Investigating the genetic basis for intelligence

Steve Hsu

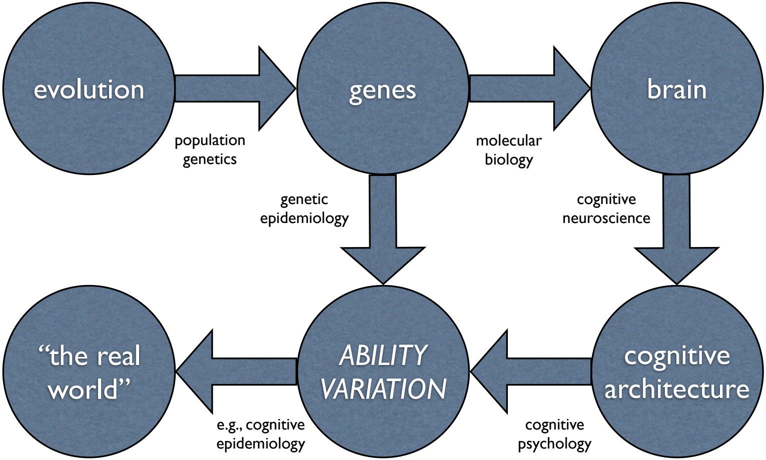
University of Oregon and BGI
www.cog-genomics.org

Outline: a multidisciplinary subject

1. What is intelligence? Psychometrics
2. g and GWAS: a project with BGI (formerly Beijing Genomics Institute)
3. The future

www.cog-genomics.org
These slides: ggoogle.pdf

Outline: a multidisciplinary subject



What is Intelligence: IQ / SAT / GRE ?

Operational perspective: who cares, as long as they have predictive power!

1. Stability / Reliability (measured value doesn't change)
2. Validity (predictive power; measures something *real*?)
3. Heritability (genetic causes)

We'll see that intelligence is comparable to height on each of these criteria.

What are IQ / SAT / GRE ?

By construction:

I. Choose a battery of n ″cognitive″ tests, e.g.,

(1) digit recall (short term memory)
(2) vocabulary
(3) math puzzles
(4) spatial rotations
. . .
(n-1) reaction time
(n) pitch recognition (music)

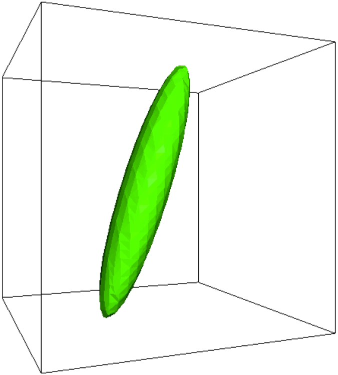
II. Test a lot of people.

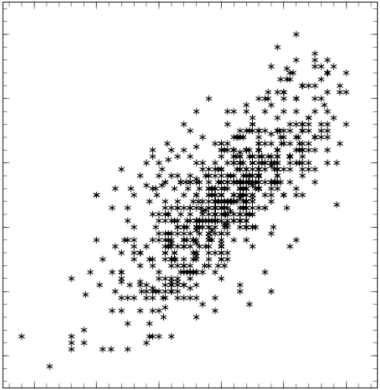
$$\{\text{individual}\} \;\rightarrow\; \text{n vector} \;\rightarrow\; \text{scalar (single number)}$$

(LOSSY) COMPRESSION!

Results

- All ”cognitive” observables seem to be *positively* correlated
- Use factor analysis or principal components to isolate direction of largest variation in the n-dimensional space





General factor of intelligence

Largest principal component of variation ≈ Dimension identified by factor analysis =

g factor = general factor of intelligence ≈ IQ ≈ SAT ≈ GRE

≈ overall goodness of cognitive functioning?

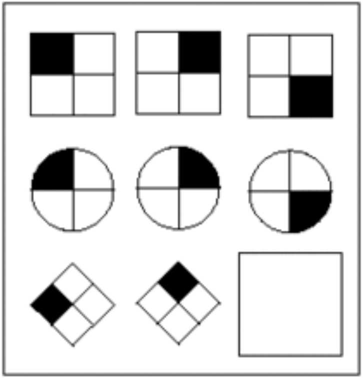
• Note these are *population level* correlations – compression may not work for a particular individual: value of g may not predict individual components of n-vector very well. But works for "typical" individuals.

• SAT, GRE heavily g-loaded: high correlation with g or IQ; "SAT is an IQ test"

IQ: mean 100, SD 15 (normally distributed)

SAT (M+V): mean 1000, SD ~ 200 (1995 "recentering")

Progressive Matrices



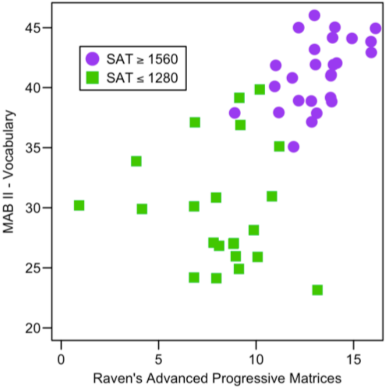
Highly g loaded but relatively culture neutral and abstract.

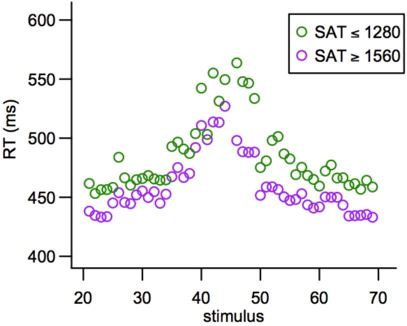
Pattern recognition and algorithmic reasoning.

Results

left fig. Vocabulary, SAT and RAPM intercorrelation.

right fig. Reaction time differences for two groups.





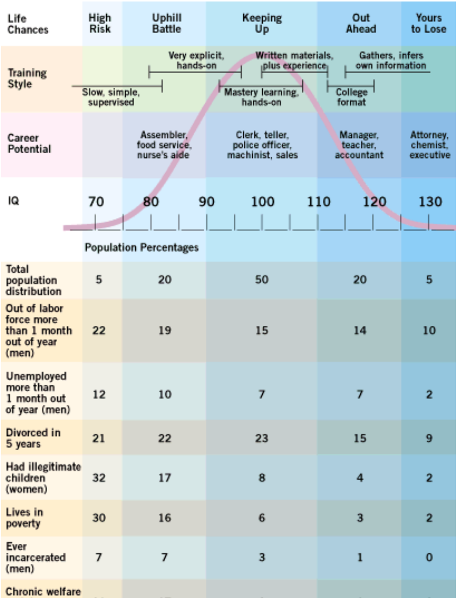
What good is it?

PRO: Among the most impressive quantitative results in all of psychology.

1. Results are stable after late adolescence (reliability). One year retest correlation .9 or higher.
2. Results are predictive (validity).
3. It's heritable (twin studies). Gulp!

CON: Only explains small fraction of variance in life outcomes.

Life outcomes



Training Style

- Slow, simple, supervised
- Very explicit, hands-on
- Mastery learning, hands-on
- Written materials, plus experience
- College format
- Gathers, infers own information

IQ: 70 80 90 100 110 120 130

Life Chances	High Risk	Uphill Battle	Keeping Up	Out Ahead	Yours to Lose
Career Potential		Assembler, food service, nurse's aide	Clerk, teller, police officer, machinist, sales	Manager, teacher, accountant	Attorney, chemist, executive
Population Percentages					
Total population distribution	5	20	50	20	5
Out of labor force more than 1 month out of year (men)	22	19	15	14	10
Unemployed more than 1 month out of year (men)	12	10	7	7	2
Divorced in 5 years	21	22	23	15	9
Had illegitimate children (women)	32	17	8	4	2
Lives in poverty	30	16	6	3	2
Ever incarcerated (men)	7	7	3	1	0
Chronic welfare					

Can further control for SES (socio-economic status) by considering sibling pairs with different IQs.

College outcomes: two factor working model

Factor 1: SAT (cognitive ability)
Factor 2: Conscientiousness, work ethic, motivation ...

These factors are only weakly correlated with each other.

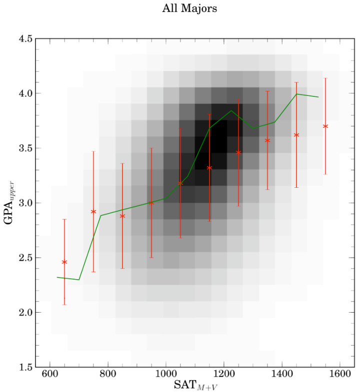
To what extent can Factor 2 compensate for Factor 1? For fixed values of SAT, what is the range of outcomes in college performance?

Are there *cognitive thresholds* for certain subjects, such that mastery is very unlikely below a certain SAT threshold (i.e., no matter how dedicated or hard working the student)?

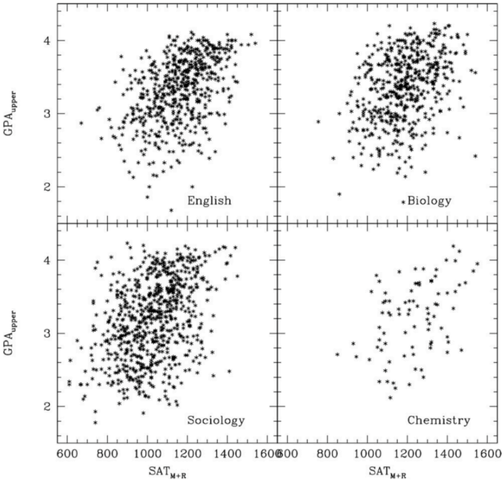
College outcomes

Data Mining the University, Hsu and Schombert, arXiv:1004.2731

Analysis of 5 years of student records at the University of Oregon.

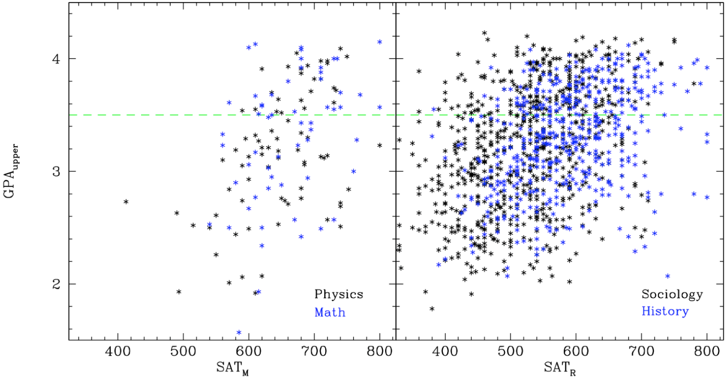


College outcomes

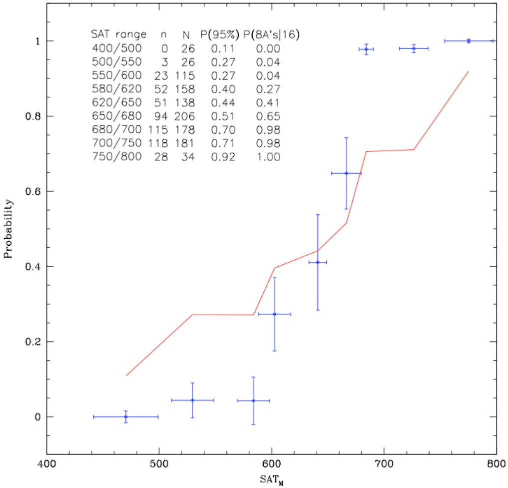


College outcomes: thresholds?

Nonlinear Psychometric Thresholds for Physics and Mathematics, Hsu and Schombert, arXiv:1011.0663



College outcomes: thresholds?



SAT range	n	N	P(95%)	P(8A's|16)
400/500	0	26	0.11	0.00
500/550	3	26	0.27	0.04
550/600	23	115	0.27	0.04
580/620	52	158	0.40	0.27
620/650	51	138	0.44	0.41
650/680	94	206	0.51	0.65
680/700	115	178	0.70	0.98
700/750	118	181	0.71	0.98
750/800	28	34	0.92	1.00

The far tail

What about the far tail?

+2 SD	130	top few percent
+3 SD	145	1 in 1000
+4 SD	160	1 in 30,000

Diminishing returns above some threshold (e.g., 120)?

OR

It's good to have a big brain ... BIGGER IS BETTER :-)

Answer: IT DEPENDS ...

The far tail

Not much data.

Roe study (1950's): 64 randomly selected eminent scientists had IQs much higher than the general population of science PhDs. Almost all of the eminent scientists in the sample scored above +(3-4) SD in at least one of M / V categories.

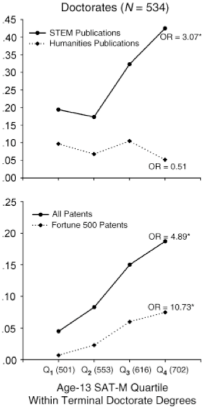
Mean score in both categories was roughly +4 SD.

Average for science PhDs around +2 SD, so eminent group highly atypical among scientists.

Positive returns to IQ > +2 SD in scientific research?

The far tail

SMPY longitudinal study. Children tested at age 13 or younger. First quartile Q1 roughly top percentile, top quartile Q4 roughly 1 in 10,000.



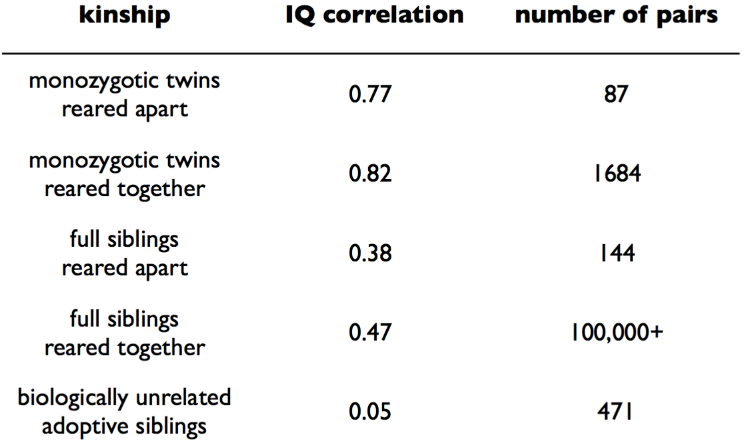
Your kids and regression

Assuming parental midpoint of n SD above the population average, the kids’ IQ will be normally distributed about a mean which is around +.6n with residual SD of about 13 points. (The .6 could actually be anywhere in the range (.5, .7), but the SD doesn’t vary much from choice of empirical inputs.)

So, e.g., for n = 4 (parental midpoint of 160 – very smart parents!), the mean for the kids would be 136 with only a few percent chance of any kid to surpass 160 (requires ~ 2 SD fluctuation). For n = 3 (parental midpoint of 145) the mean for the kids would be 127 and the probability of exceeding 145 less than 10 percent.

Heritability and Linearity

g is highly heritable and effect of individual genes is mostly linear: many genes, each of small effect. (Additive heritability about .6; broad sense heritability about .8; similar to height!)



kinship	IQ correlation	number of pairs
monozygotic twins reared apart	0.77	87
monozygotic twins reared together	0.82	1684
full siblings reared apart	0.38	144
full siblings reared together	0.47	100,000+
biologically unrelated adoptive siblings	0.05	471

Heritability and Linearity

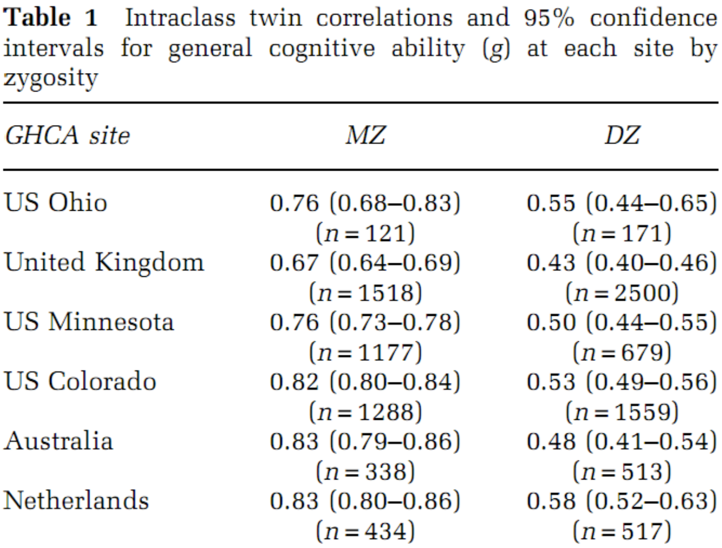


Table 1 Intraclass twin correlations and 95% confidence intervals for general cognitive ability (*g*) at each site by zygosity

GHCA site	*MZ*	*DZ*
US Ohio	0.76 (0.68–0.83) ($n = 121$)	0.55 (0.44–0.65) ($n = 171$)
United Kingdom	0.67 (0.64–0.69) ($n = 1518$)	0.43 (0.40–0.46) ($n = 2500$)
US Minnesota	0.76 (0.73–0.78) ($n = 1177$)	0.50 (0.44–0.55) ($n = 679$)
US Colorado	0.82 (0.80–0.84) ($n = 1288$)	0.53 (0.49–0.56) ($n = 1559$)
Australia	0.83 (0.79–0.86) ($n = 338$)	0.48 (0.41–0.54) ($n = 513$)
Netherlands	0.83 (0.80–0.86) ($n = 434$)	0.58 (0.52–0.63) ($n = 517$)

Linearity: many genes of small effect

1. g is normally distributed
2. genetic component is approximately linear in effect (additive heritability .6 out of .8 total)

Can think in terms of $+$ and $-$ effects from alleles.

Characterize an individual in terms of which variants they inherit at each of n sites:

$$(+++-++- \quad \cdots \quad ++--+++)$$

Coin flips with probability p_i at each site yields normal distribution as $n \to \infty$.

g and GWAS

GWAS = Genome Wide Association Study. Thus far, little success in finding genes linked to intelligence (Plomin 2010).

Candidate hits have not been successfully replicated.

Compare to the situation with height: about 200 genes found so far correlated to height. Only 5 or 10 percent of total variance associated with specific loci (“missing heritability”), but over 50 percent or more of total variance from global fit.

There is a historic opportunity to conduct the first study that finds a significant number of IQ-associated genes.

BGI: formerly Beijing Genomics Institute

Headquarters in Shenzhen, China. Raised funding of US $ 1.6 billion. Nearly 5000 employees (1000 in software development alone).

More sequencing power than any academic lab in US or Europe. Aims to become leading platform for sequencing and bioinformatics. Eventually, 1000 genomes sequenced per day at less than $ 1000 per genome.

Previous successes: participant in original Human Genome Project (1 percent), rice genome, Panda genome, Tibetan altitude adaptation, early hominid sequence, over 1000 Han genomes sequenced.

BGI

Han study: high-normal design

Seek 10^3 or more subjects with IQ +3 SD or higher (roughly 1 in 1000).

Conveniently pre-filtered population: students invited to training camps for physics, math and informatics Olympiads. Each student ranked roughly top 5 in his or her province, roughly 100 per subject per year in China.

Math ability, possibly as high as +4 SD, general intelligence probably roughly +3 SD. Randomized testing to check these estimates.

Search for volunteers among this population already under way.

Rough estimates

Simple model: N genes of equal small effect. (i.e., $N \sim 10^3$). Let + allele have slightly positive effect on IQ, and − allele have slightly negative effect.

Assume high group average is $+k$ SD, so $k \sim (3-4)$. Then difference in frequencies between high and normal groups is

$$f_+^H - f_+^N \sim \frac{k}{2\sqrt{N}}$$

How well can we measure f_+ in the two populations? Statistical fluctuations: $\frac{1}{2\sqrt{M}}$, where M is population size.

Once $M > N$, have good power to detect + alleles. (False positives: 10^3 variants of interest, 10^6 SNPs on chip; need signal to noise ratio of $> 10^3$ or so.)

Power calculations

Expected hits assuming IQ allele frequencies and effect sizes similar to height.

2000 CASES, 4000 CONTROLS

case lower threshold = 3.5 SD **total expected hits: *3.51***

	average effect					
MAF	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**
0.1	—	—	0.02	0	0.26	0
0.2	—	0.07	0.28	0	0	0
0.3	—	0.18	0.36	0	0	0
0.4	—	0.20	0	1.24	0.90	0

The Future

Expect full sequencing (not just SNP genotyping) of 10^5 to 10^6 individuals within next 5 years, paid for by science agencies of national governments. Total cost roughly US $1 billion or so ... comparable to first genome sequenced by Human Genome Project!

IF sufficient phenotype data is collected about these individuals, will have very well-powered GWAS studies within next few years – enough statistical power to capture a good fraction of total additive variance (about .6 for intelligence).

What can we do with this information?

Please help! Free genotyping!

If you are cognitively gifted, please participate in our study.
www.cog-genomics.org

Free genotyping: initially SNPs, later possibly exome or whole genome sequencing.

Learn about your ancestry and health.

Please help! Free genotyping!

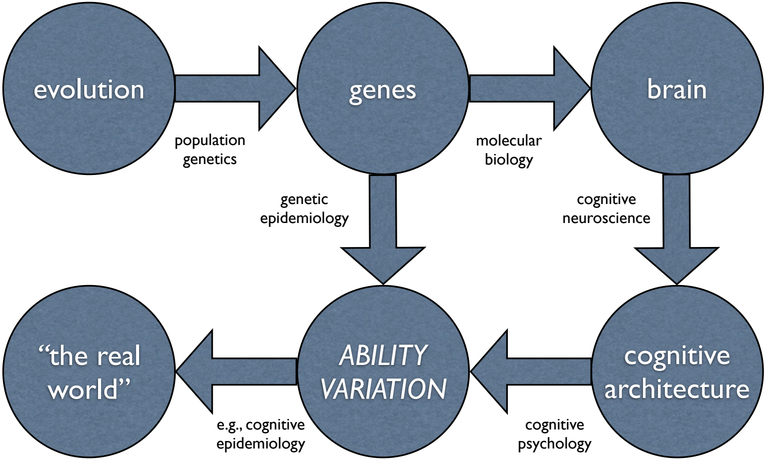
Some automatic qualifying criteria:

SAT (post 1995) M800 V760
SAT (pre 1995) M780 V700
GRE Q800 V700
ACT 35

PhD from top 5 ranked US program in physics, math, EE, theoretical computer science.

Honorable mention or better (top 50 or so in US) in Putnam competition.

Consider making a donation!



evolution

genes

brain

population genetics

molecular biology

genetic epidemiology

cognitive neuroscience

"the real world"

ABILITY VARIATION

cognitive architecture

e.g., cognitive epidemiology

cognitive psychology