



HHS Public Access

Author manuscript

Assessment. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Assessment. 2020 January ; 27(1): 136–148. doi:10.1177/1073191117744048.

Validating Online Measures of Cognitive Ability in Genes for Good, a Genetic Study of Health and Behavior

MengZhen Liu,

Department of Psychology, University of Minnesota

Gianna Rea-Sandin,

Department of Psychology, Arizona State University

Johanna Forster,

University of Trondelag

Lars Fritsche,

University of Trondelag

Katherine Brieger,

Department of Biostatistics, University of Michigan

Christopher Clark,

Department of Biostatistics, University of Michigan

Kevin Li,

Department of Biostatistics, University of Michigan

Anita Pandit,

Department of Biostatistics, University of Michigan

Gregory Zajac,

Department of Biostatistics, University of Michigan

Gonçalo R. Abecasis,

Department of Biostatistics, University of Michigan

Scott Vrieze

Department of Psychology, University of Minnesota

Abstract

Genetic association studies routinely require many thousands of participants to achieve sufficient power, yet accumulation of large well-assessed samples is costly. We describe here an effort to efficiently measure cognitive ability and personality in an online genetic study, Genes for Good. We report on the first 21,550 participants with relevant phenotypic data, 7,458 of whom have been genotyped genome-wide. Measures of crystallized and fluid intelligence reflected a two-dimensional latent ability space, with items demonstrating adequate item-level characteristics. The Big 5 Inventory questionnaire revealed the expected five factor model of personality. Cognitive

Corresponding author: Scott Vrieze, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455. vrie0006@umn.edu.

measures predicted educational attainment over and above personality characteristics, as expected. We found that a genome-wide polygenic score of educational attainment predicted educational level, accounting for 4%, 4%, and 2.7% of the variance in educational attainment, verbal reasoning, and spatial reasoning, respectively. In summary, the online cognitive measures in Genes for Good appear to perform adequately and demonstrate expected associations with personality, education, and an education-based polygenic score. Results indicate that online cognitive assessment is one avenue to accumulate large samples of individuals for genetic research of cognitive ability.

Keywords

IQ; personality; education; online assessment; polygenic score; genetic association

It has become clear that tens to hundreds of thousands of individuals are necessary to discover genetic variants associated with complex behavioral traits such as mental illness, cognitive ability, and personality (Okbay, Baselmans, et al., 2016; Ripke et al., 2014; Vrieze et al., 2014). Achieving such sample sizes is routinely done through genome-wide association study (GWAS) meta-analyses, where data is combined across many studies with existing genotypes and phenotypes. While GWAS meta-analysis has been successful in arriving at such samples, it is necessarily limited to those phenotypes that have been assessed in a sufficient number of sufficiently large studies. Time- and resource-intensive measures, such as in-person psychiatric interviews or detailed cognitive assessments, are difficult or impossible to obtain in enough individuals to support large genetic association meta-analytic studies. This fact has led to genetic association meta-analyses of proxy phenotypes. Educational attainment, measured as years of education, is a measure widely available in most psychological and medical studies, and may be a proxy for cognitive ability (Rietveld et al., 2014). A recent large genetic association meta-analysis discovered 74 genetic loci associated with educational attainment. The question remains, however, which biological, psychological, and sociological mechanisms are affected by these loci to exert influence on years of education. To understand this, new assessments on tens or hundreds of thousands of individuals will be necessary to examine novel hypotheses (Iacono, Malone, & Vrieze, 2016).

Building traditional studies of this size (e.g., with mail-in surveys or telephone/in-person interviews) requires costly initiatives or existing research infrastructure, such as a national health system or national registries available, for example, in some Scandinavian countries. In the United States no unified infrastructure exists, prompting efforts like the Precision Medicine Initiative to take on the major task of linking across health systems and implementing internet-based questionnaires. Even assuming that internet-based questionnaires provide valid measures of psychological function, anything more than token honoraria in a study of 100,000 individuals is infeasible, necessitating other incentives for participation.

In the present study, we evaluated the quality of psychological measurement in Genes for Good (<https://apps.facebook.com/genesforgood>), an online study that incentivizes

participation with an appeal to scientific altruism and participant receipt of personalized genetic and phenotypic information. Participants complete surveys about their health and behavior and, after answering a sufficient number of questionnaires, become eligible to provide a DNA sample and receive their own genome-wide genotypes.

Online measurement like that used in Genes for Good has some potential advantages over in-person or mail-in approaches. For example, online assessment allows for test administration at any time of day, in any location, and ideally over any type of internet-enabled electronic device without the social pressure or embarrassment that might be present in face-to-face testing (Condon & Revelle, 2014; Haworth et al., 2007). Multiple studies have found that few differences are apparent when comparing online to paper-and-pencil versions of several common questionnaires including standard personality questionnaires (Chuah, Drasgow, & Roberts, 2006; Haworth et al., 2007; Ployhart, Weekley, Holtz, & Kemp, 2003). For example, (Chuah et al., 2006) compared results from personality assessments that were completed either in a proctored lab setting using paper-and-pencil questionnaires (N=266) or a computer based questionnaire (N=222), or were assigned to an unproctored online computer-based questionnaire (N=240). Little difference was found in results between the lab and internet setting. A follow-up assessment was conducted two weeks later, in either the same or opposite testing condition. Test scores for all groups increased after the second assessment, indicating an expected practice effect but not one associated with the testing condition.

A similar online study of cognitive ability (Haworth et al., 2007) found strong psychometric properties of standard verbal and performance IQ in 2,500 10-year-old twin pairs (as well as a follow-up two years later). The median Cronbach's α was .89, suggesting high internal consistency, and the correlations between Internet-based and in-person testing was about .80, suggesting high validity of the online tests. However, twins completed the online cognitive tests only after being telephoned by a proctor who ensured the twins understood test instructions and talked them through the assessment while on the phone. Thus, although subjects participated online, an in-person component remained. To our knowledge, the only purely online intelligence tests have been created and validated by the International Cognitive Assessment Resource (ICAR; icar-project.com). Tests of crystallized (vocabulary, general knowledge) and fluid (a version of matrix reasoning) intelligence have been created by ICAR and the items are freely available for research. Test construction and scale validation appear acceptable and have been reported previously (Condon & Revelle, 2014, 2016).

Here, we extended this previous work by considering the psychometric properties of online personality, cognitive, and demographic measures in a large-scale genetically informative study. Data were collected during 2015–2017 in Genes for Good on the Big Five Inventory (BFI) personality questionnaire (John, Robins, & Pervin, 2008), the ICAR measures of crystallized and fluid intelligence, self-reported education, and genome-wide genotyping. We examined the psychometric properties of, and relationships between, these measures in the Genes for Good sample. Finally, we tested the extent to which genetic polymorphisms previously associated with educational attainment (Okbay, Beauchamp, et al., 2016) are associated with educational attainment and IQ in Genes for Good. Polygenic scores

generated from all variants analyzed in the SSGAC GWAS results accounted for 3.2% of variation in educational attainment (Okbay, Beauchamp, et al., 2016). Multiple other groups have used similar approaches from an earlier educational attainment GWAS (Rietveld et al., 2013) to successfully predict general cognitive ability from education-associated genetic variants (Rietveld et al., 2014).

We hypothesized, consistent with prior research, that relationships between online psychological assessments would mimic those relationships found for in-person assessments. In particular, we evaluated the following. First, consistent with a large literature, we tested whether our online data would recover a five factor personality structure and two factor cognitive ability structure reflecting crystallized and fluid intelligence. Second, also consistent with prior work, we evaluated the extent to which measures of crystallized intelligence, fluid intelligence, and personality indicators would independently predict educational attainment (Ackerman & Heggestad, 1997; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007; Soubelet & Salthouse, 2011; Sternberg, Grigorenko, & Bundy, 2001). Finally, we hypothesized that a polygenic score based on educational attainment would predict educational attainment in Genes for Good, and predict our online measures of verbal and matrix reasoning, again consistent with previous work using in-person assessments (Okbay, Beauchamp, et al., 2016; Rietveld et al., 2014).

Methods

Participants

Participants were drawn from the Genes for Good study, an online genetic study of health and behavior. Genes for Good is hosted on University of Michigan servers and is presented through a Facebook App platform to provide reliable participant authentication and a natural means for study outreach (<https://apps.facebook.com/genesforgood>). Facebook is not otherwise involved in the study and cannot view any data provided by the participants, or any data presented to the participants by the App. Facebook does know when a participant uses the App. Anyone over the age of 18 with a mailing address in the United States can theoretically participate. Participants were initially recruited through word of mouth from January to March 2015. In April 2015 recruitment increased after a press release from the University of Michigan, although no active advertising has been employed to increase recruitment rates. Participants for the present study included all individuals who consented for participation prior to March 28, 2017. To incentivize participation, participants who complete enough questionnaires are sent a saliva collection kit through the mail. Returned saliva samples are processed and genotyped. Participants are provided ancestry analysis results through the app, with a genome-wide genotype file available as a download. No other incentive or payment is provided. At the time of this writing, one must complete 15 of 28 available questionnaires about health history, which includes the cognitive tests, and 20 independent completions of eight different daily questionnaires. Daily questionnaires can only be answered once each day, so even the most motivated participant cannot qualify for genotyping until three days after they consent. Descriptive statistics including demographics for all participants recruited during this time are included in Table 1. Ethical approvals to conduct the study were granted by the University of Michigan IRB.

Materials

Demographic information was obtained through a modified version of the PhenX Demographics survey, a copy of which is available at genesforgood.org. Personality was measured with the 44-item Big Five Inventory (BFI) (John et al., 2008). Fluid and crystallized intelligence were measured with the 16-item Verbal Reasoning and 30-item Matrix Reasoning scales from the International Cognitive Ability Resource (Condon & Revelle, 2014, 2016). The Verbal Reasoning test includes items that test logic, vocabulary and general knowledge questions. The Matrix Reasoning test is similar to Raven's Progressive Matrices, a non-verbal test of fluid intelligence (Raven, 2000). The Matrix Reasoning stimuli are 3×3 arrays of geometric shapes with one of the nine shapes missing. Participants are given 8 possible solutions to the incomplete pattern. The Verbal Reasoning and Matrix Reasoning tests have each been validated in prior online research, demonstrating acceptable psychometric properties. Participants were at no point advised that their responses were timed, and completion times were generally not considered in scoring responses, with the exception of egregiously fast or slow responses indicating the possibility of non-credible responding.

As of this writing, 7,458 individuals have also been genotyped in Genes for Good using an Illumina Human Core Exome array. This array contains ~250,000 common tag single nucleotide polymorphisms (SNPs) and >250,000 additional genetic variants that are predominantly low frequency. DNA was collected with mail-in saliva collection kits. DNA extraction and genotyping was done at the University of Michigan Sequencing Core. Genotypes underwent standard quality control (variants were removed with minor allele frequency < 0.01, missing call rate > 0.1, and hardy-weinberg equilibrium exact test $p < 1e-3$; individuals were removed if missingness per individual > 0.1; failed sex check between genetic sex and self-reported sex), were phased using SHAPEIT (Delaneau, Marchini, & Zagury, 2012), and imputed to the 1000 Genomes phase 3 (The 1000 Genomes Project Consortium, 2015) whole genome sequence reference panel using Minimac3 (Das et al., 2016).

Data Analysis

Genes for Good is an online study that allows anyone in the US over the age of 18 to participate. Participation is of course voluntary, and it is unclear how study and questionnaire demand characteristics may influence who participates more or less fully in the study. We evaluated whether individuals who fully participate (to the point of being genotyped) differ from those who do not, using simple t-test and chi square test comparisons of genotyped and non-genotyped individuals. Comparisons were made for age, sex, race, income, educational level, personality, and cognitive ability. To evaluate potential limits in generalizability, participant demographics were also compared to US Census data on sex, age, education, income, and race. Census estimates from 2015 were taken from the American FactFinder website, managed by the US Census Bureau. Education level was assessed in Genes for Good by asking about educational milestones, similar to the system used by the US Census. These included the possible responses of "No high school diploma or GED", "High school graduate or GED", "Some college but no degree", "Associate degree", "Bachelor degree", and "Master's degree or higher". Responses were numerically

recoded with 1="No high school diploma or GED" up to 6="Master's degree or higher". Household income was assessed with possible responses of "Less than \$35,000", "\$35,000 to \$50,000", "\$50,000 to \$75,000", "\$75,000 to \$100,000", and "More than \$100,000". These responses were also numerically recoded in a similar way as for education.

Given the online, unsupervised nature of the study, we took several steps to evaluate data integrity. In the Verbal Reasoning and Matrix Reasoning tests, our primary measure of credible responding was response speed. For both tests, we calculated response speed for each participant response to each item. We then removed each response that was made in 5 seconds (or, for two easy and short Verbal Reasoning items, 4 seconds), under the assumption that 5 seconds is too fast to attend to item content, much less make a credible response. For Verbal Reasoning this removed 1,833 responses out of 119,269 total responses (1.5%); in Matrix Reasoning the procedure removed 5,878 responses out of 96,422 total responses (6.1%).

To investigate internal validity of the intelligence questionnaires, we calculated internal consistency statistics and conducted item response theory (IRT) analyses using the mirt library (Chalmers, 2012) in the R Environment. IRT is conceptually similar to factor analysis, in that it statistically models item response patterns as a function of a smaller number of latent factors (Reise, Ainsworth, & Haviland, 2005). General cognitive ability is an example of a latent factor hypothesized to influence participant responses on our Verbal and Matrix Reasoning tests. In IRT, the parameters of a link function are estimated, which relate level of general cognitive ability to the probability that the individual will correctly answer a given item. In general, the higher the cognitive ability, the higher the probability that a participant will answer an item correctly. Participant responses to the Verbal Reasoning and Matrix Reasoning tests were coded as binary items (correct or incorrect) and modeled with two-, three-, and four-parameter logistic models (parameters corresponding to difficulty, discrimination, guessing, and the upper bound parameter). Accurately estimating guessing and upper bound parameters requires large samples. To mitigate this issue, we placed priors on the guessing and upper bound parameters to help ensure convergence. Priors were normally distributed with means corresponding to the number of foils for each item and relatively tight standard deviations. Modifying the standard deviation of the priors did not change the model selection or interpretation outcome. If there were 8 response options, as in the Matrix Reasoning test, then the guessing prior would be $1/8^{\text{th}}$. Each model was fit to the verbal reasoning items and the matrix reasoning items, with the best model selected by the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) and Bayes Factor (Kass & Raftery, 1995). Model parameters were estimated using full information maximum likelihood and the default EM algorithm optimizer with Ramsay acceleration in the mirt package.

To verify the expected two-factor structure of the cognitive tests, we also fit to all Verbal and Matrix Reasoning items two competing models. The first model posited that all items loaded onto a single factor. The second model posited that all verbal items loaded onto a verbal factor, and the matrix items loaded onto a nonverbal factor (representing crystallized and fluid intelligence, respectively). The two factors were allowed to correlate. The two models were compared for fit using the DIC and a Bayes Factor.

To verify the expected five-factor structure of the BFI in the present sample, we conducted an exploratory factor analysis of the age- and sex-corrected BFI items. The residualized items were modeled as continuous variables and an oblimin rotation using the `fa()` function in the `psych` library of the R Environment. Scree plots and parallel analysis were conducted using the `fa.parallel()` function.

With the IRT and factor models fitted, we evaluated the external validity of the online personality and IQ measures. First, we performed a multiple regression of educational level regressed on age, sex, verbal reasoning, matrix reasoning, and sum scores representing the Big Five factors as scored according to the BFI manual (John et al., 2008). For verbal and matrix reasoning variables we used factor scores extracted from the independent verbal and matrix reasoning IRT models described above using the expected a posteriori (EAP) method (Embretson & Reise, 2013).

Finally, we evaluated the extent to which known genetic associations with educational attainment could be replicated in the Genes for Good sample, and whether those associations extended to verbal and matrix reasoning rather than education only. At the time of this writing, the Social Science Genetic Association Consortium (SSGAC) had performed the largest genome-wide association study (GWAS) to date with a discovery sample of 293,723 individuals and millions of genetic variants (Okbay, Beauchamp, et al., 2016). From these SSGAC results we computed a polygenic predictor of education, which is a sum of education-increasing alleles across many genetic variants tested in SSGAC. We considered multiple partially overlapping sets of variants to generate polygenic scores. The first set contained only the most strongly education-associated variants, where the association with educational attainment in Okbay et al. was significant at $p < 5e-8$ (number of variants = 3,264). Other sets were defined by relaxing the p-value threshold to include additional variants that were less significantly associated in the SSGAC, but may be expected to be truly associated with educational attainment and for which the direction and magnitude of association is approximately correctly estimated. In addition, common variants near one another in the human genome are often in linkage disequilibrium; that is, highly correlated and partially redundant proxies for one another. To account for this redundancy, we removed variants using the pruning and thresholding method as implemented in LDpred (Vilhjálmsdóttir et al., 2015). This method removed variants that were in linkage disequilibrium ($r^2 > 0.1$) with a more significant variant within a window containing 2600 variants (the recommended value in LDpred) around the index variant. All polygenic scores were calculated on a high-quality subset of variants with MAF $> .01$ and imputation quality score > 0.9 . (The imputation quality score can be interpreted as the squared correlation between the imputed number of alternative alleles and the true number of alternative alleles for a given variant.)

To create the polygenic score, we downloaded the genome-wide results from the SSGAC's educational attainment at (http://ssgac.org/documents/EduYears_Main.txt.gz) GWAS. After linkage disequilibrium pruning with LDpred, there were 3,264 variants included in the polygenic score with $p < 5e-8$ in the SSGAC results; 6,122 at $p < 5e-7$, 10,667 at $p < 5e-6$; 20,009 at $p < 5e-5$; 42,299 at $p < 5e-4$; 115,176 at $p < .005$; 391,083 at $p < .05$, 1,789,970 at $p < .5$; and 3,020,257 at $p=1.0$ (i.e., all variants). We evaluated the extent to which each of

these polygenic scores predicted educational attainment in Genes for Good, as well as our measures of verbal and matrix reasoning. Polygenic scores were then tested for association with educational attainment, verbal reasoning, and matrix reasoning as outcomes in linear models that also included age, age squared, sex, and the Big 5 personality scale scores.

To account for possible population stratification, we conducted all genetic analyses only in individuals of European ancestry. To identify individuals of European ancestry, a four dimensional hypercube was defined based on the first four genetic principal components from individuals of known European ancestry in the 1000 Genomes Project. Genes for Good participants were projected onto this space and those lying within the boundaries of the hypercube were taken forward for genetic analysis. To evaluate familial relatedness in the sample, we calculated a kinship matrix between all possible pairs of individuals in the Genes for Good sample. Approximately 15% of the genotyped sample was related at the level of a cousin, or of closer relation. Familial relatedness affects standard errors of effect sizes and corresponding p-values, not the point estimate of the effect size itself. We calculated polygenic risk score associations without accounting for relatedness, as the extent of relatedness observed here is not expected to have any substantive effect on our results or conclusions.

Results

Descriptive statistics for the sample are displayed in Table 1. The number of individuals responding to any given test varies widely, from 7,200 for Matrix Reasoning to 20,301 for Education and Household Income, the latter being part of the same demographics survey. The range of responses is possible in a study like Genes for Good, as participants are free to complete at their leisure whichever surveys they choose; no surveys are required.

While it is difficult to make direct comparisons between Genes for Good and the US census, due simply to the use of different demographic measurement instruments, the Genes for Good sample did deviate from the general US population on all key demographics. Genes for Good participants are younger and more likely to be female, with 76% in Genes for Good versus 51% in the US population. The median age of all US citizens was 37.6 in 2015; in Genes for Good the median age of participants, who incidentally are required to be 18 or older to join the study, was 35. In Genes for Good, 80% of participants self-identify as White whereas the census finds only 74% of US population self-identify as White. Genes for Good participants are wealthier, with 63% of the sample reporting household incomes of \$75,000 or greater versus the US median of \$53,889. The participants are also more well-educated. For example, in the US, 11.2% of the population age 25 or older has a graduate or professional degree and 13.3% are high school graduates or equivalent. In Genes for Good, 23% of those 25 or older have a graduate or professional degree and 9% have a high school diploma or equivalent.

Upon further comparison of the genotyped subsample to the remainder of participants who were not yet genotyped, individuals genotyped were more likely to be male ($\chi^2=144.1$, $df=1$, $p<2e^{-16}$), more well-educated ($t=8.4$, $df=16,060$, $p<2e^{-16}$), and have higher scores on the Verbal ($t=15.8$, $df=13,163$, $p<2e^{-16}$) and Matrix Reasoning tests ($t=10.7$, $df=7,198$, $p<2e^{-16}$). Genotyped participants were also more likely than non-genotyped participants to be

more conscientious ($t=6.0$, $df=15,788$, $p=1.5e^{-9}$), less extraverted ($t=-7.5$, $df=15,418$, $p=5.5e^{-14}$), less agreeable ($t=-4.5$, $df=15,639$, $p=4.1e^{-7}$), less neurotic ($t=-5.1$, $df=15,535$, $p=4.1e^{-7}$), and similar on openness ($t=.21$, $df=15,460$, $p=.84$). Genotyped participants were not older or younger on average ($t=-0.86$, $df=15,616$, $p=.39$), were no more or less likely to be to be White ($\chi^2=1.6$, $df=1$, $p=.21$), and had similar incomes compared to non-genotyped participants ($t=.74$, $df=15901$, $p=.46$).

Table 2 contains descriptive information for the Verbal Reasoning and Matrix Reasoning tests. Cronbach's α was .85 for both Verbal and Matrix Reasoning. The Matrix Reasoning test contained two items (MR 21 and MR 25) negatively correlated with a sumscore calculated from the remaining items ($-.12$ and $-.02$ respectively), as shown in Table 2. These items were excluded from all IRT models as they imply that as ability on the latent trait increases (e.g., as cognitive ability increases) then the probability of getting the item correct decreases. According to the DIC and Bayes Factor, the 2-parameter model fit the Verbal Reasoning data best, and the 4-parameter model fit the Matrix Reasoning data best (see Table 3). Item parameters for the Verbal and Matrix models are reported in Table 4. Generally, the majority of Verbal items were not overly difficult, with 6/16 items having difficulties less than two standard deviations below the mean. Matrix Reasoning items, on the other hand, show substantial coverage of a wide swath of the ability spectrum except for the highest abilities two or more standard deviations above the mean. Guessing parameters for both Verbal and Matrix Reasoning were close to expectation given that Verbal Reasoning had seven responses per item and Matrix Reasoning had eight. Some verbal items may have stronger response foils than others, as 10/16 had guessing parameter estimates $<.10$ (indicating that $<10\%$ of guesses for these items are correct). Guessing and upper bound parameters for all Matrix Reasoning items were highly similar. Factor scores calculated from the individual Verbal and Matrix Reasoning models were correlated .64.

When all Verbal and Matrix Reasoning items were fit under the same model, the 2-factor, correlated factor model had far superior fit to the single factor model (1-factor DIC = 244,676.3 versus 2-factor DIC = 246,178.4). The verbal and matrix factors in the 2-factor model were correlated at .81.

Full BFI response sets for participants were excluded if the within-person variance was above 3.3 or below .3, or the within-person mean was above 4.4 or below 2.5. These thresholds were chosen through simple visual inspection of distribution outliers (these distributions are displayed in Figure 1). We also calculated the difference between 16 item pairs with similar content, one of which was intentionally reverse keyed by the BFI authors. After reflecting all reverse-keyed items we summed the differences between all 16 item pairs, such that large values represent potentially non-credible responding. We removed full response sets from participants with sums over 40, representing extreme outliers from visual inspection. These three filters together resulted in the removal of 336 (1.7%) of participants who responded to the BFI questionnaire. Figure 1 shows the distributions of these three metrics, as well as the thresholds used to remove individuals.

Next, we removed all participant responses with response speeds ≤ 1 second or ≥ 5 minutes. This resulted in the exclusion of between 94 and 232 responses per item. Increasing the filter

to exclude all responses with response speeds ≥ 2 seconds increased the number of removed responses to range from 324–1990 responses per item. The wide range of missingness with a ≥ 2 second threshold suggests that fast-responding participants may have been attending to the content of the items, but some items required less thought and time to complete than others. We therefore chose the more relaxed threshold to exclude items with response times ≥ 1 second.

Most participants who started the BFI finished it. After removal of apparent non-credible responding, including too-fast responding, 19,926 participants responded to the first item of the BFI, 19,500 responded to item 10, and 19,241 participants (95%) who started the questionnaire completed the final item. Cronbach's α computed according to the BFI authors' scoring system for the openness, conscientiousness, extraversion, agreeableness, and neuroticism scales were .78, .81, .87, .78, and .87, respectively.

Prior to conducting exploratory factor analysis we linearly regressed out the effects of sex and age. The Scree test and parallel analysis on the residualized BFI data suggested 5 and 9 factors, respectively. See Figure 2. The exploratory factor loading matrix of the BFI is provided in Table 5, and is consistent with the expected five-factor solution. Additional factors fit to the data were not readily interpretable.

Next, we regressed educational level onto age, sex, Verbal Reasoning factor scores, Matrix Reasoning factor scores, and the five Big 5 sumscores computed from the BFI according to the BFI scoring instructions. Significant independent predictors of educational level included age (standardized $\beta=0.70$, $p<2e^{-16}$), age squared ($\beta=-0.60$, $p<2e^{-16}$), Verbal Reasoning ($\beta=0.33$, $p<2e^{-16}$), Matrix Reasoning ($\beta=0.16$, $p<2e^{-16}$), conscientiousness ($\beta=0.12$, $p=2e^{-16}$), openness ($\beta=.03$, $p=.009$), and extraversion ($\beta=.03$, $p=.02$). On the other hand, sex ($\beta=0.02$, $p=.17$), agreeableness ($\beta=-0.02$, $p=.06$), and neuroticism ($\beta=-0.03$, $p=.15$) did not provide incremental prediction of educational attainment. The full model, containing all predictors, accounted for 23% of the variance in educational level ($F=175.8$, $df=[10, 5,886]$, $p<2.2e^{-16}$).

The polygenic score generated from the SSGAC educational GWAS significantly predicted educational level, verbal reasoning ability, and matrix reasoning ability, for all p-value thresholds and all phenotypes, over and above covariates including sex, age, age squared, and personality. Considering only the polygenic score calculated on all variants regardless of p-value threshold, the polygenic score accounted for 4.1%, 4.3%, and 2.6% of the variance in educational attainment, Verbal Reasoning, and Matrix Reasoning, respectively. Figure 3 displays the standardized effect size of the polygenic score from the multiple regression with age, age squared, sex and the BFI five factors. Multiple standardized effects are shown for multiple polygenic scores, each calculated on a set of genetic variants as grouped by the strength of their association with educational attainment in the SSGAC GWAS results. All effects were statistically significant, with p-values ranging from $2.9e^{-6}$ to $8.1e^{-62}$, depending on the outcome and the genetic variant p-value threshold used to construct the polygenic score. Polygenic scores were also incrementally predictive of educational attainment over and above Verbal Reasoning, and Matrix Reasoning over and above all other predictors including sex, age, age squared, personality, educational attainment, as well as Verbal

Reasoning, and Matrix Reasoning, as applicable. Incremental r^2 for the polygenic score based on all genetic variants, for example, was 0.9%, 0.6%, and 0.2% with corresponding p -values of $1.5e^{-8}$, $8.1e^{-8}$, and .004.

Discussion

While online research may only work for a small set of data types and research designs—experiments that require highly controlled environments or specialized equipment is currently difficult to conduct in an online setting—it represents a promising avenue for efficient observational research. The personality and IQ measures used in Genes for Good demonstrated adequate psychometric properties, delivering expected factor structures and internally consistent scale scores, with low rates of obviously non-credible responding. The measures also behaved as expected. A five factor model of personality was recovered. The verbal and matrix reasoning measures suggested a two-factor latent space, and both incrementally predicted educational attainment. Openness, conscientiousness, extraversion, and neuroticism were all associated with educational attainment over and above cognitive ability. Finally, the polygenic risk score accounted for a significant portion of the variance in educational attainment.

The present study provides preliminary evidence on the potential of studies like Genes for Good to refine our understanding of genetic associations reported in the literature. The polygenic score produced by the SSGAC predicted not only educational attainment, but also our measures of verbal and matrix reasoning. The result replicates previous studies finding that educational attainment is genetically correlated with cognitive ability, but also that this result can be achieved using cognitive measures from a purely online study like Genes for Good.

The present study collected data on 21,550 individuals and genetic data on 7,458. While this sample size was large enough to evaluate the psychometric properties of questionnaires, it will not be enough for many behavioral genetic questions. Insufficient sample size is a major limitation in many behavioral research fields (Button et al., 2013; Open Science Collaboration, 2015), including behavioral genetics (Iacono et al., 2016). In Genes for Good, we have attempted to overcome this problem through creation of a framework by which genetic research can be conducted on a large scale. Participation in Genes for Good is entirely online and is incentivized through a sense of scientific altruism, citizen science, and return of phenotypic and genetic information. It appears that such incentives are sufficient for the collection of the behavioral data and genetic data described herein, and by extrapolation may be sufficient for collection of many types of health and psychological information. Indeed, since conducting the analyses that comprise this article, the total sample size in Genes for Good has increased to 54,489, 23,439 in the queue for genotyping, and 9,647 genotyped.

While online data collection may be viable, clearly there are limitations to the approach. One significant limitation is that the Genes for Good sample is not representative of the general population. They are younger, more educated, wealthier, and much more likely to be female. Generalizability of this sample to other samples will be limited by these factors.

Clearly, however, the pattern of results obtained in the present analyses are as one would expect from in-person assessments. This suggests either that the broad pattern of relationships among the constructs under study are not highly sensitive to some demographic variables, or perhaps that previous studies (e.g., of college undergraduates) of these constructs were biased in the similar ways as the present study. In addition to non-representativeness of the Genes for Good sample, the subsample of individuals who participate enough to be genotyped were different on key demographics than those who did not participate enough to be genotyped. They are less likely to be male, are more likely to be younger, are more well-educated, score higher on cognitive tests, and have different personalities. A restriction of range on educational attainment among the genotyped participants may attenuate associations between the polygenic score and education in the present sample, but it is unclear how else selection bias of this kind may affect relationships within and among the questionnaires and tests described herein. It remains to be seen if online testing protocols can successfully administer specialized tests of cognitive abilities, which we expect to be useful in understanding at a psychological level the mechanisms by which genetic variants associated with educational attainment or general cognitive ability may exert their influence. In one step toward this goal, in Genes for Good we have implemented a test of cognitive switching and working memory, two tests of executive function (Miyake & Friedman, 2012), as interactive tasks. Data continues to accumulate in Genes for Good, and participants are now completing these tests of executive function, as well as other surveys that have been implemented after the initial launch of the study in 2015. Continued experimentation with online assessment will help ensure that cognitive abilities—which are linked to many important health and social outcomes (Sternberg et al., 2001)—will not be omitted from large, and increasingly online, human health and behavioral research.

Acknowledgments

The authors are grateful for grant support from the University of Michigan Genomics Initiative as well as the National Institutes of Health awards R01AA023974, R01DA037904, U01DA041120, and R21DA040177.

References

- Ackerman PL, Heggestad ED. 1997; Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*. 121(2):219–245. DOI: 10.1037/0033-2909.121.2.219 [PubMed: 9100487]
- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013; Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*. 14(5):365–376. DOI: 10.1038/nrn3475 [PubMed: 23571845]
- Chalmers RP. 2012; mirt: A Multidimensional Item Response Theory Package for the *R* Environment. *Journal of Statistical Software*. 48(6)doi: 10.18637/jss.v048.i06
- Chuah SC, Drasgow F, Roberts BW. 2006; Personality assessment: Does the medium matter? No. *Journal of Research in Personality*. 40(4):359–376.
- Condon DM, Revelle W. 2014; The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*. 43:52–64.
- Condon DM, Revelle W. 2016; Selected ICAR Data from the SAPA-Project: Development and Initial Validation of a Public-Domain Measure. *Journal of Open Psychology Data*. 4(1):1.doi: 10.5334/jopd.25

- Das, S; Forer, L; Schönherr, S; Sidore, C; Locke, AE; Kwong, A; ... Fuchsberger, C. Next-generation genotype imputation service and methods. *Nature Genetics*. 2016. advance online publication
- Delaneau O, Marchini J, Zagury JF. 2012; A linear complexity phasing method for thousands of genomes. *Nature Methods*. 9(2):179–181. DOI: 10.1038/nmeth.1785
- Embretson, SE, Reise, SP. Item response theory. Psychology Press; 2013. Retrieved from <https://books.google.com/books?hl=en&lr=&id=9Xm0AAAAQBAJ&oi=fnd&pg=PR1&dq=embretson+reise+IRT&ots=Ea5UZuMR0m&sig=qrKdDm5EtdDnfzpwLtKDItnOKaY>
- Haworth CMA, Harlaar N, Kovas Y, Davis OSP, Oliver BR, Hayiou-Thomas ME, ... Plomin R. 2007; Internet cognitive testing of large samples needed in genetic research. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*. 10(4):554–563. DOI: 10.1375/twin.10.4.554 [PubMed: 17708696]
- Iacono, WG; Malone, SM; Vrieze, SI. Endophenotype best practices. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*. 2016.
- John, OP, Robins, RW, Pervin, LA. *Handbook of Personality, Third Edition: Theory and Research*. Guilford Press; 2008.
- Kass RE, Raftery AE. 1995; Bayes factors. *Journal of the American Statistical Association*. 90(430): 773–795.
- Miyake A, Friedman NP. 2012; The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current Directions in Psychological Science*. 21(1):8–14. DOI: 10.1177/0963721411429458 [PubMed: 22773897]
- Okbay, A; Baselmans, BM; De Neve, J-E; Turley, P; Nivard, MG; ... Fontana, MA; , et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*. 2016. Retrieved from <http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3552.html>
- Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, ... Benjamin DJ. 2016; Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*. 533(7604):539–542. DOI: 10.1038/nature17671 [PubMed: 27225129]
- Open Science Collaboration. 2015; Estimating the reproducibility of psychological science. *Science*. 349(6251):aac4716–aac4716. DOI: 10.1126/science.aac4716 [PubMed: 26315443]
- Ployhart RE, Weekley JA, Holtz BC, Kemp C. 2003; Web-Based and Paper-and-Pencil Testing of Applicants in a Proctored Setting: Are Personality, Biodata, and Situational Judgment Tests Comparable? *Personnel Psychology*. 56(3):733–752. DOI: 10.1111/j.1744-6570.2003.tb00757.x
- Raven J. 2000; The Raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*. 41(1):1–48. [PubMed: 10945921]
- Reise SP, Ainsworth AT, Haviland MG. 2005; Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*. 14(2):95–101.
- Rietveld CA, Esko T, Davies G, Pers TH, Turley P, Benyamin B, ... Koellinger PD. 2014; Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*. 111(38):13790–13794. DOI: 10.1073/pnas.1404623111
- Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, ... Koellinger PD. 2013; GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science*. 340(6139):1467–1471. DOI: 10.1126/science.1235488 [PubMed: 23722424]
- Ripke S, Neale BM, Corvin A, Walters JT, Farh KH, ... Holmans PA, et al. 2014; Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 511(7510):421. [PubMed: 25056061]
- Roberts BW, Kuncel NR, Shiner R, Caspi A, Goldberg LR. 2007; The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science*. 2(4):313–345. DOI: 10.1111/j.1745-6916.2007.00047.x [PubMed: 26151971]
- Soubelet A, Salthouse TA. 2011; Personality–cognition relations across adulthood. *Developmental Psychology*. 47(2):303–310. DOI: 10.1037/a0021816 [PubMed: 21142358]
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. 2002; Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 64(4):583–639.

- Sternberg RJ, Grigorenko E, Bundy DA. 2001; The Predictive Value of IQ. *Merrill-Palmer Quarterly*. 47(1):1–41. DOI: 10.1353/mpq.2001.0005
- The 1000 Genomes Project Consortium. 2015; A global reference for human genetic variation. *Nature*. 526(7571):68–74. DOI: 10.1038/nature15393 [PubMed: 26432245]
- Vilhjálmsón BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, ... Zheng W. 2015; Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*. 97(4):576–592. DOI: 10.1016/j.ajhg.2015.09.001 [PubMed: 26430803]
- Vrieze SI, Malone SM, Vaidyanathan U, Kwong A, Kang HM, ... Zhan X, et al. 2014; In search of rare variants: preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes. *Psychophysiology*. 51(12):1309–1320. [PubMed: 25387710]

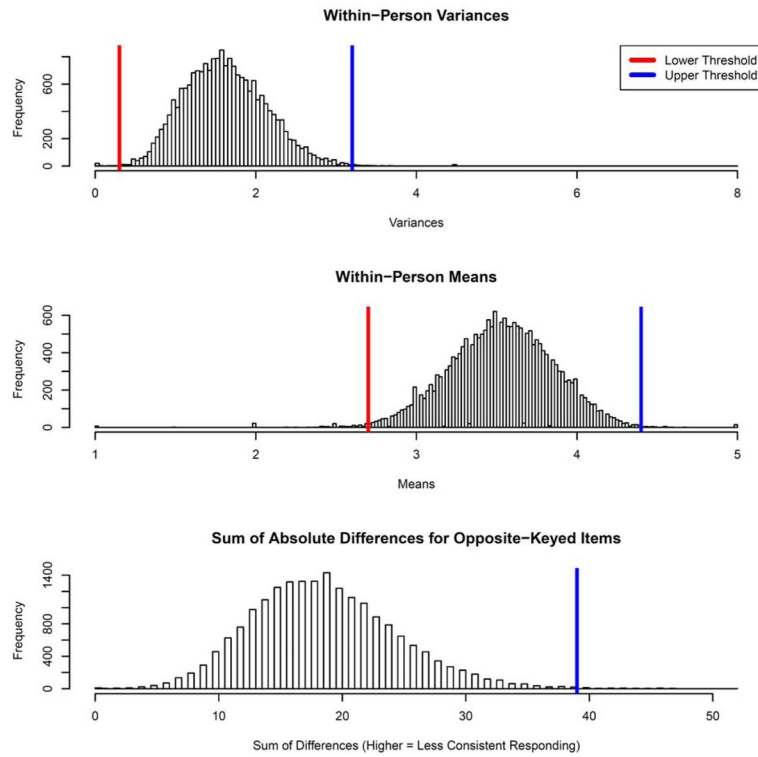


Figure 1. BFI Quality Control Filters. Participants were excluded if they scored below or above the lower and upper thresholds on these three quality control metrics computed from the Big Five Inventory. See text for details.

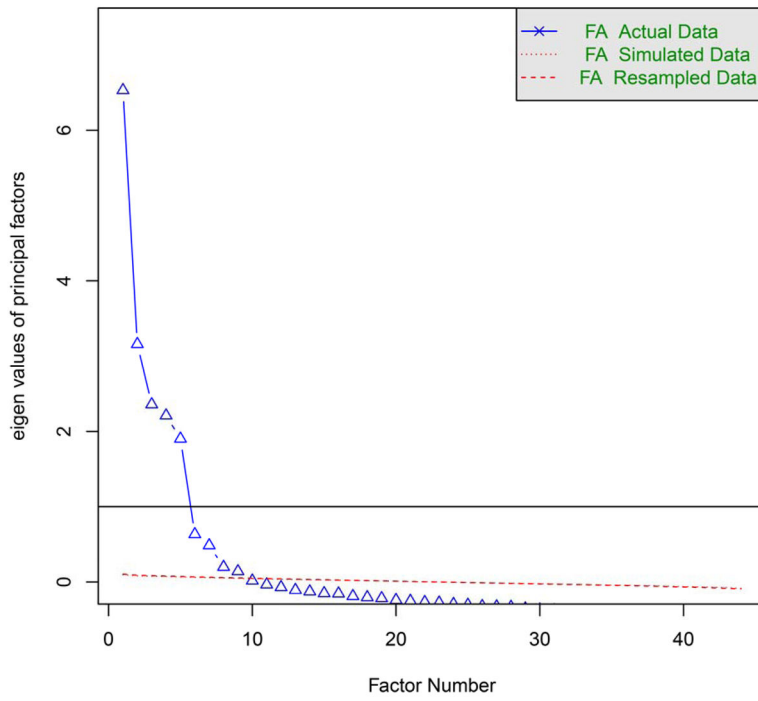


Figure 2. Scree Plot and Parallel Analysis of Age- and Sex-Residualized BFI Items.

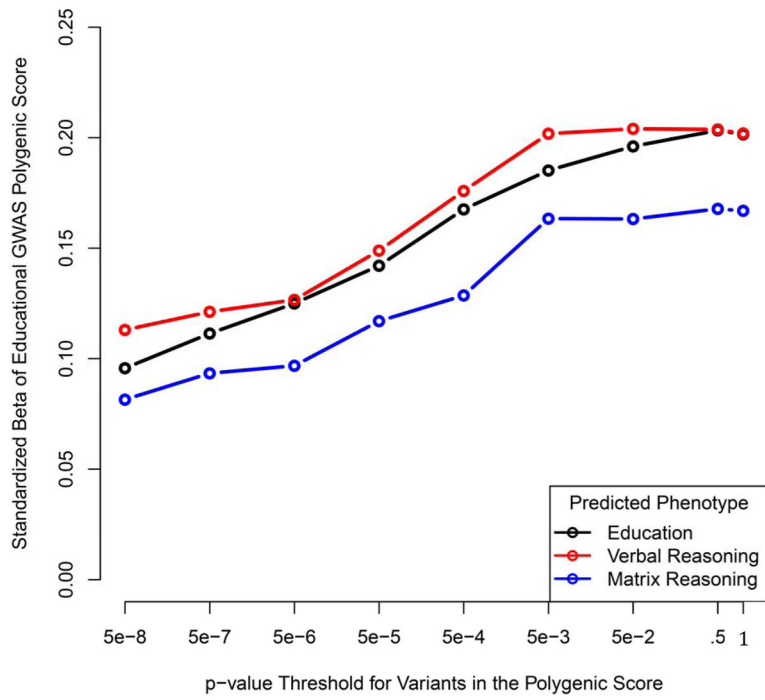


Figure 3. Education Polygenic Score Prediction of Educational Attainment, Verbal Reasoning, and Matrix Reasoning. All points are highly statistically significant.

Table 1

Descriptive Statistics on the Genes for Good Sample

Dataset	N	Age		% Male	% White	Income		Years of Education	
		M	SD			M	SD	M	SD
Genotyped Subsample	7,458	37.1	13.9	28%	80%	2.9	1.5	4.2	1.4
Verbal Reasoning	13,975	38.0	14.0	24%	81%	2.8	1.5	4.1	1.4
Matrix Reasoning	7,200	38.3	14.0	23%	81%	2.8	1.5	4.2	1.4
Personality	20,262	37.3	13.6	23%	81%	2.9	1.5	4.1	1.4
Education & Income	20,301	37.3	13.6	24%	81%	2.9	1.5	4.1	1.4
European Ancestry with Complete Genotypes and Phenotypes	2,683	39.3	14.4	23%	93%	2.9	1.5	4.3	1.4

Note: The total number of participants with at least one non-missing observation was 21,550. “% White” is the percentage of individuals who self-reported as only “White” and no other race/ethnicity.

Table 2

Verbal and Matrix Reasoning Descriptive Statistics

Item	N	% Missing	% Female	% Correct	Cor with Sumscore ^d	Response Time (seconds)			
						25 th ile	Median	75 th ile	Mean (SD)
VR 1	13,609	0	76	98	.07	-	-	-	-
VR 2	13,285	3	76	94	.28	23	31	44	39 (39)
VR 3	13,284	3	76	83	.18	9	12	18	17 (22)
VR 4	13,230	3	76	91	.26	8	11	15	14 (20)
VR 5	13,318	2	76	63	.42	20	28	43	38 (43)
VR 6	13,231	3	76	69	.53	24	37	60	51 (51)
VR 7	13,074	4	76	57	.34	34	50	74	63 (58)
VR 8	13,035	4	76	64	.61	14	23	41	36 (45)
VR 9	12,928	5	76	54	.46	24	37	59	52 (55)
VR 10	13,024	4	76	92	.20	14	19	26	24 (26)
VR 11	12,969	5	76	44	.35	12	17	27	27 (36)
VR 12	12,913	5	76	71	.49	25	36	52	46 (47)
VR 13	12,951	5	76	35	.34	13	18	28	25 (29)
VR 14	12,350	9	77	65	.28	8	13	23	20 (26)
VR 15	12,796	6	76	28	.16	11	16	28	24 (30)
VR 16	12,799	6	76	14	.29	26	48	90	76 (85)
MR 1	7199	0	77	55	.25	-	-	-	-
MR 2	7014	3	78	67	.36	19	29	45	39 (42)
MR 3	6941	4	77	85	.33	9	13	19	18 (25)
MR 4	6875	5	78	94	.21	8	10	14	13 (18)
MR 5	6965	3	77	31	.09	16	23	35	31 (36)
MR 6	6880	4	78	75	.41	25	37	57	49 (50)
MR 7	6546	9	77	40	.47	44	79	131	105 (100)
MR 8	6534	9	78	53	.43	26	41	66	56 (56)
MR 9	6491	10	78	79	.50	19	28	43	38 (39)
MR 10	6331	12	78	75	.45	24	36	57	50 (56)
MR 11	6387	11	78	82	.46	14	20	31	28 (31)

Item	N	% Missing	% Female	% Correct	Cor with Sumscore ^d	Response Time (seconds)			
						25 th %ile	Median	75 th %ile	Mean (SD)
MR 12	6251	13	77	59	.55	24	40	71	61 (72)
MR 13	6126	15	78	27	.38	25	46	85	72 (88)
MR 14	6104	15	78	81	.50	13	19	31	29 (42)
MR 15	6045	16	77	75	.47	22	37	64	56 (64)
MR 16	5846	19	78	58	.52	30	52	86	70 (70)
MR 17	5783	20	78	37	.46	28	55	100	83 (94)
MR 18	5874	18	78	59	.53	19	30	51	44 (57)
MR 19	5791	20	77	33	.41	23	44	84	71 (86)
MR 20	5743	20	77	40	.51	18	34	68	61 (79)
MR 21	5675	21	77	2	-.12	19	32	58	52 (67)
MR 22	5501	24	78	36	.18	22	52	121	101 (132)
MR 23	5482	24	77	18	.28	21	44	90	76 (97)
MR 24	5336	26	77	31	.34	25	57	122	99 (123)
MR 25	5247	27	78	16	-.02	18	44	99	82 (108)
MR 26	5369	25	77	20	.20	17	28	49	42 (51)
MR 27	5402	25	77	34	.24	21	36	62	51 (54)
MR 28	5450	24	78	33	.50	19	37	69	58 (73)
MR 29	5424	25	78	60	.39	12	18	29	26 (35)
MR 30	5420	25	77	34	.34	19	37	72	60 (73)

VR = verbal reasoning; MR = matrix reasoning; SD = standard deviation. Correlation between item and sumscore based on all other items. All results computed after too-fast and too-slow responses were removed.

IRT Model Fit Comparisons for Verbal and Matrix Reasoning

Table 3

Test	Model	Log-posterior	Estimated parameters	DIC	BF Relative to 2PL	BF Relative to 3PL
Verbal Reasoning	2PL	-101,571.2	32	203,206.4		
	3PL	-101,626.2	48	203,348.4	7.4e ³³	
	4PL	-101,599.3	64	203,326.5	1.5e ¹²	~0
Matrix Reasoning	2PL	-87,958.1	56	176,028.2		
	3PL	-87,571.9	84	175,311.8	~0	
	4PL	-87,516.0	112	175,256.1	~0	~0

2PL, 3PL, and 4PL indicate the number of parameters in the model, difficulty, discrimination, guessing, and upper bound, respectively. DIC = Deviance Information Criterion. BF = Bayes Factor; small values, which all reported values are, indicate support for the more complex model.

Table 4

IRT parameter estimates for Verbal and Matrix Reasoning

Item	Parameter Estimates			
	a	B	g	u
VR 1	0.65	-6.23	0	1
VR 2	1.52	-2.32	0	1
VR 3	0.57	-2.90	0	1
VR 4	1.18	-2.41	0	1
VR 5	1.29	-0.55	0	1
VR 6	2.09	-0.62	0	1
VR 7	0.98	-0.30	0	1
VR 8	2.70	-0.41	0	1
VR 9	1.43	-0.12	0	1
VR 10	0.95	-2.95	0	1
VR 11	1.04	0.31	0	1
VR 12	1.87	-0.70	0	1
VR 13	1.07	0.75	0	1
VR 14	0.79	-0.85	0	1
VR 15	0.46	2.22	0	1
VR 16	1.08	2.01	0	1
MR 1	0.69	-0.07	.13	.95
MR 2	2.11	-0.90	.09	.86
MR 3	2.28	-1.43	.11	.96
MR 4	2.45	-2.11	.12	.98
MR 5	1.40	2.49	.26	.95
MR 6	1.80	-.865	.13	.97
MR 7	2.19	0.38	.10	.88
MR 8	1.62	-0.10	.09	.92
MR 9	3.34	-0.84	.10	.99
MR 10	2.11	-0.74	.13	.98
MR 11	3.53	-0.97	.10	.98
MR 12	3.09	-0.06	.14	.98
MR 13	2.65	1.02	.11	.90
MR 14	3.31	-0.85	.12	.99
MR 15	2.33	-0.64	.15	.98
MR 16	2.38	-0.06	.12	.97
MR 17	2.59	0.65	.12	.92
MR 18	2.52	-0.10	.11	.97
MR 19	2.72	0.93	.14	.97
MR 20	3.29	0.58	.12	.98
MR 22	0.72	1.67	.15	.95
MR 23	3.11	1.56	.10	.95

Item	Parameter Estimates			
	a	B	g	u
MR 24	1.81	1.16	.13	.95
MR 26	1.52	2.03	.12	.95
MR 27	0.83	1.47	.12	.95
MR 28	4.22	0.78	.12	.97
MR 29	1.40	-0.13	.15	.97
MR 30	1.47	1.08	.13	.96

Note: a = discrimination, b = difficulty, g = guessing parameter, and u = upper bound. Based on model comparison results, the guessing parameter g and upper bound parameter u was fixed for all Verbal Reasoning items.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

Five-factor exploratory factor loadings of the Big Five Inventory

Item	I see myself as someone who... ^a	Expected Scale	Factor				
			1 (E)	2 (N)	3 (C)	4 (A)	5 (O)
1	Is talkative	E	.70	.10	-.10	-.01	.06
6 ^R	Is reserved	E	.80	.01	-.06	.02	-.05
11	Is full of energy	E	.28	-.26	.10	-.05	.16
16	Generates a lot of enthusiasm	E	.50	-.02	.11	.23	.20
21 ^R	Tends to be quiet	E	.86	.05	-.01	-.02	-.06
26	Has an assertive personality	E	.50	-.05	.26	-.25	.16
31 ^R	Is sometimes shy, inhibited	E	.70	-.19	.03	-.09	-.04
36	Is outgoing, sociable	E	.73	-.08	.04	.15	.04
4	Is depressed, blue	N	-.11	.56	-.12	.00	.06
9 ^R	Is relaxed, handles stress well	N	.04	.72	-.03	-.04	-.05
14	Can be tense	N	-.02	.67	.13	-.15	.06
19	Worries a lot	N	-.04	.73	.04	.06	.02
24 ^R	Is emotionally stable, not easily upset	N	.08	.69	-.08	-.06	-.05
29	Can be moody	N	.03	.58	-.02	-.16	.04
34 ^R	Remains calm in tense situations	N	.06	.51	-.13	-.04	-.18
39	Gets nervous easily	N	-.18	.65	-.03	.13	-.03
3	Does a thorough job	C	.01	.11	.70	.07	.01
8 ^R	Can be somewhat careless	C	-.13	-.06	.48	.08	-.10
13	Is a reliable worker	C	.01	.02	.58	.17	-.06
18 ^R	Tends to be disorganized	C	-.04	-.08	.61	-.09	-.14
23 ^R	Tends to be lazy	C	.05	-.11	.50	-.01	.03
28	Perseveres until the task is finished	C	-.02	-.02	.66	.00	.09
33	Does things efficiently	C	.01	-.01	.66	.05	.01
38	Makes plans and follows through with them	C	.06	-.06	.57	.02	-.02
43 ^R	Is easily distracted	C	-.14	-.24	.39	-.15	-.07
2 ^R	Tends to find fault with others	A	-.05	-.27	-.07	.47	-.02
7	Is helpful and unselfish with others	A	.09	.07	.15	.56	.07

Item	I see myself as someone who... ^a	Expected Scale	Factor				
			1 (E)	2 (N)	3 (C)	4 (A)	5 (O)
12 ^R	Starts quarrels with others	A	-.17	-.14	.10	.51	-.06
17	Has a forgiving nature	A	.02	-.13	-.09	.55	.05
22	Is generally trusting	A	.08	-.12	-.03	.38	-.03
27 ^R	Can be cold and aloof	A	.26	-.08	-.03	.47	-.09
32	Is considerate and kind to almost everyone	A	.03	.07	.09	.70	.02
37 ^R	Is sometimes rude to others	A	-.10	-.13	.02	.57	-.03
42	Likes to cooperate with others	A	.10	-.01	.12	.52	-.02
5	Is original, comes up with new ideas	O	.14	-.06	.12	.02	.59
10	Is curious about many different things	O	-.07	-.11	-.15	-.12	.52
15	Is ingenious, a deep thinker	O	-.12	.02	.04	-.08	.62
20	Has an active imagination	O	.00	.06	-.17	-.02	.64
25	Is inventive	O	.07	-.07	.11	.02	.64
30	Values artistic, aesthetic experiences	O	.07	.18	.05	.28	.39
35 ^R	Prefers work that is routine	O	.13	-.14	-.10	-.01	.23
40	Likes to reflect, play with ideas	O	-.11	-.03	-.08	.01	.64
41 ^R	Has few artistic interests	O	.05	.09	.02	.19	.30
44	Is sophisticated in art, music, or literature	O	.02	.06	.00	.11	.49
Factor Correlations							
			1 (E)	2 (N)	3 (C)	4 (A)	5 (O)
Factor 1	E	1					
Factor 2	N	-.16	1				
Factor 3	C	.15	-.27	1			
Factor 4	A	.10	-.21	.25	1		
Factor 5	O	.11	-.05	.04	.04	1	

Note: ^R = item is reverse coded. Five factor model with oblimin rotation. All loadings 0.3 are in bold.