Statistical Methods for ICAR Automated Reports

All analyses are based on a data set consisting of observations of 616,270 participants. Due to the study design, only a very small proportion of the existing items appears in each test leading to about 98% missing values. After a brief summary of the participants, the used methods to handle the mass of NA's are described. [1]

1 The Participants

34.5% of the participants are male, 64.21% female and 0.18% identify as another sex.

The geographical distribution of the participants is shown in the following table. The majority is from America or Asia, mainly from the USA (226044), India (75039) and the Philippines (35432).

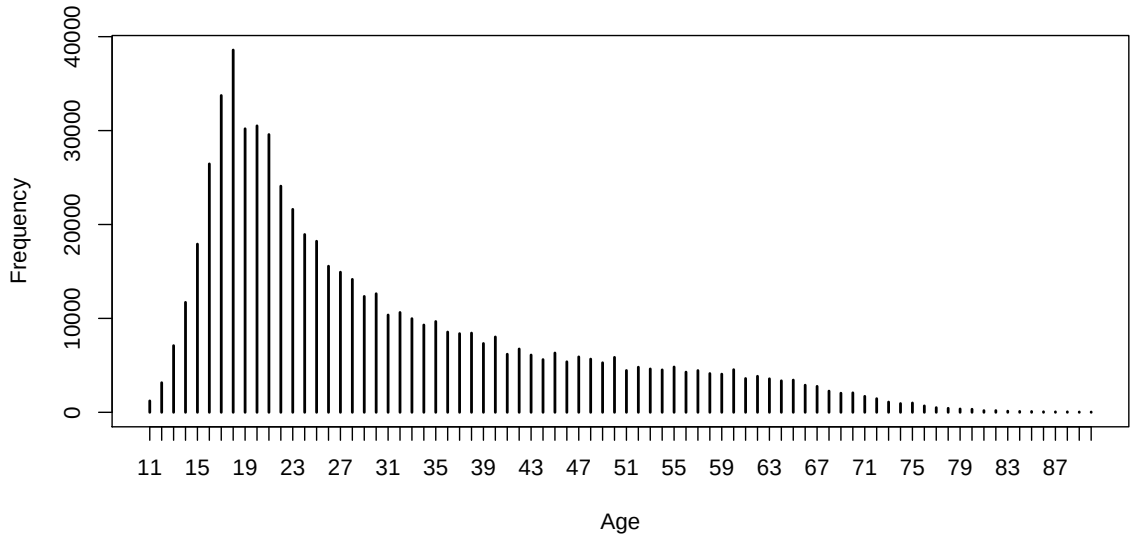
Africa	America	Asia	Europe	Oceania
54,631	262,661	185,412	45,579	12,254

The next table indicates the education of the participants. Most are currently in college/university, currently in graduate or professional school or have a graduate or professional school degree.

The distribution of the participant's age in years is shown in the following barplot. With a range from 11 to 90, every age group is present in the data set. However, primarily teenager and young adults between 15 and 25 years attended the test. The arithmetic mean is about 31 and the median 25 years.

[1]Contact: Prof. Dr. Philipp Doebler, doebler@statistik.tu-dortmund.de

Less than 12 years	45,126
High school graduate	61,029
Currently in college/university	108,003
Some college/university, but did not graduate	39,107
Associate degree (2 yr)	20,517
College/university degree (4 yr)	119,743
Currently in graduate or professional school	20,303
Graduate or professional school degree	101,823



2 Statistical Methods

Firstly, the item types are analyzed descriptively by counting how often each item has been answered (N), the arithmetic means of the items which is equivalent to the percentage of correct answered questions and the standard deviations.

Next, a correlation matrix is calculated for each item type based on the pairwise tetrachoric correlations of all items belonging to this type. As mentioned above, there are a lot of missing values which often leads to the problem of too little pairwise observations. In this case, some correlations cannot be obtained which complicates further analyses. One way to handle that is imputation.

So in the first step, the tetrachoric correlation matrix is calculated using a modified variant of the `tetrachoric()` function in the R package `psysch` (Revelle, 2018), which allows NA's in the resulting correlation matrix in the cells with too little pairwise observations.

If the tetrachoric correlation is incomplete, it is imputed by replacing a missing value

by the arithmetic mean of the row. The effective sample size is set to the mean number of pairwise observations. Note that the imputation is only done when the median of the number of pairwise observations is at least 100.

Now, the item types can be analyzed based on the obtained tetrachoric correlations. We calculated Cronbach's alpha with `psych::alpha` as an estimate for the reliability and applied the Spearman Brown formula. Additionally, `psych::irt.fa` is used for an item response analysis by exploratory factor analysis of the tetrachoric correlations.

The four item types Letter and Number Series, Verbal Seasoning, Matrix Reasoning and 3D-Rotation were also applied pairwise to confirmatory 2-factor analyses using `cfa()` of the R package `lavaan` (Rosseel, 2012). The tetrachoric correlation matrix of those paired item types were also imputed the way described above.

References

- Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, Version 1.8.12, `https://CRAN.R-project.org/package=psych`,
- Rosseel, Yves (2012): lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, **48**(2), 1-36, `http://www.jstatsoft.org/v48/i02/`