MANKIND QUARTERLY 2021 61:4 792-819

Intelligence and General Psychopathology in the Vietnam Experience Study: A Closer Look

Emil O. W. Kirkegaard*
Ulster Institute for Social Research, London, UK

Helmuth Nyborg
Professor emeritus Aarhus University, Denmark (1968-2007)

* Corresponding author: the.dfx@gmail.com

Prior research has indicated that one can summarize the variation in psychopathology measures in a single dimension, labeled P by analogy with the *g* factor of intelligence. Research shows that this P factor has a weak to moderate negative relationship to intelligence. We used data from the Vietnam Experience Study to reexamine the relations between psychopathology assessed with the MMPI (Minnesota Multiphasic Personality Inventory) and intelligence (total n = 4,462: 3,654 whites, 525 blacks, 200 Hispanics, and 83 others). We show that the scoring of the P factor affects the strength of the relationship with intelligence. Specifically, item response theory-based scores correlate more strongly with intelligence than sum-scoring or scale-based scores: *r*'s = -.35, -.31, and -.25, respectively. We furthermore show that the factor loadings from these analyses show moderately strong Jensen patterns such that items and scales with stronger loadings on the P factor also correlate more negatively with intelligence (*r* = -.51 for 566 items, -.60 for 14 scales). Finally, we show that training an elastic net model on the item data allows one to predict intelligence with extremely high precision, *r* = .84. We examined whether these predicted values worked as intended with regards to cross-racial predictive validity, and relations to other variables. We mostly find that they work as intended, but seem slightly less valid for blacks and Hispanics (*r*'s .85, .83, and .81, for whites, Hispanics, and blacks, respectively).

Keywords: Vietnam Experience Study, MMPI, General psychopathology factor, Intelligence, Cognitive ability, Machine learning, Elastic net, Lasso, Random forest, Crud factor

Since the dawn of scientific psychiatry it has been noted that there are positive associations ("comorbidity") between clinical diagnoses of mental disorders (Caspi et al., 2014; Haltigan, 2019; Pettersson et al., 2013). This covariation has generally been considered a problem to the typological thinking or classification of diagnoses (nosology) since it results in unclear boundaries between disorders that were supposed in many theories to be distinct. To intelligence researchers, however, this pattern is not surprising as they are used to thinking of general factors in datasets, as well as generalist (pleiotropic) genes that cause such patterns (Trzaskowski et al., 2013). Accordingly, and with explicit analogy to the *g* factor of intelligence, some researchers have conducted factor analyses of psychopathology measures such as behavioral ratings or clinical diagnoses (Laceulle et al., 2015; Martel et al., 2017; Patalay et al., 2015; Tackett et al., 2013). Given the observed positive correlations among disorders or symptom indicators (the positive manifold), factor analysis invariably finds a general factor that 'accounts for' part of the observed variation.[1] Recently, this factor has accordingly been named general psychopathology factor, or P factor for short, not to be confused with the Psychoticism factor from Eysenck's personality model.[2]

Research on the P factor has found it to be moderately to highly heritable based on both pedigree data and DNA (Neumann et al., 2016; Pettersson et al., 2013, 2016; Rietz et al., 2020). Furthermore, some studies have analyzed correlations between the P factor and measures of intelligence finding small to moderate negative correlations. WAIS (Wechsler Adult Intelligence Scale), adults: -.19, WISC-R (Wechsler Intelligence Scale for Children), age 7-11: -.15, SB (Stanford Binet intelligence test), age 5: -.17 (Caspi et al., 2014); age 6-8: -.14 (Neumann et al., 2016). However, the nature of these relations has not so far been explored in detail. Thus, the first goal of the present study was to conduct such an exploration.

---

[1] Despite the causal-sounding language, this simply means that one can hypothesize or posit a factor and a vector of factor loadings such that one can reproduce the observed correlations among the indicators, at least to some degree.

[2] In fact, historical discussions and findings of this factor go back many decades (Barron, 1953; Choca et al., 1986; Gourlay, 1980; Pukrop et al., 2001; Walters et al., 2008).

One prior study based on the same dataset as ours exists (Irwing et al., 2012). This study, however, used only the MMPI scales to extract a general factor, even though item data are also available. Furthermore, that study interpreted the general factor from the scales as reflecting a general factor of personality ('GFP'), not the general factor of psychopathology (P factor). Considering the mixed nature of the MMPI items, it is unclear what this general factor represents exactly. The MMPI items were not designed to measure a broad trait with measurement purity (i.e., free of other trait content insofar as possible), as is usually done in modern scale development (Yarkoni, 2010, 2015). Rather, items were chosen such as to best discriminate between clinical and general populations, no matter what the item content was (Cox et al., 2009). Furthermore, a number of studies using both 'normal range' personality data and psychopathology measures have found that the general factors from these are highly correlated. There is also a strong relation to trait emotional "intelligence" (Alegre et al., 2019; Musek, 2017; Oltmanns et al., 2018; Rosenström et al., 2019; Smith et al., 2020), as also indicated by the large personality gaps between the general population and psychiatric groups (Kotov et al., 2010). We remain agnostic about the specific interpretation of these general factors.

In a different strand of research, it has been shown that supervised machine learning models trained on large item-level datasets are able to predict various variables of interest (e.g. education, age) much better (sometimes >100%) than scores generated from ad hoc methods or common factor based models (Cutler et al., 2019; Mõttus et al., 2017). So far, the research has been conducted using personality and cognitive ability data, but not psychiatric data, or a set of mixed personality-interest-psychiatric questions like those in the MMPI. The present study presented an opportunity to try with such a dataset, i.e., apply machine learning methods to our item data to examine predictive validity, and this was thus the second goal.

Data

We used archival data from the Vietnam Experience Study (VES, https://www.cdc.gov/nceh/veterans/default1c.htm). The VES is a US military dataset based on a sample of 4,462 enlisted men (3,654 whites, 525 blacks, 200 Hispanics, 49 Amerindian/Native Americans, and 34 Asians). They were inducted in the military between 1965 and 1971, and a follow-up interview was conducted in 1985-1986. The purpose of the study was to examine reports of negative health effects of participation in the Vietnam War, in particular, exposure to chemical warfare (Centers for Disease Control, 1989). Because of this, the study is a matched group design with approximately 55% of the sample being Vietnam

veterans, and the remainder being soldiers who were stationed elsewhere, such as South Korea. The dataset is in the public domain (i.e., it is not copyrighted, or available only upon request or application) but no complete repository exists for it. The supplementary materials contain a copy of the dataset that was used in the present study.

Generally speaking, the measurements were of very high quality. The total dataset contains thousands of variables derived from objective ability and skill testing, clinical interviews, self-rated scales, blood analyses, and even some measures of their spouses and children. The children were measured due to reports of birth defects following service in Vietnam.

The subjects took a large and diverse battery of cognitive tests, some of which were taken at the time of induction and others at the follow-up. In total, there are 19 tests which cover domains such as psychomotor ability (e.g. drawing), delayed recall, verbal ability, and spatial ability (block design). These have been described in detail elsewhere (Kirkegaard & Nyborg, 2020), and the appendix contains a summary of the tests and the factor analysis. As in the other studies, we computed the *g* factor from the 19 tests and saved it for further analysis.

The subjects were administered the MMPI (Minnesota Multiphasic Personality Inventory), version 1975, at the follow-up measurement. This was administered as a group test, where subjects sat in a room and filled out the answer sheets. The MMPI version used in this study consists of 566 self-reported items dealing with various aspects of behavior, interests, and personality (Dahlstrom et al., 1975). Importantly, the latter includes both healthy variation in personality as well as psychopathologically relevant variation. The appendix contains the 25 first items of the MMPI-2 from which the reader may form an impression. While the focus is on psychopathology-related personality, various researchers have devised scoring methods (simple models) for estimating e.g. Big Five (OCEAN) traits from the MMPI data (Cortina et al., 1992; Han et al., 1996). While the MMPI can be scored in many ways, the following scales were used in this study, shown in Table 1.

Table 1. *Scales scored from the MMPI, with loadings on the general psychopathology (P) factor.*

Scale[1]	Abbreviation	Items	Loading
Hypochondriasis	Hs	32	0.65
Depression	D	57	0.74
Hysteria	Hy	60	0.53
Psychopathic Deviate	Pd	50	0.62

Scale[1]	Abbreviation	Items	Loading
Masculinity/Femininity	Mf	56	0.30
Paranoia	Pa	40	0.69
Psychasthenia	Pt	48	0.88
Schizophrenia	Sc	78	0.91
Hypomania	Ma	46	0.31
Social Introversion	Si	69	0.46
Lie (social desirability)	L	15	-0.17
F	F	60	0.80
F (back)	Fb	40	(not used)
K	K	30	-0.35
Ego Strength	Es	52	-0.77

[1] Descriptions of some of the scales from Framingham, (2016):

Psychasthenia: "person's inability to resist specific actions or thoughts, regardless of their maladaptive nature."

Lie (social desirability): "intended to identify individuals who are deliberately trying to avoid answering the MMPI honestly and in a frank manner. The scale measures attitudes and practices that are culturally laudable, but rarely found in most people."

F: "intended to detect unusual or atypical ways of answering the test items, like if a person were to randomly fill out the test."

F (back): same as F but only based on the last 40 items, a measure of test fatigue,

K: "designed to identify psychopathology in people who otherwise would have profiles within the normal range."

Note that some items are used in multiple scales (akin to cross-loadings). Loadings from factor analysis.

Results

We factor analyzed the 14 scales of the MMPI to obtain a single factor, similar to the prior study (Irwing et al., 2012). All the scales had non-zero loadings, though not all were in the same direction (mean loadings = .40, range -.77 to .36, variance accounted for = 39%). The loadings are given in Table 1, above. To test robustness, we tried all the factor analysis and scoring method choices available in the **psych** package (Revelle, 2020).[3] The scores from these correlated near

[3] Specifically, the *fa()* function in **psych** contains a setting for the factor extraction method (loading estimation) and the factor scoring method. There are 5 extraction methods and 8 scoring methods, so there are theoretically 40 sets of resulting scores. However, frequently, some of these methods or their combinations result in errors. In the present

unity (mean $r = 1.00$) meaning there was negligible method variance. We saved the scores (P scales) from this analysis for further use.

While there were no missing data for the scales, there were some for the items (0.2% missing cells). We imputed these with 0's and computed the sum of affirmative answers for each person as a simple index of overall psychopathology (P sum, sumscore).[4] We also used item response theory (IRT) factor analysis (FA) on the items to produce a better estimate of the underlying trait, using the 2PL model from the **mirt** package (Chalmers et al., 2020).[5] The difference to the sumscore is that this approach allows for variation in the factor loadings of items, including potentially negative loadings (reverse scoring).[6] The three MMPI scoring methods were then compared to the *g* scores from above, as shown in Figure 1.

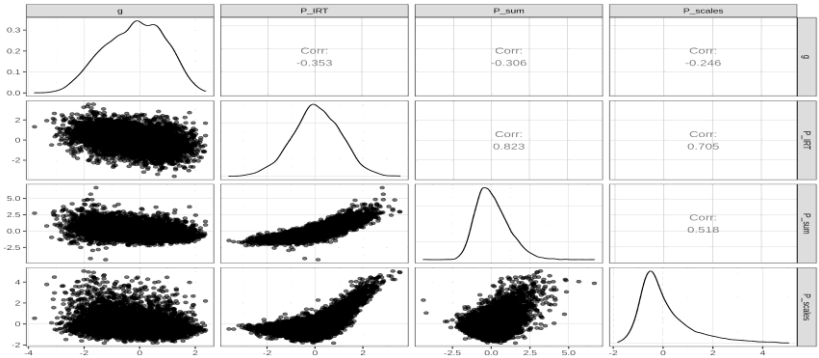

Figure 1. *Pairwise correlations between MMPI-based general psychopathology (P) scores and intelligence (g) from 19 tests. IRT = item response theory-based, sum = sumscores, scales = factor analysis of 14 scales.*

study, 32 of the 40 combinations produced a result. This method is implemented in the *fa_all_methods()* in the **kirkegaard** package (https://github.com/Deleetdk/kirkegaard/).

[4] We also calculated the average score without this imputation. These correlated 1.00 with the imputed version.

[5] Item response theory is a theoretical approach to modeling item responses specifically, not test/scale level. Item data requires other methods to model because the measured data are not continuous, usually dichotomous for ability data, but may be polytomous (ordinal). See DeMars (2010) for a brief and easily readable introduction. Item response theory factor analysis is factor analysis carried out on the estimated item correlations free from bias from the item format (latent correlations, usually tetrachoric).

[6] Formally speaking, sumscores are or can be seen as a factor score too. The sumscore is identical to a factor model where all items load on the general factor and have loadings of 1 (McNeish & Wolf, 2020).

The results show that the three scores' approaches of the P factor are fairly strongly related (r's .52 to .82), but not near-unity as might be expected from this number of items and scales. Furthermore, the scores from each method were related to the *g* factor, but differentially. The scoring method based on IRT produced the strongest correlation (r = -.35), the sumscore in between (-.31), and the scale-level factor analysis the weakest (-.25). The scatterplots also show some level of nonlinearity in the data, in particular between P from scales and IRT.

Similarly, we can look at the race gaps in MMPI-based P scores (Hall et al., 1999; Lynn, 2002). Because of the correlation to intelligence, all else equal, we would expect small to medium sized race gaps in MMPI scores due to the large race gaps in intelligence. Table 2 shows the results.

Table 2. *Racial means in P factor scores and intelligence (g). Scores were standardized to white mean = 0, sd = 1.*

Group	*g*	P scales	P scales expected	P sum	P sum expected	P IRT	P IRT expected
Black	-1.27	0.24	0.31	0.55	0.39	0.37	0.45
Hispanic	-0.78	0.46	0.19	0.37	0.24	0.37	0.27
White	0	0	0	0	0	0	0

The observed race gaps in P scores were in the expected direction, but not entirely of the expected size considering the relationship to intelligence. Thus, it seems that there are other factors at play with regard to P factor gaps than intelligence. We used regression analysis as an alternative approach to examining this. The results are shown in Tables 3a, 3b.

Table 3a. *Regression models for predicting P scores from MMPI. Standardized β with standard deviation in parentheses. * p < .01, ** p < .005, *** p < .001.* g = *intelligence. Numerical variables are standardized to white mean = 0, sd = 1.*

	Main effects only		
Outcome →	Scales	Sum	IRT
Intercept	-0.01 (0.017)	0.00 (0.017)	0.00 (0.016)
g	-0.24*** (0.016)	-0.28*** (0.016)	-0.33*** (0.015)
Black	-0.04 (0.052)	0.20*** (0.052)	-0.05 (0.048)
Hispanic	0.22** (0.077)	0.14 (0.077)	0.07 (0.072)
g * Black			
g * Hispanic			
R^2 adj.	0.062	0.097	0.124
N	4238	4238	4238

Table 3b. *Regression models for predicting P scores from MMPI. Standardized β with standard deviation in parentheses. * p < .01, ** p < .005, *** p < .001.* g = *intelligence. Numerical variables are standardized to white mean = 0, sd = 1.*

	With interactions		
Outcome →	Scales	Sum	IRT
Intercept	-0.01 (0.017)	0.00 (0.017)	0.00 (0.016)
g	-0.23*** (0.017)	-0.27*** (0.017)	-0.33*** (0.016)
Black	-0.07 (0.081)	0.15 (0.081)	0.01 (0.076)
Hispanic	0.08 (0.100)	0.12 (0.100)	0.02 (0.094)
g * Black	-0.03 (0.055)	-0.04 (0.055)	0.04 (0.051)
g * Hispanic	-0.19 (0.086)	-0.02 (0.086)	-0.06 (0.080)
R^2 adj.	0.063	0.096	0.123
N	4238	4238	4238

The results show that for the scales and sum-based scores there was an extra effect of race beyond that of *g*, though in two different models (for Hispanics for scales-based, and for Blacks with sum-based scores). However, when the IRT scores were used, race no longer made any difference, and the beta for intelligence was also the strongest (-.33). The addition of an intelligence * race interaction term did not alter this result, and the interaction terms did not improve model fits. Thus, the conclusion is that when IRT-based scores are used, the population differences in intelligence explain almost perfectly the observed P factor gaps, at least, insofar as we can see with the statistical precision of this study.

With regards to the scales and items, one might wonder what the Jensen pattern looks like (Rushton, 1998). The Jensen pattern is the relationship between each indicator's (item or scale) factor loading and its relationship to some variable of interest, such as the P factor.[7] Figures 2 and 3 show scatterplots of these relationships.

[7] This method is usually called *the method of correlated vectors* (MCV), as named by Arthur R. Jensen, who formalized the method based on an idea from Charles Spearman (Jensen, 1998). However, this name is misleading because mathematically speaking, any correlation is a correlation of vectors, so the name describes nothing more than the regular correlation. Furthermore, the idea is more general and can also be used with multiple regression models and other multivariate methods (Al-Bursan et al., 2018). For this reason, it is more apt to name it *Jensen's method*, in his honor.

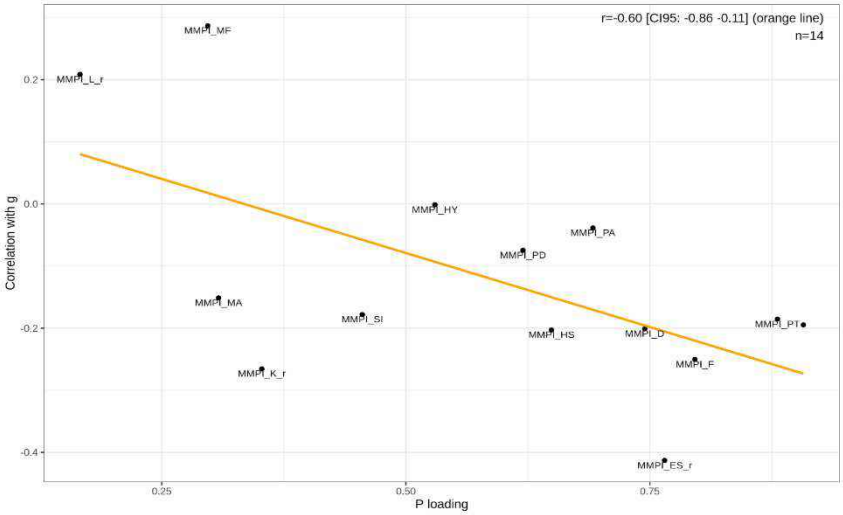

Figure 2. *Relationship between each scale's P factor loading and its correlation with intelligence. Indicators with negative factor loadings were reversed to avoid variance inflation (shown by "_r" in the scale names).*

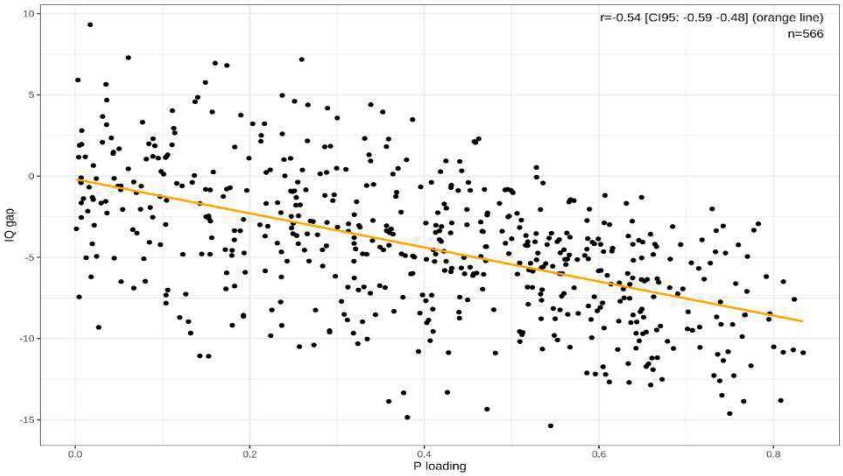

Figure 3. *Relationship between P factor loading and the IQ gap for each item (gap between those who answer yes and no). Indicators with negative factor loadings were reversed. The appendix gives the top 25 most IQ-associated items. Without reversing, the correlations are -.73 and -.76, for scales and items,*

respectively.

The results show moderately strong Jensen patterns such that indicators with stronger P factor loadings also had stronger, more negative relations to intelligence. If the underlying P factor is related to intelligence, this pattern is expected to be close to 1.00 given no sampling error. Thus, given our relatively large sample size, and resulting small sampling error, the results suggest that there is some true heterogeneity in the relations. In other words, the relationship to intelligence is more than just a function of an item's P loading. This also has the implication that a supervised model trained on the item data to predict intelligence could do much better than just extracting the latent trait of the predictors and estimating the outcome from that.[8] To test out this idea, we trained an elastic net model on the MMPI items to predict intelligence.[9] We used standard 10 fold cross-validation using the **tidymodels** meta-package to accomplish this (https://www.tidymodels.org/, Kuhn et al., 2020). The resulting out-of-sample predictions for the best choice of hyperparameters are shown in Figure 4.[10]

The resulting model accuracy was surprisingly high, $r = .84$, more than double that of the IRT-based score ($r = -.35$, direction is irrelevant here). In fact, the

[8] In fact, the latent trait model will only do best when the Jensen pattern has a true correlation of or close to 1.00. Empirically, such correlations exist, but rarely reach 1.00 even with statistical corrections for downwards biases, and this means there is some room for the supervised model to do better. Put another way, predictions based on latent traits are making a strong assumption: the variance shared among the indicators is exactly the same variance that is responsible for the predictive validity. This assumption seems to be largely correct, though not entirely so (Kirkegaard, 2018).

[9] Elastic net is a linear regression model with two hyperparameters. The first is the penalty, which shrinks all regression weights (betas) towards 0 with the aim of reducing overfitting. The second is the mixture, which controls the behavior of the shrinking towards 0. When mixture is closer to 1, the betas tend to become exactly 0, and the model is thus relatively sparser (only some predictors are used in the model). When the mixture is set to 1, the model is called the lasso, and when it is 0, it is called ridge regression. The values of the hyperparameters are usually tuned by cross-validation (James et al., 2013).

[10] We used the default settings of **tidymodels** which is to perform random search for the optimal hyperparameters and choose the optimal values from the same cross-validation. This approach causes a small amount of overfitting, but the amount was negligible in the present study (the standard errors of the cross-validated model validities were very small, approximately 0.006).

predictive accuracy is so high that using this prediction would be superior to using any single test of intelligence, especially an abbreviated test.[11]

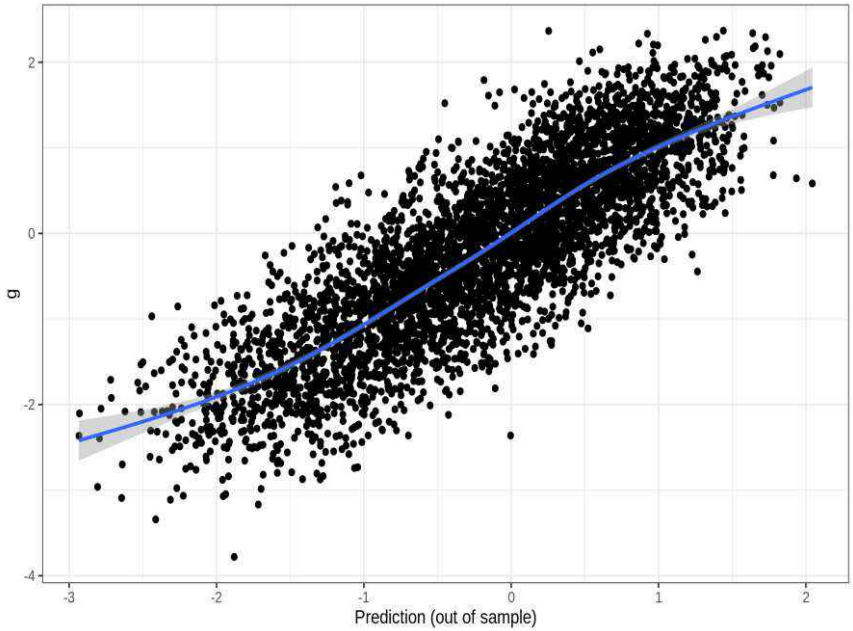

Figure 4. *Out of sample predictions for intelligence from MMPI items based on the elastic net model, r = .84. Blue line is a LOESS smoothing fit for visual aid.*

To further examine predictive accuracy, we trained a lasso model to see if a relatively sparse model could be obtained. The validity of the lasso model, however, was essentially identical to the elastic net one, and the optimal lasso fit was not very sparse (363 out of 556 items used).[12] Figure 5 shows the relationship between model accuracy and model hyper-parameters for the elastic net.

[11] Of the 19 tests used to extract *g* in this study, only the AFQT composite given at induction correlated at the same level with the *g* factor (.83, without including itself) as the MMPI based prediction. Thus, this level of model accuracy is roughly equivalent using the AFQT battery to measure intelligence (a battery of 4 tests). For single tests, the ACB verbal test had a correlation of .80 with the *g* factor (not including itself).

[12] The similarity of the results is not surprising considering that the lasso is a special case of the elastic net, as noted in a prior footnote.

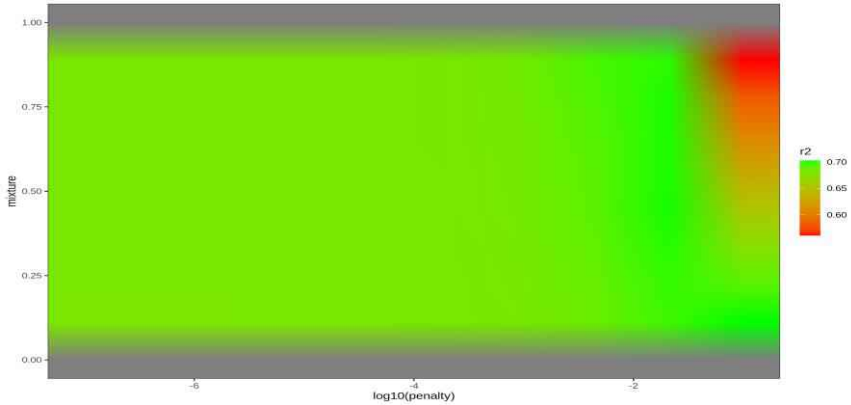

Figure 5. *Elastic net hyperparameters and model R^2 in cross-validation. Smoothing done by LOESS.*

We see that accuracy is highest when some penalization is present and mixture is low (towards ridge regression, i.e., non-sparse). We also fit a random forest model, but this had slightly inferior performance to the elastic net and lasso models (out of sample $r = .78$).

To examine to which degree the model could be abbreviated without loss of accuracy, i.e., based on fewer items, we calculated the cross-validated correlations for different numbers of items, shown in Figure 6.

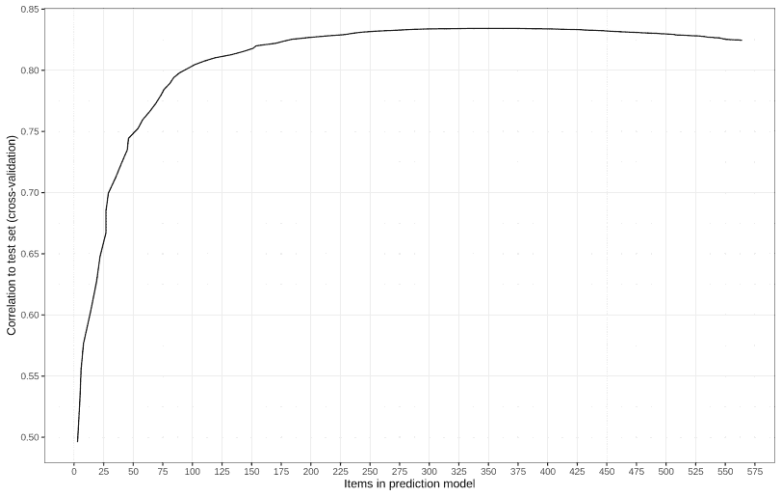

Figure 6. *Model accuracy by number of MMPI items included in the model (cross-validated).*

It is seen that about 90 items are needed to reach a correlation accuracy of .80, whereas only 3 items are needed to reach .50. This may be surprising, but some items have absolute correlations to *g* of around .40, so it is unsurprising that combining three of them yields a model accuracy at .50. To see whether the predicted *g* scores were biased with regards to demographics or variables of interest, we correlated the original *g* scores and the MMPI-based predicted values with various known correlates of intelligence, shown in Table 4.

By comparing the two columns with *g* and predicted *g*, we see that the values are very similar. The predicted *g* values based on MMPI data behave close to exactly as the real *g* values (correlation between the columns r = 1.00).[13] This is reassuring as it shows the model isn't introducing construct irrelevant variance into the predictions. In fact, the correlation sizes are not different either, suggesting that the remaining variance in the *g* scores is not predictive of outcomes.

Table 4. *Correlations between measured intelligence (g), MMPI-based predicted intelligence, and various outcomes; predicted g = final predictions from elastic net model (not out of sample).*

	Measured g	Predicted g	Educ.	Income	Unempl.	Height	BMI
predicted g	0.86						
Education	0.55	0.53					
Income	0.40	0.40	0.35				
Unemployed	-0.22	-0.23	-0.15	-0.48			
Height	0.14	0.14	0.09	0.09	-0.03		
BMI	-0.05	-0.05	-0.02	0.02	-0.02	-0.01	
Age	0.08	0.09	0.16	0.18	-0.11	0.01	0.06

With regards to demographic bias, we tested whether the predicted *g* values worked differently across races. To do this, we fit regression models with the real *g* scores as the outcome, and the predicted g scores as the predictor along with race. Results are shown in Table 5.

[13] Similarly, one might ask whether the correlations to the tests used to extract intelligence are similar to their loadings on the *g* factor. The answer is yes, the relationship is nearly perfect, *r* = .97.

Table 5. *Regression models comparing the predictive validity of the predicted g scores. Predicted variable is measured* g. ** p < .01, ** p < .005, *** p < .001. Standard errors in parentheses.*

Predictor/Model	1	2	3
Intercept	0.19*** (0.008)	0.22*** (0.008)	0.22*** (0.008)
predicted_g	0.98*** (0.009)	0.95*** (0.009)	0.96*** (0.010)
Black		-0.25*** (0.025)	-0.35*** (0.038)
Hispanic		-0.10* (0.037)	-0.15** (0.050)
predicted_g * Black			-0.11*** (0.031)
predicted_g * Hispanic			-0.08 (0.048)
R^2 adj.	0.741	0.747	0.747
N	4379	4379	4379

We see that adding the race variable (Model 2-3 from 1) has only a minor effect on the model fit (R^2 gain is .006). Because of our large sample size, we still see that the race variables are detected as useful in the model (small p values), both main and (for blacks) interaction effects. Thus, prediction of *g* from MMPI items was not race-invariant. Figure 7 shows the resulting predictive bias.

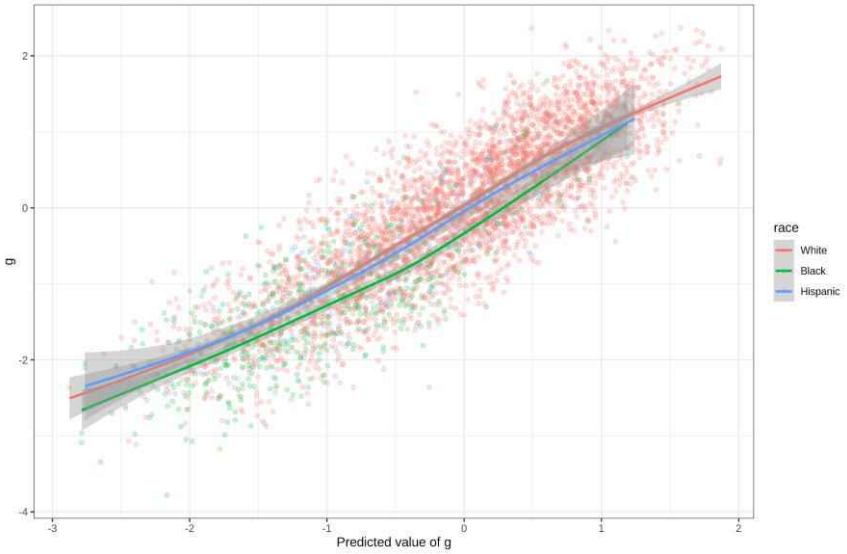

Figure 7. *Predictive validity by race. Lines by LOESS.*

We see that the lines are somewhat different. The line for blacks is somewhat below the white line except for the high end of g. In terms of correlations, the values are .85, .83, and .81, for whites, Hispanics, and blacks, respectively.

Discussion

We examined the relationships between the general psychopathology factor (P factor) scored from the MMPI data, and intelligence extracted from a battery of 19 diverse cognitive tests. We find that the scoring method of MMPI data has a large effect on the strength of the association to intelligence. When we extracted the P factor from the 14 MMPI provided scales, the scores correlated at -.25 with intelligence, when we computed a sumscore from the items, the resulting factor correlated -.31 with intelligence, and when we fit an item response theory (IRT, 2PL) model to the item data, the scores from this correlated at -.35 with intelligence. We examined race gaps in the MMPI scores as a validation test, and found that these could be partially (scales-based and sumscores) or entirely (IRT scores) explained as being a function of the intelligence gaps. Furthermore, there was a moderately strong Jensen pattern in the data such that indicators with stronger loadings on the P factor also had more negative relations to intelligence. This pattern was seen for both scales (r = -.60) and item (r = -.51) factor analysis. The pattern suggests that the underlying construct of the P factor, or some related metric (e.g. network loading, (Christensen & Golino, 2020)) is related to intelligence, though not overwhelmingly so.

To examine whether predictions could be made more accurate by supervised learning, we fit machine learning algorithms (elastic net, lasso, and random forest) to predict intelligence from the MMPI items. Since prior research had indicated that penalized regression approaches work well, we began with an elastic net model. The accuracy of this model was surprisingly high, r = .84 in the out-of-sample predictions (standard 10-fold cross-validation). We then fit a lasso (L1 penalization) model to see if we could construct a relatively sparse model, but unfortunately, this was not very successful (363 out of 556 items used at the optimal R^2). Finally, we fit a random forest model. This performed slightly worse than the elastic net (r = .78). The failure of the random forest model to do better than the elastic net indicates that nonlinear and interaction effects are not important in a given dataset for the purpose of prediction. In other words, the additive assumption is supported for this dataset and outcome variable. Our variables were binary, so linearity is not a potential issue. This finding replicates the result found in a prior study using items from cognitive ability tests to predict an assortment of life outcomes or personal characteristics such as education, income, age, sex (Cutler et al., 2019; see appendix for a summary of this study).

The accuracy of the intelligence prediction is so high that unless one has multiple high-quality tests in a dataset, it would be more accurate to impute from MMPI items. If this finding holds and can be generalized to other traits, it has important implications for data collection and expansion. It suggests that it is possible to train machine learning algorithms to predict unobserved traits in datasets as long as these contain a large number of diverse items one can predict from. If the method works, it could augment our existing datasets with new traits for use in many analyses. Similarly, it could be used to improve the estimation of existing traits (i.e. reduce measurement error). One far-reaching idea here is that we could collect new datasets that contain items found in an old dataset as well as traits of interest that are missing in the old datasets. We then train a model to predict the desired but missing traits on the new data, and apply it to the old dataset. In this way, we can potentially choose and add new variables to existing datasets where the subjects are no longer available for measurement (e.g. deceased). The question of whether this approach would be successful depends on the ability of the predictive models to generalize across traits and datasets that were perhaps collected across decades in time.

A related approach to the one advocated above is combining and abbreviating scales (Eisenbarth et al., 2015; Yarkoni, 2010, 2015). Currently, most scales were and are developed with factorial purity in mind, often simply measured with Cronbach's alpha. People strive to create items that measure one trait only (weak cross-loadings) and measure it well (high loading). While this is good in some ways, it is highly inefficient. From an informative theoretic perspective, high correlations between items in a survey indicate redundancy. Cronbach's alpha is probably mainly used because it is taken as a proxy for what we really want, test-retest reliability (Revelle & Condon, 2019). However, it is possible to have a test with alpha reliability of approximately 0 while having very high retest reliability. If instead one had a pool of items that measured multiple or even many traits at once, one could develop models on these to simultaneously score the various traits. This could potentially result in quite short surveys or tests that measure a lot of aspects just as well as now by exploiting the cross-loadings properly. Based on this perspective, what one should strive for in items is in fact high cross-loadings and high test-retest reliability. A variety of algorithms have been suggested for shortening tests without causing increased construct validity, by directly training to retain a subset of items that retain the scale's relations to criterion variables. This approach is similar to the one used here, even more advanced (Raborn & Leite, 2020; Schroeders et al., 2016). We found, however, without such direct criteria related training that the results nevertheless did not indicate any construct contamination or invalidity.

Another way to think about the results is in terms of Paul Meehl's crud factor, which is the tendency for everything to be correlated with everything else, although usually weakly.[14] In fact, Edward L. Thorndike had already described a stronger version of this idea in 1920: "... in human nature good traits go together. To him that hath a superior intellect is given also on the average a superior character; the quick boy is also in the long run more accurate; the able boy is also more industrious. There is no principle of compensation whereby a weak intellect is offset by a strong will, a poor memory by good judgment, or a lack of ambition by an attractive personality. Every pair of such supposed compensating qualities that have been investigated has been found really to show correspondence." (Thorndike, 1920). Another way to state the same is to say that the true distribution of correlations among variables is importantly moved away from zero, forming a kind of mixed distribution, perhaps with 2 modes around +/- .10 or thereabouts. Thus, the reason why the results in the present study are so strong is that we sampled a very large collection of items with a sufficiently large sample size to enable us to find the most predictive items without much overfitting. In fact there are a number of single items with latent correlations to *g* above |.40|, which is quite impressive for single items.

The main limitations of the current study are as follows. First, we used a single dataset, so it is not known to which extent the results will generalize across datasets. Because our sample size is relatively large, pure sampling error is not a likely problem with our general findings. Thus, an important task for future work is finding datasets that contain the same pool of item data, and examining predictive validity across datasets. It is particularly important to examine whether cross-dataset validity is stable for datasets collected decades apart, or whether it decays quickly; and whether or not it is stable across countries. In fact, whether this is the case is equivalent to asking whether the relations between traits that the models rely upon change much over time and are different in different places. If not, then the models should remain valid for datasets collected decades apart.

Second, we only had relatively large samples for whites, Hispanics and blacks. It's possible that the results work differently for other populations or racial groups. Our analyses of predictive invariance of the predicted intelligence scores revealed some departure from perfection. In fact, such results are expected on theoretical grounds in some cases (Borsboom et al., 2008). Specifically, unequal

[14] In fact, Meehl (1986, 1990) ascribes this name to David T. Lykken in a book chapter published before his well-known 1990 article. He had, however, discussed the idea without the name decades earlier. For a historical review of the general "everything is correlated" idea, see Branwen (2014).

variances across subgroups will result in deviations from predictive invariance. The appendix contains descriptive statistics for the primary variables.

Third, the dataset here concerns only men, only people who had military experience, and covered a fairly narrow age range. It is possible that this lack of diversity in demographics has impacted the results. In fact, a general narrowing of diversity in a sample is expected to lower trait correlations (range restriction). Thus, based on this limitation, one might expect results to be stronger with more diverse samples.

Acknowledgements and supplementary materials

We would like to thank the US Department of Defense for collecting this detailed dataset and putting it in the public domain. All study materials can be found in the OSF repository at https://osf.io/dbn4k/. Additionally, the R notebook from the study is available at https://rpubs.com/EmilOWK/VES_MMPI.

References

Al-Bursan, I.S., Kirkegaard, E.O.W., Fuerst, J., Bakhiet, S.F.A., Al Qudah, M.F., Hassan, E.M.A.H. & Abduljabbar, A.S. (2018). Sex differences in 32,347 Jordanian 4th graders on the National Exam of Mathematics. *Journal of Individual Differences* 40: 71-81. https://doi.org/10.1027/1614-0001/a000278

Alegre, A., Pérez-Escoda, N. & López-Cassá, E. (2019). The relationship between trait emotional intelligence and personality. Is trait EI really anchored within the Big Five, Big Two and Big One frameworks? *Frontiers in Psychology 10: 866.*

Barron, F. (1953). An ego-strength scale which predicts response to psychotherapy. *Journal of Consulting Psychology* 17: 327333. https://doi.org/10.1037/h0061962

Bayroff, A.G. & Fuchs, E.F. (1970). *The Armed Services Vocational Aptitude Battery*. U.S. Army Behavior and Systems Research Laboratory. https://apps.dtic.mil/docs/citations/AD0706832

Borsboom, D., Romeijn, J.-W. & Wicherts, J.M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods* 13: 75-98. https://doi.org/10.1037/1082-989X.13.2.75

Branwen, G. (2014). *Everything is correlated*. https://www.gwern.net/Everything

Buechley, R. & Ball, H. (1952). A new test of "validity" for the group MMPI. *Journal of Consulting Psychology* 16: 299-301. https://doi.org/10.1037/h0053897

Caspi, A., Houts, R.M., Belsky, D.W., Goldman-Mellor, S.J., Harrington, H., Israel, S., Meier, M.H., Ramrakha, S., Shalev, I., Poulton, R. & Moffitt, T.E. (2014). The p factor:

One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science* 2: 119-137. https://doi.org/10.1177/2167702613497473

Centers for Disease Control (1989). *Health status of Vietnam veterans*, Vol. I, Synopsis. Atlanta, GA: US Department of Health and Human Services.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C.F., Meade, A., Schneider, L., King, D., Liu, C.-W. & Oguzhan, O. (2020). *mirt: Multidimensional Item Response Theory* (1.32.1) [Computer software]. https://CRAN.R-project.org/package=mirt

Choca, J.P., Peterson, C.A. & Shanley, L.A. (1986). Factor analysis of the Millon Clinical Multiaxial Inventory. *Journal of Consulting and Clinical Psychology* 54: 253-255. https://doi.org/10.1037//0022-006x.54.2.253

Christensen, A.P. & Golino, H. (2020). Statistical equivalency of factor and network loadings [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xakez

Cortina, J.M., Doherty, M.L., Kaufman, G. & Smith, R.G. (1992). The "Big Five" Personality Factors in the Ipi and Mmpi: Predictors of Police Performance. *Personnel Psychology* 45: 119-140. https://doi.org/10.1111/j.1744-6570.1992.tb00847.x

Cox, A.C., Weed, N.C. & Butcher, J.N. (2009). *The MMPI-2: History, Interpretation, and Clinical Issues*. Oxford Univ. Press. https://doi.org/10.1093/oxfordhb/9780195366877.013.0014

Cutler, A., Dunkel, C.S., McLoughlin, S. & Kirkegaard, E.O.W. (2019). Machine learning psychometrics: Improved cognitive ability validity from supervised training on item level data. *International Society for Intelligence Research, Minneapolis, MN, USA*. https://www.researchgate.net/publication/334477851_Machine_learning_psychometrics_Improved_cognitive_ability_validity_from_supervised_training_on_item_level_data

Dahlstrom, W., Welsh, G. & Dahlstrom, L. (1975). *An MMPI Handbook Volume I Clinical Interpretation. A Revised Edition*. Univ. Of Minnesota Press.

DeMars, C. (2010). *Item Response Theory*. Oxford University Press.

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. https://doi.org/10.1037/pas0000032

Elwood, R.W. (1995). The California Verbal Learning Test: Psychometric characteristics and clinical application. *Neuropsychology Review* 5(3): 173-201.

Framingham, J. (2016, May 17). *Minnesota Multiphasic Personality Inventory (MMPI)*. //psychcentral.com/lib/minnesota-multiphasic-personality-inventory-mmpi/

Gourlay, N. (1980). Psychiatric group dimensions within the pen structure. *Personality and Individual Differences* 1: 121-127. https://doi.org/10.1016/0191-8869(80)90029-X

Greve, K.W., Stickle, T.R., Love, J.M., Bianchini, K.J. & Stanford, M.S. (2005). Latent structure of the Wisconsin Card Sorting Test: A confirmatory factor analytic study. *Archives of Clinical Neuropsychology* 20: 355-364. https://doi.org/10.1016/j.acn.2004.09.004

Hall, G.C.N., Bansal, A. & Lopez, I.R. (1999). Ethnicity and psychopathology: A meta-analytic review of 31 years of comparative MMPI/MMPI-2 research. *Psychological Assessment* 11: 186-197. https://doi.org/10.1037/1040-3590.11.2.186

Haltigan, J.D. (2019). Editorial: Putting practicality into "p": Leveraging general factor models of psychopathology in clinical intervention. *Journal of the American Academy of Child and Adolescent Psychiatry* 58: 751-753. https://doi.org/10.1016/j.jaac.2019.03.005

Han, K., Weed, N.C. & McNeal, T.P. (1996). Searching for Conscientiousness on the MMPI-2. *Journal of Personality Assessment* 67: 354-363. https://doi.org/10.1207/s15327752jpa6702_10

Irwing, P., Booth, T., Nyborg, H. & Rushton, J.P. (2012). Are *g* and the General Factor of Personality (GFP) correlated? *Intelligence* 40: 296-305. https://doi.org/10.1016/j.intell.2012.03.001

James, G., Witten, D., Hastie, T. & Tibshirani, R. (eds.) (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Jensen, A.R. (1998). *The g Factor: The Science of Mental Ability*. Praeger.

Kirkegaard, E.O.W. (2018, Oct. 3). The *g* factor and principal components regression. *Clear Language, Clear Mind*. https://emilkirkegaard.dk/en/?p=7414

Kirkegaard, E.O.W. & Nyborg, H. (2020). Pupil size and intelligence: A large-scale replication study. *Mankind Quarterly* 60: 525-538.

Kotov, R., Gamez, W., Schmidt, F. & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin* 136: 768-821. https://doi.org/10.1037/a0020327

Kuhn, M., Wickham, H. & RStudio (2020). *tidymodels: Easily Install and Load the "Tidymodels" Packages* (0.1.0) [Computer software]. https://CRAN.R-project.org/package=tidymodels

Laceulle, O.M., Vollebergh, W.A.M. & Ormel, J. (2015). The structure of psychopathology in adolescence: Replication of a general psychopathology factor in the TRAILS Study. *Clinical Psychological Science* 3: 850-860. https://doi.org/10.1177/2167702614560750

Leckliter, I.N., Matarazzo, J.D. & Silverstein, A.B. (1986). A literature review of factor analytic studies of the WAIS-R. *Journal of Clinical Psychology* 42: 332-342. https://doi.org/10.1002/1097-4679(198603)42:2<332::AID-JCLP2270420220>3.0.CO;2-2

Lynn, R. (2002). Racial and ethnic differences in psychopathic personality. *Personality and Individual Differences* 32: 273-316. https://doi.org/10.1016/S0191-8869(01)00029-0

Martel, M.M., Pan, P.M., Hoffmann, M.S., Gadelha, A., do Rosário, M.C., Mari, J.J., Manfro, G.G., Miguel, E.C., Paus, T., Bressan, R.A., Rohde, L.A. & Salum, G.A. (2017). A general psychopathology factor (P factor) in children: Structural model analysis and external validation through familial risk and child global executive function. *Journal of Abnormal Psychology* 126: 137-148. https://doi.org/10.1037/abn0000205

McNeish, D. & Wolf, M.G. (2020). Thinking twice about sum scores. *Behavior Research Methods* 52: 2287-2305. https://doi.org/10.3758/s13428-020-01398-0

Meehl, P.E. (1986). What social scientists don't understand. In: D.W. Fiske & R.A. Shweder (eds.), *Metatheory in Social Science: Pluralisms and Subjectivities*, pp. 315-338.

Meehl, P.E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports* 66: 195-244. https://doi.org/10.2466/pr0.1990.66.1.195

Merrill, R.M. & Heathers, L.B. (1956). The relation of the MMPI to the Edwards Personal Preference Schedule on a college counseling center sample. *Journal of Consulting Psychology* 20: 310-314. https://doi.org/10.1037/h0047297

Mõttus, R., Bates, T., Condon, D.M., Mroczek, D. & Revelle, W. (2017). *Your personality data can do more: Items provide leverage for explaining the variance and co-variance of life outcomes*. https://doi.org/10.31234/osf.io/4q9gv

Musek, J. (2017). *The General Factor of Personality*. Academic Press.

Neumann, A., Pappa, I., Lahey, B.B., Verhulst, F.C., Medina-Gomez, C., Jaddoe, V.W., Bakermans-Kranenburg, M.J., Moffitt, T.E., van IJzendoorn, M.H. & Tiemeier, H. (2016). Single nucleotide polymorphism heritability of a General Psychopathology Factor in children. *Journal of the American Academy of Child & Adolescent Psychiatry* 55: 1038-1045. https://doi.org/10.1016/j.jaac.2016.09.498

Oltmanns, J.R., Smith, G.T., Oltmanns, T.F. & Widiger, T.A. (2018). General factors of psychopathology, personality, and personality disorder: Across domain comparisons. *Clinical Psychological Science* 6: 581-589. https://doi.org/10.1177/2167702617750150

Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P. & Wolpert, M. (2015). A general psychopathology factor in early adolescence. *British Journal of Psychiatry* 207: 15-22. https://doi.org/10.1192/bjp.bp.114.149591

Pettersson, E, Anckarsäter, H., Gillberg, C. & Lichtenstein, P. (2013). Different neurodevelopmental symptoms have a common genetic etiology. *Journal of Child Psychology and Psychiatry* 54: 1356-1365. https://doi.org/10.1111/jcpp.12113

Pettersson, E., Larsson, H. & Lichtenstein, P. (2016). Common psychiatric disorders

share the same genetic origin: A multivariate sibling study of the Swedish population. *Molecular Psychiatry* 21: 717-721. https://doi.org/10.1038/mp.2015.116

Pukrop, R., Gentil, I., Steinbring, I. & Steinmeyer, E. (2001). Factorial structure of the German version of the Dimensional Assessment of Personality Pathology–Basic questionnaire in clinical and nonclinical samples. *Journal of Personality Disorders* 15: 450-456. https://doi.org/10.1521/pedi.15.5.450.19195

Raborn, A. & Leite, W. (2020). *ShortForm: Automatic Short Form Creation* (0.4.6) [Computer software]. https://CRAN.R-project.org/package=ShortForm

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research* (1.9.12.31) [Computer software]. https://CRAN.R-project.org/package=psych

Revelle, W. & Condon, D.M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment* 31: 1395-1411. https://doi.org/10.1037/pas0000754

Rietz, E.D., Pettersson, E., Brikell, I., Ghirardi, L., Chen, Q., Hartman, C., Lichtenstein, P., Larsson, H. & Kuja-Halkola, R. (2020). ADHD is more closely linked to neurodevelopmental than externalizing and internalizing disorders: A genetically informed multivariate Swedish population study. *MedRxiv*, 2020.02.26.20028175. https://doi.org/10.1101/2020.02.26.20028175

Rosenström, T., Gjerde, L.C., Krueger, R.F., Aggen, S.H., Czajkowski, N.O., Gillespie, N.A., Kendler, K.S., Reichborn-Kjennerud, T., Torvik, F.A. & Ystrom, E. (2019). Joint factorial structure of psychopathology and personality. *Psychological Medicine* 49: 2158-2167. https://doi.org/10.1017/S0033291718002982

Ruff, R.M. & Parker, S.B. (1993). Gender- and age-specific changes in motor speed and eye-hand coordination in adults: Normative values for the finger tapping and grooved pegboard tests. *Perceptual and Motor Skills* 76(3S): 1219-1230. https://doi.org/10.2466/pms.1993.76.3c.1219

Rushton, J.P. (1998). The “Jensen effect” and the “Spearman-Jensen hypothesis” of Black-White IQ differences. *Intelligence* 26: 217-225. https://doi.org/10.1016/S0160-2896(99)80004-X

Schroeders, U., Wilhelm, O. & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS ONE* 11(11): e0167110. https://doi.org/10.1371/journal.pone.0167110

Shin, M.-S., Park, S.-Y., Park, S.-R., Seol, S.-H. & Kwon, J.S. (2006). Clinical and empirical applications of the Rey–Osterrieth Complex Figure Test. *Nature Protocols* 1(2): 892-899. https://doi.org/10.1038/nprot.2006.115

Smith, G.T., Atkinson, E.A., Davis, H.A., Riley, E.N. & Oltmanns, J.R. (2020). The General Factor of Psychopathology. *Annual Review of Clinical Psychology* 16: 75-98. https://doi.org/10.1146/annurev-clinpsy-071119-115848

Tackett, J.L., Lahey, B.B., Hulle, C.V., Waldman, I., Krueger, R.F. & Rathouz, P.J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology* 122: 1142-1153. https://doi.org/10.1037/a0034151

Thorndike, E.L. (1920). Intelligence and its uses. *Harper's Magazine* 140: 227-235.

Tombaugh, T.N. (2006). A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Archives of Clinical Neuropsychology* 21: 53-76. https://doi.org/10.1016/j.acn.2005.07.006

Trzaskowski, M., Shakeshaft, N.G. & Plomin, R. (2013). Intelligence indexes generalist genes for cognitive abilities. *Intelligence* 41: 560-565. https://doi.org/10.1016/j.intell.2013.07.011

Walters, G.D., Knight, R.A., Grann, M. & Dahle, K.-P. (2008). Incremental validity of the Psychopathy Checklist facet scores: Predicting release outcome in six samples. *Journal of Abnormal Psychology* 117: 396-405. https://doi.org/10.1037/0021-843X.117.2.396

Witt, J.C. (1986). Review of the Wide Range Achievement Test-Revised. *Journal of Psychoeducational Assessment* 4: 87-90. https://doi.org/10.1177/073428298600400110

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality* 44: 180-198. https://doi.org/10.1016/j.jrp.2010.01.002

Yarkoni, T. (2015). *Internal consistency is overrated, or how I learned to stop worrying and love shorter measures, Part I*. http://www.talyarkoni.org/blog/2015/01/26/internal-consistency-is-overrated-or-how-i-learned-to-stop-worrying-and-love-shorter-measures-part-i/

Appendix

MMPI item examples:

The MMPI items are often analyzed but rarely presented. Since they are copyrighted, we cannot present them all here. However, the fair use clause of US copyright law allows us to present some of them for discussion. The list below contains the first 25 items from the MMPI 1975 version used in this study.

1. I like mechanics magazines.
2. I have a good appetite.
3. I wake up fresh and rested most mornings.
4. I think I would like the work of a librarian.
5. I am easily awakened by noise.

6. I like to read newspaper articles on crime.
7. My hands and feet are usually warm enough.
8. My daily life is full of things that keep me interested.
9. I am about as able to work as I ever was.
10. There seems to be a lump in my throat much of the time.
11. A person should try to understand his dreams and be guided by or take warning from them.
12. I enjoy detective or mystery stories.
13. I work under a great deal of tension.
14. I have diarrhea once a month or more.
15. Once in a while I think of things too bad to talk about.
16. I am sure I get a raw deal from life.
17. My father was a good man.
18. I am very seldom troubled by constipation.
19. When I take a new job, I like to be tipped off on who should be gotten next to.
20. My sex life is satisfactory.
21. At times I have very much wanted to leave home.
22. At times I have fits of laughing and crying that I cannot control.
23. I am troubled by attacks of nausea and vomiting.
24. No one seems to understand me.
25. I would like to be a singer.

MMPI top IQ associated items:

The table below gives the top 25 items with the strongest association with IQ, measured as the IQ gap size between those who answer "yes" and those who answer "no". In empirical fact, these all have negative gaps and positive P loadings, but this is not always the case. The first item with a positive relationship is the 30th. The 7th and 14th items are the same, and show similar associations. There are 15 pairs of duplicated items in the MMPI, designed to measure consistency to guard against random answers (Buechley & Ball, 1952; Merrill & Heathers, 1956).

	Item	Item	Prevalence	P loading	IQ gap
1.	Someone has been trying to poison me.	MM010151	0.01	0.54	-15.37
2.	I was a slow learner in school.	MM010260	0.38	0.38	-14.85
3.	I commonly hear voices without knowing where they come from.	MM010184	0.04	0.75	-14.62
4.	Dirt frightens or disgusts me.	MM010510	0.08	0.47	-14.34
5.	I hear strange things when I am alone.	MM010350	0.06	0.77	-13.86
6.	Sexual things disgust me.	MM010470	0.03	0.36	-13.86
7.	I am sure I get a raw deal from life.	MM010315	0.06	0.81	-13.81

	Item	Item	Prevalence	P loading	IQ gap
8.	I believe I am being followed.	MM010123	0.02	0.74	-13.49
9.	A windstorm terrifies me.	MM010392	0.11	0.38	-13.34
10.	In walking I am very careful to step over sidewalk cracks.	MM010213	0.09	0.43	-13.31
11.	People say insulting and vulgar things about me.	MM010364	0.11	0.66	-12.85
12.	The future is too uncertain for a person to make serious plans.	MM010395	0.20	0.63	-12.70
13.	I have had attacks in which I could not control my movements or speech but in which I knew what was going on around me.	MM010194	0.06	0.61	-12.67
14.	I am sure I get a raw deal from life.	MM010016	0.09	0.74	-12.60
15.	At times I have enjoyed being hurt by someone I loved.	MM010363	0.03	0.67	-12.51
16.	I believe I am a condemned person.	MM010202	0.04	0.75	-12.28
17.	I have often been frightened in the middle of the night.	MM010559	0.12	0.73	-12.27
18.	Sometimes I am strongly attracted by the personal articles of others such as shoes, gloves, etc., so that I want to handle or steal them though I have no use for them.	MM010085	0.01	0.61	-12.21
19.	I feel uneasy indoors.	MM010365	0.10	0.60	-12.18
20.	I believe my sins are unpardonable.	MM010209	0.05	0.59	-12.12
21.	Once a week or oftener I feel suddenly hot all over, without apparent cause.	MM010047	0.10	0.66	-11.91
22.	I cannot understand what I read as well as I used to.	MM010159	0.23	0.60	-11.74
23.	I have certainly had more than my share of things to worry about.	MM010338	0.42	0.65	-11.72
24.	I feel that I have often been punished without cause.	MM010157	0.13	0.77	-11.67
25.	I am troubled by attacks of nausea and vomiting.	MM010023	0.03	0.66	-11.56

Cognitive ability battery

This text is copied from (Kirkegaard & Nyborg, 2020).

1. Grooved Pegboard Test (GPT, right hand): A measure of manual dexterity and fine motor speed (Ruff & Parker, 1993). The speed score is the reciprocal of the number of seconds taken to place a set of pegs in a grooved hole as quickly as possible.
2. GPT (left hand).
3. Paced Auditory Serial Addition Test (PASAT): A measure of mental control, speed, and computational and attentional abilities (Tombaugh,

2006). The subject mentally adds a sequence of numbers in rapid succession. Score is the total number of correct responses.

4. Rey-Osterrieth Complex Figure Drawing (CFD): A measure of visuospatial ability and memory (Shin et al., 2006). The direct copy score (CFDD) is given from a subject reproducing a complex spatial figure while the figure is in full view.
5. CFD, copy from immediate recall. The immediate recall score (CFDI) is given from a subject reproducing a complex spatial figure immediately after being shown it.
6. CFD, copy from delayed recall. The delayed recall score (CFDL) is given from a subject being exposed to a complex spatial figure and, after 20 minutes of other activities, drawing it.
7. Wechsler Adult Intelligence Scale-Revised (WAIS-R), general information (Leckliter et al., 1986). A test of general knowledge.
8. WAIS-R, block design. A test of spatial ability.
9. Word List Generation Test (WLGT). A measure of verbal fluency. The subject generates as many words as possible which begin with the letters F, A, and S for 60 seconds. The score is the total number of words generated.
10. Wisconsin Card Sort Test (WCST). A measure of executive function (Greve et al., 2005). The score is the ratio of correct responses to countable responses.
11. Wide Range Achievement Test (WRAT). Measures ability to read aloud a list of single words (untimed) (Witt, 1986).
12. California Verbal Learning Test (CVLT). A measure of verbal learning and memory (Elwood, 1995). The subject recalls a list of 16 words over 5 repeated learning trials. The score is the total correct over 5 trials.
13. Army Classification Battery (ACB). A verbal test administered at induction (ACBVE) (Bayroff & Fuchs, 1970).
14. ACB verbal. Administered at the follow-up interview (ACBVL).
15. ACB arithmetic reasoning test. An arithmetic test administered at induction (ACBAE).
16. ACB arithmetic. Administered at the follow-up interview (ACBAL).
17. Pattern Analysis Test (PAT). A measure of pattern recognition administered at induction.
18. General Information Test (GIT). A test of general knowledge administered at induction.
19. Armed Forces Qualification Test (AFQT). A general aptitude battery. This measure is the total score on four subtests (word knowledge,

paragraph comprehension, arithmetic reasoning, mathematics knowledge) administered at induction.

Five of the tests (13, 15, 17-19) were given at induction and the remaining at the follow-up interview. Factor loadings are given below.

Test	*g*-loading	Test	*g*-loading
VE time1	0.82	PASAT	0.57
AR time1	0.81	WLGT	0.49
PA	0.70	Copy direct	0.47
GIT	0.69	Copy immediate	0.55
AFQT	0.85	Copy delayed	0.55
VE time2	0.82	CVLT	0.42
AR time2	0.82	WCST	0.46
WAIS BD	0.67	GPT left	0.34
WAIS GI	0.76	GPT right	0.33
WRAT	0.73		

Variance explained was 0.42.

Descriptive statistics for primary variables

Trait	Group	N	Mean	SD	Median	Mad	Skew	Kurtosis
g	White	3555	0.00	1.00	0.05	1.05	-0.28	-0.43
g	Black	502	-1.27	0.86	-1.32	0.83	0.39	0.07
g	Hispanic	181	-0.78	0.89	-0.81	0.93	0.16	-0.50
g	Asian	34	-0.20	1.16	-0.12	1.24	-0.42	-0.75
g	Native	48	-0.41	1.05	-0.49	1.09	0.35	-0.39
predicted *g*	White	3654	0.00	1.00	0.12	1.05	-0.38	-0.36
predicted *g*	Black	525	-1.22	0.93	-1.24	0.99	0.26	-0.36
predicted *g*	Hispanic	200	-0.86	0.92	-0.86	0.94	-0.04	-0.34
predicted *g*	Asian	34	-0.42	0.87	-0.55	0.93	-0.30	-0.47
predicted *g*	Native	49	-0.64	1.05	-0.73	1.08	0.44	-0.72
P IRT	White	3654	0.00	1.00	0.00	0.99	-0.07	0.02
P IRT	Black	525	0.37	0.91	0.36	0.90	-0.09	-0.09
P IRT	Hispanic	200	0.37	1.06	0.29	1.12	0.18	-0.41
P IRT	Asian	34	0.35	1.15	0.32	1.14	-0.10	-0.52
P IRT	Native	49	0.58	1.06	0.54	0.95	-0.16	0.14

Summary of Cutler et al. (2019)

Because this conference presentation is somewhat obscure and the study is relevant to the present, we reproduce the summary of the study here for ease of reference.

Has psychometrics overlooked machine learning methods? We investigated whether machine learning methods could improve the scoring of cognitive item data. To do this, we collected 7 large datasets of item data (total n = 37k). Most datasets provided response-level data i.e. which response subject gave, not just binary (correct/incorrect). Datasets collectively had many outcomes of interest, but we focused on a small number that mostly overlapped between datasets: age, sex, educational attainment and income (all [quasi-]continuous aside from sex). We then applied standard psychometric scoring methods, sum scores and item response theory (IRT) scores, to the data as well as a variety of supervised machine learning methods including random forest, lasso/ridge regression, deep neural networks, as well as unpenalized ordinary least squares (OLS) for comparison. Parameters were tuned using efficient leave one out cross-validation on a training set. Performance was measured on 20% hold out data.

Our results indicate that machine learning methods regularly outperform standard psychometric scoring methods. Across datasets, a mean gain in validity of 47%, 17%, and 13% were seen for age, education, and income, respectively. This gain was fairly consistent across datasets and test item types, i.e. machine learning methods were able to use all tested item types to extract extra validity, including vocabulary, memory, verbal fluency, matrices/Raven's, general knowledge, and math.

Of the machine learning methods, many were roughly equivalent. Ridge regression was overall the best method. This indicates that: 1) Sparsity of effects is not a good assumption for these data, i.e. that each response option was unique in utility. In other words: the different 'distractors' (wrong options) in multiple choice questions are differentially informative, not equivalent as assumed by the binary scoring methods. 2) Interactions between items do not seem to play important roles for predictive purposes, in line with traditional psychometric results.

We conclude that standard scoring approaches of cognitive data are missing extra validity present in the data, sometimes a lot of it, depending on the outcome. This finding has implications for tests used for practical purposes (such as selection, dementia screening), where their validity has likely been underestimated.