takes less time to use.

2 Abbreviating the test

There are several methods that can be used to select items to use for a shorter test. One method used in a recent study Eisenbarth et al. (2015) uses a genetic (evolutionary) algorithm to search the composition space for a good selection of items. Genetic algorithms do not try every single combination of items, instead they search the composition space in a semi-directed fashion. Unfortunately, this means that it is possible that the algorithm ends up in a local maximum and thus fails to find the best combination of items. Whether this will happen or not depends on how the space 'looks like'; whether there are many smaller hills or one large mountain. Figure 1 shows an example of this.

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

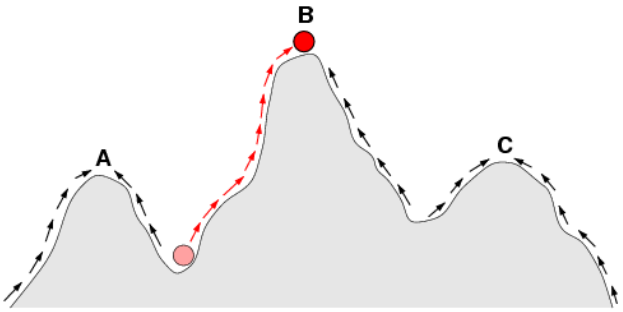

Figure 1: Simple example of a fitness landscape. The arrows show the direction the genetic algorithm would go. B is the global maximum while A and C are local maximums. Picture from Wikipedia: https://en.wikipedia.org/wiki/Fitness_landscape

The advantage of genetic algorithms is that since one doesn't have to search thru the entire space, they are computationally feasible to use when it is not feasible to try all the possibilities (which may be an infinite number). In contrast, exhaustive search tries all the possibilities and necessarily finds the best combination.

We wanted a reasonably short test and decided to create a 5-item version (hence ICAR5). There is no reason for this specific number, but we note that the 3-item *Cognitive Reflection Test* Toplak et al. (2011) achieved a sizable correlation with the ICAR16 in one of our studies (r=.51; (Kirkegaard & Nordbjerg, 2015)). However, we wanted a stronger correlation than that. Because our item pool consisted of only 16 items (found in the appendix of (Condon & Revelle, 2014)), our search space was not so large as to be computationally infeasible to search thru using exhaustive search (choose 5 of 16 without duplicates = 4368), thus we used exhaustive search.

We used the datasets (N's = 72 and 54) from the two previous studies (Kirkegaard & Bjerrekær, 2016; Kirkegaard & Nordbjerg, 2015) to calculate the correlation between the all the possible 5-item tests and the ICAR16 (the criterion correlation). Then we correlated the criterion correlations across the two datasets. This gave a correlation of only .20. We suspected that this was due to the small sample sizes (too much error in the estimates) and thus looked for a larger dataset for the ICAR16. We found this in the **psych** package for R (Revelle, 2015) which contains a dataset (*ability*) that has N=1449. Some data was missing however, but the subset of complete data is N=1248. We then calculated the criterion correlations for this dataset as well and correlated it with the other two. The correlations across datasets are shown in Table 1. In general, the intercorrelations are not strong, presumably owing to the small sample size of the two Danish datasets.[1] We used the largest dataset for all further analysis.

Table 1: Correlations between datasets' criterion correlations. DK1 = dataset from Kirkegaard & Nordbjerg (2015), DK2 = dataset from Kirkegaard & Bjerrekær (2016), and psych = dataset from **psych** package (Revelle, 2015).

	DK1	DK2	psych
DK1	1.00	0.20	0.30
DK2	0.20	1.00	0.15
psych	0.30	0.15	1.00

Figure 2 shows the distribution of criterion correlations, that is, the correlations between all the possible 5-item tests and the ICAR16.

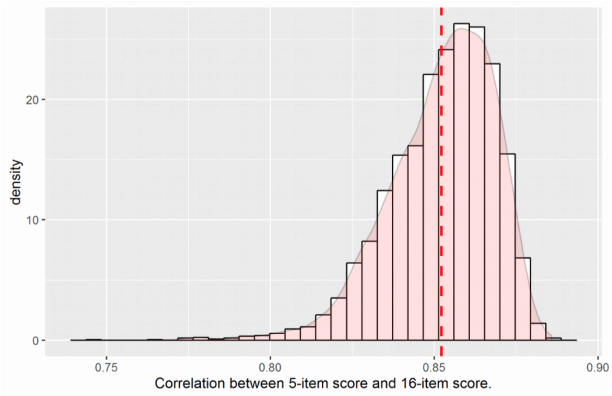

Figure 2: Distribution of correlations between 5-item versions and the 16-item version. The dashed red line shows the mean. The red curve is a density fit (for details, see the details for the *geom_density* function in the **ggplot2** package.).

The distribution is skewed with a long left tail (skew = -.96). The mean and median correlations were .85. The strongest correlation was .89.

2.1 *Item composition and correlation to 16-item version*

The ICAR16 consists of 4 verbal reasoning items (VR), 4 letter-number/alphanumeric items (LN), 4 matrix reasoning items (MR), and 4 3D-rotation dice items (R3D). It is interesting to examine the item compositions of the 5-item versions and their correlations to the criterion correlations. To do this we counted the number of items of each type in the tests. Table 2 shows the correlations. Interestingly, there were some relationships. The compositions with more alphanumeric items tended to have stronger criterion correlations and those with more matrix reasoning items weaker correlations.

[1] We did some testing to verify this by splitting up the *ability* dataset into 2 equal sized, non-overlapping subsets and then correlated the criterion correlations from them. This resulted in a very strong correlation of about .90.

Table 2: Correlations between item composition and criterion correlations. N=4368, the standard error is approximately .015.

	Criterion correlations	verbal reasoning	alpha-numeric	matrix reasoning	dice rotation
Criterion correlations		0.08	0.33	-0.30	-0.12
verbal reasoning	0.08		-0.33	-0.33	-0.33
alphanumeric	0.33	-0.33		-0.33	-0.33
matrix reasoning	-0.30	-0.33	-0.33		-0.33
dice rotation	-0.12	-0.33	-0.33	-0.33	

2.2 The 30 best compositions

We sorted the compositions for their criterion correlation. Table 3 shows the top 30 compositions.

Inspecting the items makes it clear that specific items, not just those of a given type, tend to be found in the best tests (e.g. VR4). Furthermore, one can note that almost all the tests have either two alphanumeric items or two dice rotation items. Unfortunately, these items take the longest to complete. Altho we did not measure the completion time for the items, our guess is that the verbal reasoning items take the shortest to complete, and so we wanted a combination with two of these items. The best combination to have that feature is 797 (rank 27 out of 4368, shown in green). This combination has a marginally lower criterion correlation (.8858 vs. .8800) but we judged that it was worth the trade-off. Thus, we selected this combination as our ICAR5.

3 Validation study

To make sure that the abbreviated test worked as intended, we carried out a validation study.

All analyses were done in R. R is a free language that enjoys wide and increasing use for statistical computing (among other things). The source code for the analyses can be found in the supplementary materials (*scripts* folder).

3.1 Participants

The ICAR16 has previously been tested on college students with a mean age in the early 20s (Condon & Revelle, 2014), but as far as we know, no one has tested the ICAR items on younger persons. We made a guess about the age required to solve some of the items and decided on 6th grade (about age 13 at the time of testing).

The tests were given to 236 pupils in the Danish school system from 6th to 10th grade from two different schools. For the first school (VPR), JDB asked the headmaster if the school wanted to participate in the study. For the second school (GS), JDB asked a teacher for permission to hand out the ICAR5 test sheet and instruct the pupils. According to national GPA data, one of the schools was a little above average, while the other was a little below.

Figure 3 shows the age distribution of the participants.

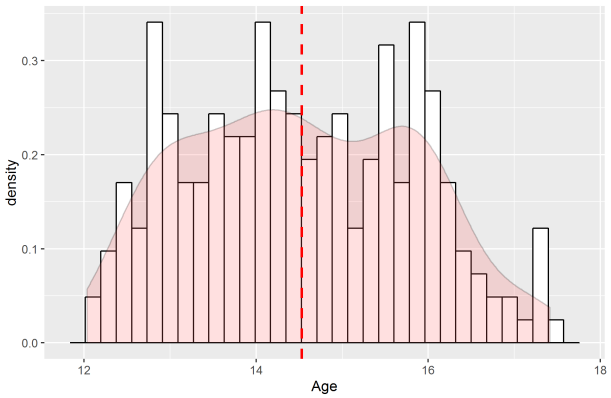

Figure 3: Age distribution of participants. The dashed line shows the mean age (14.5). The standard deviation of age was 1.31. The red curve is a density fit (for details, see the details for the *geom_density* function in the **ggplot2** package.).

Table 4 shows a breakdown of the students by grade level and school.

3.2 Administration

JDB administered the tests so the administration was standardized. However for the 7th, 8th (A and B) and 9th grade at GS, test administration was done by a teacher, who were told to give the same instructions as was given to the rest of the pupils. Said teacher was present at one of the previous test sessions as to minimize errors while administering the tests.

Table 3: The 30 best 5-item tests according to criterion correlation. The green row shows the chosen item combination. VR = verbal reasoning, LN = letter-number/alphanumeric, MR = matrix reasoning, and R3D = 3D dice rotation.

Combination	Crit. correlation	Verbal reasoning	alpha-numeric	Matrix reasoning	Dice rotation	Items
1088	0.8858	1	2	1	1	VR4 LN33 LN58 MR47 R3D6
1087	0.8851	1	2	1	1	VR4 LN33 LN58 MR47 R3D4
1172	0.8846	1	2	1	1	VR4 LN34 LN58 MR47 R3D6
2876	0.8845	1	2	1	1	VR17 LN34 LN58 MR45 R3D4
1161	0.8838	1	2	1	1	VR4 LN34 LN58 MR45 R3D6
2918	0.8833	1	1	1	2	VR17 LN34 MR45 R3D4 R3D6
968	0.8830	1	2	1	1	VR4 LN7 LN58 MR47 R3D6
1171	0.8824	1	2	1	1	VR4 LN34 LN58 MR47 R3D4
2877	0.8823	1	2	1	1	VR17 LN34 LN58 MR45 R3D6
2882	0.8819	1	2	1	1	VR17 LN34 LN58 MR46 R3D4
1166	0.8818	1	2	1	1	VR4 LN34 LN58 MR46 R3D4
2678	0.8813	1	2	1	1	VR17 LN7 LN58 MR46 R3D4
1202	0.8812	1	1	1	2	VR4 LN34 MR45 R3D4 R3D6
2875	0.8812	1	2	1	1	VR17 LN34 LN58 MR45 R3D3
1160	0.8811	1	2	1	1	VR4 LN34 LN58 MR45 R3D4
967	0.8811	1	2	1	1	VR4 LN7 LN58 MR47 R3D4
1082	0.8807	1	2	1	1	VR4 LN33 LN58 MR46 R3D4
1227	0.8806	1	1	1	2	VR4 LN34 MR47 R3D4 R3D6
2677	0.8805	1	2	1	1	VR17 LN7 LN58 MR46 R3D3
2672	0.8804	1	2	1	1	VR17 LN7 LN58 MR45 R3D4
3372	0.8804	1	2	1	1	VR19 LN34 LN58 MR45 R3D6
1263	0.8803	1	1	2	1	VR4 LN58 MR46 MR47 R3D4
2979	0.8801	1	1	2	1	VR17 LN58 MR46 MR47 R3D4
962	0.8801	1	2	1	1	VR4 LN7 LN58 MR46 R3D4
2162	0.8800	1	2	1	1	VR16 LN34 LN58 MR45 R3D6
2803	0.8800	1	2	1	1	VR17 LN33 LN58 MR47 R3D4
797	0.8800	2	1	1	1	VR4 VR19 LN58 MR46 R3D4
1193	0.8799	1	1	2	1	VR4 LN34 MR45 MR47 R3D6
929	0.8798	1	2	1	1	VR4 LN7 LN34 MR45 R3D6
2798	0.8796	1	2	1	1	VR17 LN33 LN58 MR46 R3D4

Table 4: Overview of participants by grade level and school.

Grade level	GS	VPR
6	39	19
7	33	19
8	32	19
9	32	20
10	0	16

3.2.1 The test sheet

To facilitate the testing, we made an A4 size test sheet for the students. The sheet can be found in the supplementary materials (In Danish). We asked students to give their date and year of birth but not their name since we had no need for a persistent identifier and for privacy reasons. The tests were filled out with pen and paper and were later digitized manually. We noted which exact item the students had chosen so as to make it possible to examine patterns in their incorrect choices (Section 3.3.3).

3.2.2 The administration

Before handing out the test, the pupils were given oral instructions to minimize confusion for how to proceed with the test. The instructions were as follows:

> There are 5 questions, each with multiple answers. You are to mark only one answer per question, meaning you will have a total of five answers when you are done. At the very top of the test, write down today's date and your birthday.
>
> While taking the test, please be quiet, so your fellow pupils are not disturbed. Remain quiet till everyone is done with the test. When you have marked five answers please raise your hand, so that your test can be collected.
>
> If you are in doubt, mark the answer which you feel is the most correct.
>
> As for question 3, there might be a difficult word for some of you. When encountering an 'alphanumerical' series, you are to convert the letters into numbers like:
>
> A = 1
> B = 2
> C = 3
>
> And so on. Then you are supposed to find the next logical step in this number series. An example of a number series could be:
>
> 1 2 3 4 (5)
> 2 4 6 8 (10)
>
> You may take as long as you need to finish this test.

The last part was done to ensure that the pupils had understood the third question. From a pilot run, nearly all students in a 6th grade asked for the meaning of this question, so it was assessed that this question had to be explained in more detail.

3.3 Analysis of responses

3.3.1 Descriptive analysis

Items were scored as correct or incorrect (1 or 0). Table 5 shows descriptive statistics for the items across all participants.

Table 5: Descriptive statistics for items. VR = verbal reasoning, LN = letter-number/alphanumeric, MR = matrix reasoning, and R3D = 3D dice rotation.

Item	Mean	SD	Skew	Kurtosis
VR.4	0.61	0.49	−0.44	−1.82
VR.19	0.52	0.50	−0.06	−2.01
LN.58	0.34	0.47	0.69	−1.53
MR.46	0.42	0.50	0.31	−1.91
R3D.4	0.17	0.38	1.74	1.04

Table 6 shows the correlation matrix.

Table 6: Item intercorrelations. Tetrachoric correlations below the diagonal, Pearson correlations above.[2] VR = verbal reasoning, LN = letter-number/alphanumeric, MR = matrix reasoning, and R3D = 3D dice rotation.

Item	VR.4	VR.19	LN.58	MR.46	R3D.4
VR.4		0.29	0.27	0.02	0.08
VR.19	0.45		0.26	0.14	0.21
LN.58	0.45	0.42		0.18	0.32
MR.46	0.03	0.22	0.28		0.20
R3D.4	0.15	0.39	0.54	0.36	

As expected, all correlations were positive. However, some were only barely so.

[2] Tetrachoric correlations are estimates of what the Pearson correlations would have been if the data has been continuous instead of dichotomous (0/1). They were calculated using the *tetrachoric* function in the **psych** package.

3.3.2 Item response theory analysis

We factor analyzed the data using item response theory factor analysis as implemented in the *irt.fa* function in the **psych** package (Revelle, 2015). This consists of first finding the tetrachoric/polychoric correlations between the items and then factor analyzing them using standard methods. This corresponds to a 2 parameter analysis based on the cumulative normal distribution (2PN; (Revelle, 2016, p. 251)).

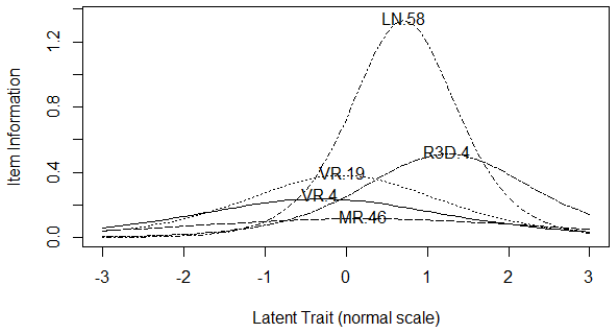

Figure 4: Item information from the ICAR5. VR = verbal reasoning, LN = letter- number/alphanumeric, MR = matrix reasoning, and R3D = 3D dice rotation.

Figure 4 shows the item information plot (as outputted from the *irt.fa* function from the **psych** package).

It can be seen that that the alphanumeric item was the best at discriminating between students and the matrix reasoning item relatively useless. This is unfortunate because this item takes a long time to complete. We note that this was the only item for which more explicit instructions were given. This could be because the students were confused as to how to solve the other items. However, the item difficulties in Table 5 show that the students were generally able to solve the other items just as well or better as the matrix reasoning item, so this does not seem to be the case.

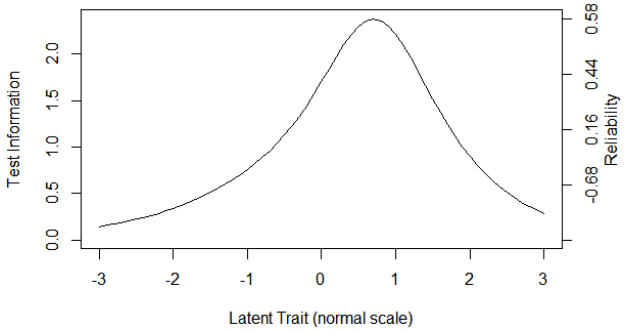

Figure 5: Test information plot for the ICAR5.

From figure 4 it can be seen that the test lacks items with good discriminative ability for persons with an ability level of about -1. This can be seen more clearly if one looks at the test information plot, shown in Figure 5.

3.3.3 Error analysis

Tables 7 thru 11 gives the counts of specific responses (in columns) by item and grade level (in rows) in percent (% omitted). Green marks the correct item.

Table 7: Responses to VR.4 (verbal reasoning item 4) by grade level. All numbers are in % row-wise.

Grade level / response	-	2	3	4	5	6	7
6	0	14	10	26	40	9	2
7	4	0	15	8	58	10	6
8	2	10	0	25	59	4	0
9	2	0	4	2	87	6	0
10	0	0	19	12	69	0	0

In general, the correct item was usually also the most chosen item, but did not necessarily receive the majority of the responses. The lower grades responses were more varied than the higher grades responses, except for the 10th grade.

3.3.4 Scoring methods

We scored the items using item response theory factor analysis, standard factor analysis[3] and using simple summed scores (all items weighed 1). Table 12 shows the correlations between the scoring methods.

As expected, the correlations were near 1. We used the simple sums for the following analysis because these would be the scores that would likely be used in practice due to their ease of calculation and the high correlations with the more sophisticated scores.

3.3.5 Relationship to age

Figure 6 shows a scatterplot of age and score. As expected, there was a strong positive effect of age on scores. 12% of participants got the lowest score and 7% the highest score, so there is some need for a lower floor and a higher ceiling.

3.3.6 Relationship to grade level/class

We calculated the mean score by class and grade level, shown in Figure 7.

[3] We used the default settings for the *fa* function in the **psych** package for R, that is, factor extraction is done using minimum residuals (least squares) and scored using the regression method.

Table 8: Responses to VR.19 (verbal reasoning item 19) by grade level. These are the names of the week in Danish. Fredag = Friday, Lørdag = Saturday, Mandag = Monday, Onsdag = Wednesday, Søndag = Sunday, and Tirsday = Tuesday. All numbers are in % row-wise.

Grade level / response	Fredag	Lørdag	Mandag	Onsdag	Søndag	Tirsday
6	3	12	36	5	3	40
7	4	2	35	8	8	44
8	2	0	57	2	2	37
9	0	2	81	2	4	12
10	0	0	50	12	0	38

Table 9: Responses to LN.58 (letter-number item 58) by grade level. All numbers are in % row-wise.

Grade level / response	-	H	I	J	L	M	N
6	2	12	16	14	5	38	14
7	2	6	17	6	2	42	25
8	6	14	8	10	0	27	35
9	2	2	2	10	0	17	67
10	0	19	19	25	0	19	19

Table 10: Responses to MR.46 (matrix reasoning item 46) by grade level. All numbers are in % row-wise.

Grade level / response	A	B	C	D	E	F
6	17	28	9	10	24	12
7	8	35	10	15	27	6
8	14	49	4	10	16	8
9	8	62	2	10	17	2
10	19	38	12	19	6	6

Table 11: Responses to R3D.4 (3D rotation item 4) by grade level. All numbers are in % row-wise.

Grade level / response	-	A	B	C	D	E	F	G	H
6	0	5	5	5	5	5	5	52	17
7	4	4	8	10	23	4	2	35	12
8	2	2	18	6	20	6	4	29	14
9	2	0	40	4	10	4	6	33	2
10	0	6	12	19	12	0	6	44	0

As expected, there is a general upwards trend commensurate with the increase seen for age. The one 10th grade class is an outlier. This is probably due to a selection effect. 10th grade is not mandatory in Denmark and the less academically able, and hence those with lower cognitive ability, students tend to take it.

Table 12: Correlations (Pearson) between scores derived using simple sums, standard FA (factoranalysis), and IRT (item response theory factor analysis).

Method	Simple sums	IRT	Standard FA
Simple sums	1	0.93	0.97
IRT	0.93	1	0.98
Standard FA	0.97	0.98	1

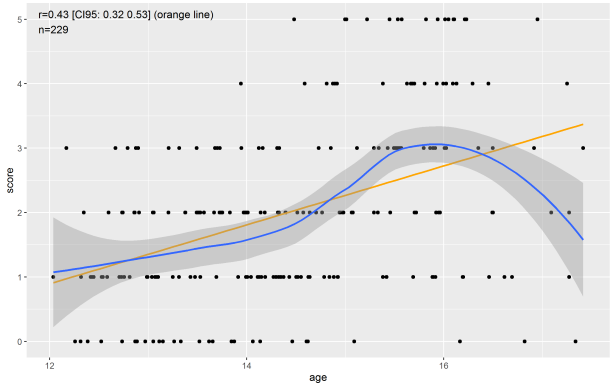

Figure 6: Scatterplot of age and score. The blue line is based on local regression.

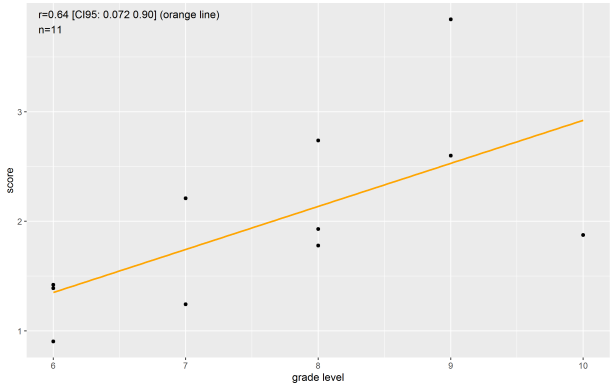

Figure 7: Mean score by grade level and class.

4 Discussion and conclusion

We were able to construct an abbreviated version of the ICAR test that shows the expected relationship

to age. However, analysis showed that the test is too difficult for students below approximately 8th grade (Danish standards).

Analysis showed that the matrix item did not perform as expected. Given the observed floor effect, one may want to swap it with an easier item, perhaps another matrix item so that the maximal diversity of items can be retained.

We did not have other criteria variables than age and grade level to validate the test against. Future studies should use a broader collection of criterion variables such as grade point average and parental educational attainment.

We did not measure the time each item takes to complete. This was not possible with our research design, but can be done somewhat easily with computerized testing. Instead we used our informed opinion to guess the administration times for each item. It took roughly 10-15 minutes for the pupils to complete the ICAR5.

Supplementary material and acknowledgements

Data, high quality figures and R code can be found in the supplementary materials available at https://osf.io/yqe6p/. Thanks to Davide Piffer, Nick Mendieta and Bob Williams for reviewing.

References

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64. doi: http://doi.org/10.1016/j.intell.2014.01.004

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the psychopathic personality inventory–revised (ppi-r). *Psychological Assessment*, *27*, 194–202. doi: http://doi.org/10.1037/pas0000032

Kirkegaard, E. O. W., & Bjerrekær, J. D. (2016). Country of origin and use of social benefits: A pilot study of stereotype accuracy in denmark. *Open Differential Psychology*. (Retrieved from http://openpsych.net/ODP/2016/04/country-of-origin-and-use-of-social-benefits-a-pilot-study-of-stereotype-accuracy-in-denmark/)

Kirkegaard, E. O. W., & 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/)

Revelle, W. (2015). *Procedures for psychological, psychometric, and personality research (version 1.5.4)*. (Retrieved 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/)

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.