Submitted: 19th of March 2015
Published: 31st of July 2015

Validating a Danish translation of the International Cognitive Ability Resource sample test and Cognitive Reflection Test in a student sample

Emil O. W. Kirkegaard* Oliver Nordbjerg†

Open Differential Psychology

Abstract

We translated the International Cognitive Ability Resource sample test (ICAR16) and the Cognitive Reflection Test (CRT) into Danish. We administered the test online to a student sample (N=72, mean age 17.4). Factor analysis revealed a general factor. The summed score of all test items correlated .42 with GPA. Item difficulties correlated .85 with those reported in the Internet norming sample. Method of correlated vectors analysis showed positive relationships between g-loading of items/subtests and their correlation with GPA (r=.53/.85). Model comparisons revealed that for predicting GPA the CRT did not have incremental validity over the ICAR16, but the evidence was not strong.

Keywords: ICAR, intelligence, general cognitive ability, grade point average, g-factor, educational achievement, method of correlated vectors, Danish

1 Introduction

A large fraction of cognitive and personality tests are privately owned and are usually very expensive to obtain legally (e.g. the Wechsler test is owned by Pearson). There is no good reason for this; it is a barrier to psychological science and the practical use of testing. Given the growing open science movement (see e.g. [1]), it is not surprising that some have been working on replacing proprietary tests with public domain tests with equal validity. One such test is the International Cognitive Ability Resource developed by Condon and Revelle.[2] A sample test consisting of 16 items is presently available for research.

Another recently invented test is the Cognitive Reflection Test (see e.g. [3]). Although not invented in the spirit of open science, the test consists of a mere 3 items which can be found in many places on the internet, so it is presumably not copyright-protected.

Since we want to contribute to the on-going development of free psychology tools and have a Danish language test to use for future projects, the aim of this study was to psychometrically validate the Danish translation of the two tests using a student sample.

2 Data source and sample description

We put together a survey using *Google Forms* consisting of questions about the participants':

- gender (M/F)
- age
- type of gymnasie[1]
- year they started in gymnasiet
- grade they are in
- total grade point average (GPA)
- GPA in science classes (math and psych included)
- GPA in language classes

* Department of Culture and Society, University of Aarhus. Corresponding author, E-mail: emil@emilkirkegaard.dk
† E-mail: nordbjerg.oliver@gmail.com

[1] Gymnasie is a secondary education taken by approximately two thirds of a generation. In years of education, it is grade 10-13 or 9-12 depending on whether the student took the optional 10th grade or not. In US terms, it is similar but not identical to high school. The gymnasie is meant as a preparation for further education, so it is somewhat selected for academic ability and hence general cognitive ability.

- GPA in physical education
- the 3 items from the CRT
- the 16 items from the ICAR16

One of us (ON) posted this on a student intranet visited by hundreds of students from different schools. He posted it twice with about a week's delay and obtained a total of 72 responses.

Mean age was 17.4 (SD=1.25), sample was 76% female. 92% of the students attended STX, with the rest attending various other kinds (HTX 1.4%, HF, 5.6%, HHX 1.4%).[2] All students began their studies between 2012 and 2014 roughly evenly spread and they were also about equally spread throughout grades as expected. Repeating grades in uncommon in Danish gymnasier.

Note that since testing was not controlled, the test settings were probably not identical for each test taker, which introduces error into the correlations.

After collecting data we noticed that we had forgotten to translate the text on the four rotation item images. It may have made these items a bit harder for the students, but not that much. Two recent surveys found that Danes have the best command of English of any non-native English group.[4, 5]

3 Scoring items

Since no fact-sheet regarding the correct answers was provided, each author took the test independently and then compared their answers. There were no cases of disagreement and so we coded our answers as correct. We used simple functions in *Google Calc* to score each item as either 1 (correct) or 0 (incorrect). Mean scores and SD for each item is shown in Table 1.

Table 2 shows the test means and SDs.

4 Internal structure

All statistical analyses were done in R.[3]

To examine the internal structure, we used both traditional factor analysis (FA) and item response theory factor analysis (IRT FA) on all the cognitive items to extract 1 factor.

[2] In Denmark there are 4 main types of gymnasier. STX is the standard type, HTX is technology-oriented, HHX is trade-oriented. Lastly, HF is a shorter 2-year program which gives a roughly equivalent degree.

[3] R is a free, powerful, easy to use programming language designed for data mining and statistics. See http://www.r-project.org/

Table 1: Mean and SD for each item.

Item	Item mean	Item SD
CRT1	0.36	0.48
CRT2	0.42	0.50
CRT3	0.33	0.47
VR.04	0.69	0.46
VR.16	0.64	0.48
VR.17	0.83	0.38
VR.19	0.72	0.45
LN.07	0.53	0.50
LN.33	0.49	0.50
LN.34	0.39	0.49
LN.58	0.32	0.47
MR.45	0.42	0.50
MR.46	0.49	0.50
MR.47	0.57	0.50
MR.55	0.31	0.46
R3D.03	0.25	0.44
R3D.04	0.31	0.46
R3D.06	0.38	0.49
R3D.08	0.22	0.42

Table 2: Overall means and SDs by test.

Items	Mean	SD
ICAR16+CRT	8.65	4.52
ICAR16	7.54	3.83
CRT	1.11	1.15

Although popular, principal components analysis has been shown to give misleading results in some cases.[6] For this reason, we used another extraction method which by default is MinRes (minimum residuals), but it does not appear to make a large difference which method is used.[6] The functions *fa()* and *irt.fa()* from the **psych** package were used for extraction.[7]

The difference between FA and IRT FA is that the latter is done on a correlation matrix calculated using tetrachoric correlations. A tetrachoric correlation estimates the Pearson correlation between two normally distributed continuous latent variables that are assumed to underlie dichotomous variables such as correct/incorrect items.

Factor loadings are shown in Table 3. The factor congruence across extraction methods was 1.00.

Next we compared the item means and SDs (shown in Table 1) with those published by Condon and Revelle.[2] The correlations were .85 and .63 indicating high construct reliability across languages and samples.

Table 3: Item loadings. Variance explained in FA = 23 %, in IRT FA = 38%.

Item	FA loadings	IRT FA loadings
CRT1	0.51	0.69
CRT2	0.40	0.51
CRT3	0.62	0.76
VR.04	0.34	0.51
VR.16	0.56	0.74
VR.17	0.41	0.62
VR.19	0.36	0.52
LN.07	0.58	0.74
LN.33	0.56	0.70
LN.34	0.61	0.76
LN.58	0.24	0.35
MR.45	0.32	0.41
MR.46	0.27	0.36
MR.47	0.46	0.63
MR.55	0.48	0.61
R3D.03	0.59	0.76
R3D.04	0.51	0.64
R3D.06	0.48	0.61
R3D.08	0.46	0.59

The internal reliability metrics are shown in Table 4.

Since Condon and Revelle's paper only reported multi-factor loadings, we contacted them to obtain loadings for a one-factor solution which they provided for their sample 2 (these can be found in the supplementary material). Using these, we calculated the factor congruence which was .977, showing excellent factor structure agreement across samples and languages. Note that this analysis was done using only the ICAR16 items and used the IRT FA method.

5 Cognitive scores and GPA

We investigated three different ways of summing the scores:

1) unweighted sum

2) factor scores from FA

3) factor scores from IRT FA

We also calculated the unweighted sums of ICAR16 and CRT alone. We then correlated these with each other and with total GPA, results shown in Table 5.

Weighted sums showed very little incremental ability above the unweighted sum. Including the 3 CRT items along with the 16 ICAR16 items showed little incremental validity over ICAR16 (ICAR16 alone .405, ICAR16+CRT .416).

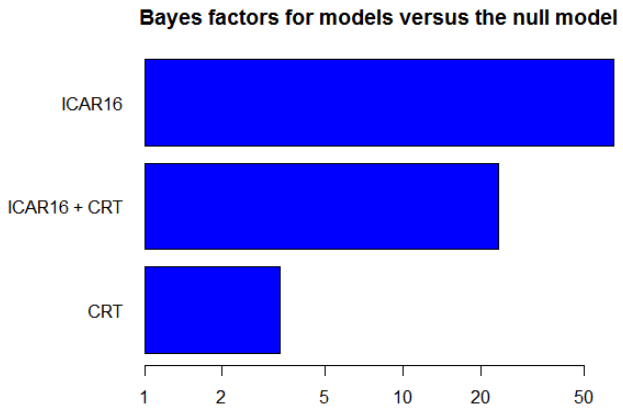

Figure 1: Bayesian Factors versus the null model.

6 Incremental validity of the CRT

6.1 OLS regression

To examine possible incremental validity of CRT over ICAR, we regressed GPA on ICAR and CRT with ordinary least squares (OLS) regression. For this purpose we used the unweighted sums of both tests. Results are in Table 6.

Both predictors had positive betas; however, the beta for CRT was only .11 and the confidence interval included 0. Due to the small sample, this may either be because it has no incremental validity or because power was too low to detect it.

Another method is to compare the adjusted R^2 value of the possible models. These are shown in Table 7.

The adjusted R^2 value takes into account the number of predictors used in a model. In the table, we see that the model without CRT was better than the one with, but that the difference is fairly small.

6.2 Bayesian linear regression

As an alternative, we used Bayesian linear regression to examine the same modeling question as above. We used the **BayesFactor** package, see [8]. The results are shown in Figure 1.

The results were similar to before in that the ICAR16 model was better than the others. The CRT was again redundant. One can calculate a measure of how strong the evidence is of one model over another by calculating the Bayes' factor (the probability of the

Table 4: Internal reliability measures of tests.

Test	Conbach's alpha	Guttman's lambda	Mean interitem r
ICAR16+CRT	.84	.88	.21
ICAR16	.81	.85	.21
CRT	.69	.62	.43

Table 5: Correlations of scores with total GPA.

Variable	IRT.FA.scores	unwt.sum.scores	CRT	ICAR16	GPA
FA.scores	0.998	0.988	0.700	0.958	0.417
IRT.FA.scores		0.996	0.700	0.967	0.417
unwt.sum.scores			0.685	0.976	0.416
CRT				0.509	0.286
ICAR16					0.405

Table 6: Dependent variable: GPA. Parameter estimates from OLS regression.

Predictor	Std. beta	CI95%
ICAR16	.35	0.10 to 0.60
CRT	.11	-0.15 to 0.36

Table 7: Model comparison using OLS regression.

Model	R2 adjusted
ICAR16	0.152
CRT	0.082
ICAR16+CRT	0.149

data given one model dividing by the probability of the data given the other). There are multiple rules of thumb about interpreting the value, but values below 3 or 5 are said to be "barely worth mentioning".[9] The value for our study is 2.78.

7 Jensen's method

In order to examine whether the g-factor is driving the relationship between cognitive scores and GPA, we calculated the correlation between the g-loading of each item and each item's correlation with GPA (method of correlated vectors).[10] If g is driving the relationship, this correlation should be moderately to strongly positive.[11] Results are shown in Figure 2.

We also ran the analysis with the loadings from regular factor analysis, and results were similar (r=.49).

Since item scores have quite a bit of noise, especially with a small sample, we ran the MCV at the subtest level by grouping the items after content type. Results are shown in Figure 3.

Generally, the results support the idea that g is driving relationship between cognitive scores and GPA.

8 Discussion and conclusion

One reviewer criticized the study for not regressing out the effect of age. One could do this, but given the small variation (SD = 1.25) of age in the sample, the expected effect size of age was minute. However, we did calculate the correlation of the IRT FA scores with age, which was .11. The partial correlation of the IRT FA scores with GPA controlling for age was virtually identical at .404 vs. 417 without partialing out age.

The 16 item ICAR test shows good internal reliability, construct validity and criteria validity when translated to Danish and used on a student sample. Future studies should compare its predictive validity to standard proprietary tests. Judging from our results and the results reported by Condon and Revelle, the test appears to be ready for practical use.

Three methods of analyzing incremental validity of the CRT above the ICAR16 failed to find evidence of this and two of them produced evidence to the contrary, but due to the small sample, the evidence is not very strong.

Supplementary material and data

The source code, figures, data and the Danish versions of the CRT and ICAR16 can be found at the

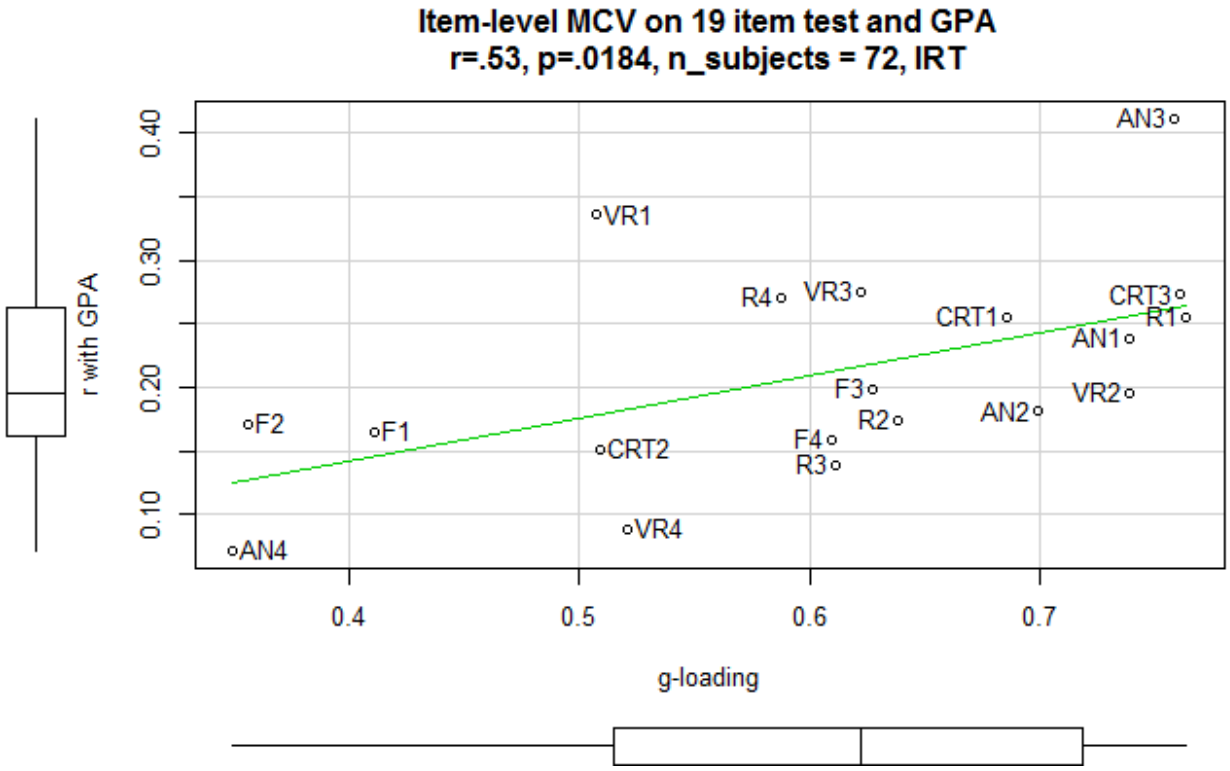

Figure 2: Item-level method of correlated vectors with item response theory factor loadings and GPA.

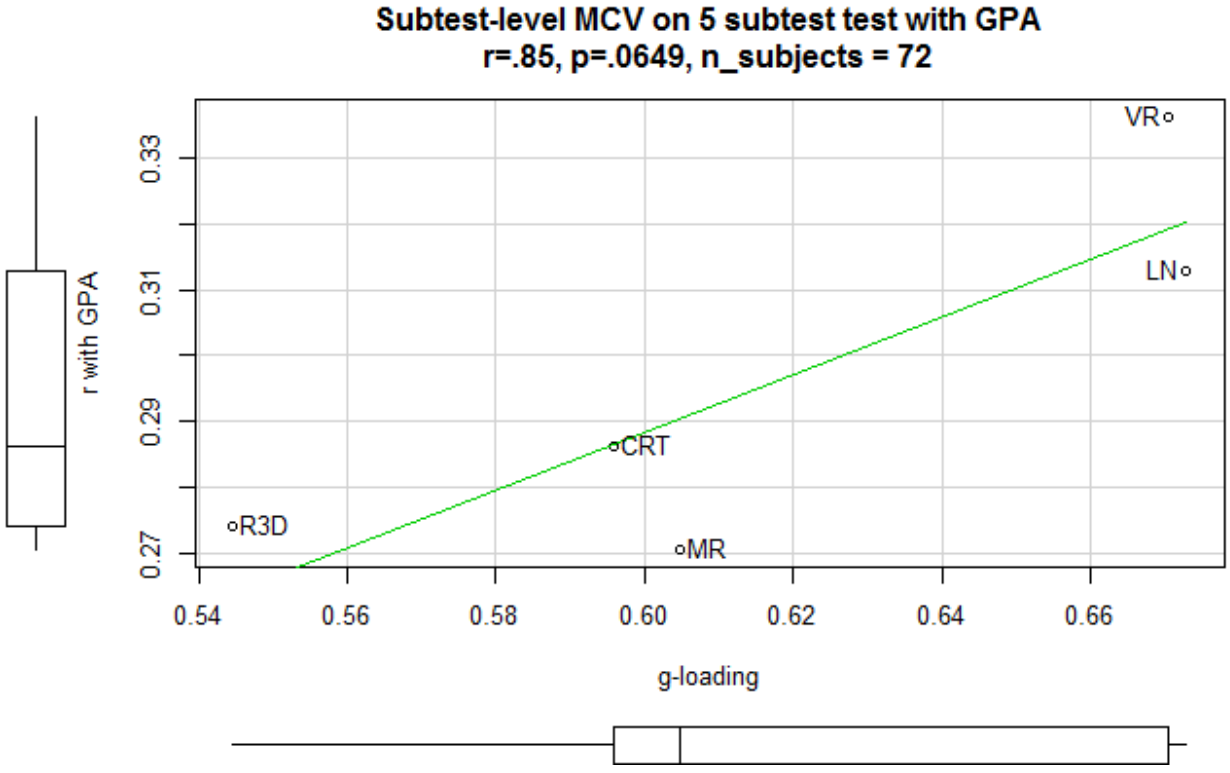

Figure 3: Subtest-level method of correlated vectors with GPA.

Open Science Framework repository. The peer-review discussion can be found on the journal's forum.

Author contributions

Both designed the questionnaire and both proof-read the paper before submission. E conceived of the study, did the data analysis and wrote the paper. O posted the questionnaire on the intranet.

References

[1] Bartling S, Friesike S. Opening Science: The Evolving Guide on How the Web is Changing Research, Collaboration and Scholarly Publishing. GitHub.com; 2014.

[2] Condon DM, Revelle W. The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. Intelligence. 2014;43:52–64.

[3] Toplak ME, West RF, Stanovich KE. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & Cognition. 2011;39(7):1275–1289.

[4] First E. EF English Proficiency Index 2014; 2014. Available from: http://www.ef-danmark.dk/epi/downloads/.

[5] your vocab T. Results by country; 2011. Available from: http://testyourvocab.com/blog/2011-09-10-Results-by-country.

[6] Kirkegaard EOW. The international general socioeconomic factor: Factor analyzing international rankings. Open Differential Psychology. 2014;.

[7] William Revelle. Overview of the psych package; 2015. Available from: http://cran.r-project.org/web/packages/psych/index.html.

[8] Etz A. Using Bayes Factors to Get the Most out of Linear Regression: A Practical Guide Using R. Winnower. 2015;.

[9] Bayes factor; 2015. Page Version ID: 663791788. Available from: https://en.wikipedia.org/w/index.php?title=Bayes_factor&oldid=663791788.

[10] Jensen AR. The g factor: the science of mental ability. Westport, Conn.: Praeger; 1998.

[11] Frisby CL, Beaujean AA. Testing Spearman's hypotheses using a bi-factor model with WAIS-IV/WMS-IV standardization data. Intelligence. 2015 Jul;51:79–97. Available from: http://www.sciencedirect.com/science/article/pii/S0160289615000549.