Is Psychometric g a Myth?

APRIL 3, 2013 / DALLIARD / 80 COMMENTS

As an online discussion about IQ or general intelligence grows longer, the probability of someone linking to statistician Cosma Shalizi's essay *g, a Statistical Myth* approaches 1. Usually the link is accompanied by an assertion to the effect that Shalizi offers a definitive refutation of the concept of general mental ability, or psychometric *g*.

In this post, I will show that Shalizi's case against *g* appears strong only because he misstates several key facts and because he omits all the best evidence that the other side has offered in support of *g*. His case hinges on three clearly erroneous arguments on which I will concentrate.

Contents

I. Positive manifold

Shalizi writes that when all tests in a test battery are positively correlated with each other, factor analysis will necessarily yield a general factor. He is correct about this. All subtests of any given IQ battery are positively correlated, and subjecting an IQ correlation matrix to factor analysis will produce a first factor on which all subtests are positively loaded. For example, the 29 subtests of the revised 1989 edition of the Woodcock-Johnson IQ test are correlated in the following manner (click for larger image):

Table 1.4: Pearsonian intercorrelation matrix, combined kindergarten to adult sample (decimals omitted). 29 variables from the Woodcock-Johnson psycho-educational battery — revised, $N = 1425$ (correlations corrected for age).

Variable:		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Memory for Names	1	1000																												
Memory for Sentences	2	279	1000																											
Visual Matching	3	213	254	1000																										
Incomplete Words	4	167	255	191	1000																									
Visual Closure	5	148	103	178	176	1000																								
Picture Vocabulary	6	404	403	202	267	229	1000																							
Analysis-Synthesis	7	275	324	280	205	161	323	1000																						
Visual-Auditory Learning	8	542	343	267	192	205	382	376	1000																					
Memory for Words	9	208	559	221	245	046	225	215	246	1000																				
Cross Out	10	170	241	621	168	241	242	291	265	203	1000																			
Sound Blending	11	245	323	245	367	133	323	265	332	335	246	1000																		
Picture Recognition	12	293	216	212	123	234	256	233	299	155	257	212	1000																	
Oral Vocabulary	13	388	534	310	319	234	632	419	405	364	315	389	304	1000																
Concept Formation	14	306	382	306	236	206	325	484	376	227	305	275	269	458	1000															
Memory for Names (Delayed Recall)	15	721	236	155	168	129	383	269	460	173	123	242	236	359	284	1000														
Visual-Auditory Learning (Delayed Recall)	16	345	164	162	120	192	255	269	460	110	168	192	275	271	323	446	1000													
Numbers Reversed	17	259	416	384	227	129	255	368	321	401	309	316	206	396	354	225	182	1000												
Sound Patterns	18	233	257	204	221	109	269	271	259	243	229	294	168	331	299	222	214	282	1000											
Spatial Relations	19	280	266	278	158	265	317	389	369	189	343	225	288	388	404	240	289	311	294	1000										
Listening Comprehension	20	331	469	266	334	204	576	349	344	279	263	351	256	642	375	294	221	308	274	320	1000									
Verbal Analogies	21	379	454	334	228	242	522	455	445	310	344	355	322	639	496	377	330	403	304	465	526	1000								
Calculation	22	256	331	435	142	132	299	423	347	252	358	293	208	471	401	249	242	413	257	376	374	483	1000							
Applied Problems	23	337	416	419	206	175	439	470	388	312	388	360	273	603	489	313	268	438	315	486	524	631	655	1000						
Science	24	380	437	260	285	233	633	368	364	246	280	323	246	658	389	362	270	336	260	385	619	544	440	570	1000					
Social Studies	25	371	477	298	262	200	626	386	374	270	278	323	255	693	411	348	245	332	256	344	638	595	508	617	702	1000				
Humanities	26	390	447	308	281	252	622	343	414	297	284	355	283	665	359	368	283	326	282	340	572	598	427	536	633	672	1000			
Word Attack	27	281	370	356	263	119	316	303	366	322	255	484	202	468	329	269	228	398	316	312	326	415	422	450	346	354	398	1000		
Quantitative Concepts	28	342	427	408	205	162	497	437	416	309	361	320	244	615	413	337	280	433	299	440	513	624	656	728	602	637	576	471	1000	
Writing Fluency	29	225	350	494	193	123	260	309	347	266	410	358	196	398	335	197	194	365	229	276	285	394	420	426	293	336	409	488	434	1000

All the subtest intercorrelations are positive, ranging from a low of 0.046 (Memory for Words – Visual Closure) to a high of 0.728 (Quantitative Concepts – Applied Problems). (See Woodcock 1990 for a description of the tests.) This is the reason why we talk about general intelligence or general cognitive ability: individuals who get a high score on one cognitive test tend to do so on all kinds of tests regardless of test content or type (e.g., verbal, numerical, spatial, or memory tests), while those who do bad on one type of cognitive test usually do bad on all tests.

This phenomenon of positive correlations among all tests, often called the "positive manifold", is routinely found among all collections of cognitive ability tests, and it is one of the most replicated findings in the social and behavioral sciences. The correlation between a given pair of ability tests is a function of the shared common factor variance (*g* and other factors) and imperfect test reliabilities (the higher the reliabilities, the higher the correlation). All cognitive tests load on *g* to a smaller or greater degree, so all tests covary at least through the *g* factor, if not other factors.

John B. Carroll factor-analyzed the WJ-R matrix presented above, using confirmatory analysis to successfully fit a ten-factor model (*g* and nine narrower factors) to the data (Carroll 2003):

Table 1.5: LISREL estimates of orthogonal factor loadings for 29 variables on 10 factors (decimals omitted).

Stratum:	3	2	2	2	2	2	2	2	2	2	
Factor:	*g*	Glr	Gsm	Gs	Ga	Gv	Gc	Gf	Gq	Lang	h^2
Factor No.:	1	2	3	4	5	6	7	8	9	10	
01 MEMNAM	**478**	**695**	—	—	—	—	—	—	—	—	712
02 MEMSEN	**587**	—	**396**	—	—	—	—	—	—	—	501
03 VISMAT	**499**	—	—	**709**	—	—	—	—	—	—	752
04 INCWDS	**340**	—	—	—	**308**	—	—	—	—	—	210
05 VISCLO	**279**	—	—	—	—	**472**	—	—	—	—	301
06 PICVOC	**566**	—	—	—	—	—	**531**	—	—	—	602
07 ANLSYN	**591**	—	—	—	—	—	—	**213**	—	—	395
08 VISAUD	**579**	**343**	—	—	—	—	—	—	—	—	453
09 MEMWDS	**424**	—	**782**	—	—	—	—	—	—	—	791
10 CRSOUT	**478**	—	—	**539**	—	—	—	—	—	—	519
11 SNDBND	**490**	—	—	—	**642**	—	—	—	—	—	652
12 PICREC	**398**	—	—	—	—	**260**	—	—	—	—	226
13 ORALVO	**749**	—	—	—	—	—	**377**	—	—	—	703
14 CNCPTF	**623**	—	—	—	—	—	—	**543**	—	—	683
15 MMNADR	**439**	**729**	—	—	—	—	—	—	—	—	724
16 VSAUDR	**404**	**320**	—	—	—	—	—	—	—	—	266
17 NMRVRS	**571**	—	**203**	—	—	—	—	—	—	—	367
18 SNDPAT	**436**	—	—	—	**144**	—	—	—	—	—	211
19 SPAREL	**580**	—	—	—	—	**219**	—	—	—	—	384
20 LISCMP	**619**	—	—	—	—	—	**424**	—	—	—	563
21 VBLANL	**761**	—	—	—	—	—	**162**	**052**	—	—	608
22 CALCUL	**652**	—	—	—	—	—	—	—	**432**	—	612
23 APLPRB	**783**	—	—	—	—	—	—	—	**335**	—	725
24 SCIENC	**651**	—	—	—	—	—	**491**	—	—	—	665
25 SOCSTU	**686**	—	—	—	—	—	**488**	—	—	—	709
26 HUMANI	**661**	—	—	—	—	—	**448**	—	—	**107**	649
27 WDATCK	**587**	—	—	—	**273**	—	—	—	—	**197**	458
28 QUANCN	**743**	—	—	—	—	—	**177**	—	**400**	—	743
29 WRIFLU	**549**	—	—	**286**	—	—	—	—	—	**685**	852
SMSQ	9515	1235	810	875	602	338	1341	343	459	519	16037
%CCV	59.33	7.70	5.05	5.45	3.75	2.10	8.36	2.13	2.23	3.23	100.00

Measures of goodness of fit for the whole model:
CHI-square with 343 degrees of freedom = 1488.60 ($p = 0.000$)
Goodness of fit index = 0.931; Adjusted goodness of fit index = 0.912
Root mean square residual = 0.039

Note: Analysis of the correlation matrix of Table 1.4, which see for full names of variables. Factor Names (as given by McGrew *et al.*, 1991): *g*: General Intellectual Ability; Glr: Long-Term Retrieval; Gsm: Short-Term Memory; Gs: Processing Speed; Ga: Auditory Processing; Gv: Visual-Spatial Thinking; Gc: Comprehension-Knowledge; Gf: Fluid Reasoning: Gq: Mathematics; Lang: Language. h^2: Communality or Squared Multiple Correlation; SMSQ: Sums of Squares; %CCV: Percentages of Common Factor Covariance.

Loadings on the *g* factor range from a low of 0.279 (Visual Closure) to a high of 0.783 (Applied Problems). The *g* factor accounts for 59 percent of the common factor variance, while the other nine factors together account for 41 percent. This is a routine finding in factor analyses of IQ tests: the *g* factor explains more variance than the other factors put together. (Note that in addition to the common factor variance, there is always some variance specific to each subtest as well as variance due to random measurement error.)

II. Shalizi's first error

Against the backdrop of results like the above, Shalizi makes the following claims:

> The correlations among the components in an intelligence test, and between tests themselves, are all positive, because that's how we design tests. *[…]* So making up tests so that they're positively correlated and discovering they have a dominant factor is just like putting together a list of big square numbers and discovering that none of them is prime — it's necessary side-effect of the construction, nothing more.

[…]

> What psychologists sometimes call the "positive manifold" condition is enough, in and of itself, to guarantee that there will appear to be a general factor. Since intelligence tests are made to correlate with each other, it follows trivially that there must appear to be a general factor of intelligence. This is true whether or not there really is a single variable which explains test scores or not.

[…]

> By this point, I'd guess it's impossible for something to become accepted as an "intelligence test" if it doesn't correlate well with the Weschler [sic] and its kin, no matter how much intelligence, in the ordinary sense, it requires, but, as we saw with the first simulated factor analysis example, that makes it inevitable that the leading factor fits well.

Shalizi's thesis is that the positive manifold is an artifact of test construction and that full-scale scores from different IQ batteries correlate only because they are designed to do that. It follows from this argument that if a test maker decided to disregard the *g* factor and construct a battery for assessing several independent abilities, the result would be a test with many zero or negative correlations among its subtests. Moreover, such a test would not correlate highly with traditional tests, at least not positively. Shalizi alleges that there are tests that measure intelligence "in the ordinary sense" yet are uncorrelated with traditional tests, but unfortunately he does not gives any examples.

Inadvertent positive manifolds

There are in fact many cognitive test batteries designed without regard to *g*, so we can put Shalizi's allegations to test. The Woodcock-Johnson test discussed above is a case in point. Carroll, when reanalyzing data from the test's standardization sample, pointed out that its technical manual "reveals a studious neglect of the role of any kind of general factor in the WJ-R." This dismissive stance towards *g* is also reflected in Richard Woodcock's article about the test's theoretical background (Woodcock 1990). (Yes, the Woodcock-Johnson test was developed by a guy named Dick Woodcock, together with his assistant Johnson. You can't make this up.) The WJ-R was developed based on the idea that the *g* factor is a statistical artifact with no psychological relevance. Nevertheless, all of its subtests are intercorrelated and, when factor analyzed, it reveals a general factor that is no less prominent than those of more traditional IQ tests. According to the WJ-R technical manual, test results are to be interpreted at the level of nine broad abilities (such as Visual Processing and Quantitative Ability), not any general ability. Similarly, the manual reports factor analyses based only on the nine factors. But when Carroll reanalyzed the data, allowing for loadings on a higher-order *g* factor in addition to the nine factors, it turned out that most of the tests in the WJ-R have their highest loadings on the *g* factor, not on the less general ("broad") factors which they were specifically designed to measure.

While the WJ-R is not meant to be a test of *g*, it does provide a measure of "broad cognitive ability", which correlates at 0.65 and 0.64 with the Stanford-Binet and Wechsler full-scale scores, respectively (Kamphaus 2005, p. 335). Typically, correlations between full-scale scores from different IQ tests are around 0.8. ~~The WJ-R broad cognitive ability scores are probably less *g*-loaded than those of other tests, because they are based on unweighted sums of scores on subtests selected solely on the basis of their content diversity; hence the lower correlations, I believe.~~ The lower than expected correlation appears to be due to range restriction in the sample used. In any case, the WJ-R is certainly not uncorrelated with traditional tests. (The WJ-III, which is the newest edition of the test, now recognizes the *g* factor.)

The WJ-R serves as a forthright refutation of Shalizi's claim that the positive manifold and inter-battery correlations emerge by design rather than because all cognitive abilities naturally intercorrelate. But perhaps the WJ-R is just a giant fluke, or perhaps its 29 tests correlate as a carryover from the previous edition of the test which had several of the same tests but was not based on anti-*g* ideas. Are there other examples of psychometricians accidentally creating strongly *g*-loaded tests against their best intentions? In fact, there is a long history of such inadvertent confirmations of the ubiquity of the *g* factor. This goes back at least to the 1930s and Louis Thurstone's research on "primary mental abilities".

Thurstone and Guilford

In a famous study published in 1938, Thurstone, one of the great psychometricians, claimed to have developed a test of seven independent mental abilities (verbal comprehension, word fluency, number facility, spatial visualization, associative memory, perceptual speed, and reasoning; see Thurstone 1938). However, the *g* men quickly responded, with Charles Spearman and Hans Eysenck publishing papers (Spearman 1939, Eysenck 1939) showing that Thurstone's independent abilities were not independent, indicating that his data were compatible with Spearman's *g* model. (Later in his career, Thurstone came to accept that perhaps intelligence could best be conceptualized as a hierarchy topped by *g*.)

The idea of non-correlated abilities was taken to its extreme by J.P. Guilford who postulated that there are as many as 160 different cognitive abilities. This made him very popular among educationalists because his theory suggested that everybody could be intelligent in some way. Guilford's belief in a highly multidimensional intelligence was influenced by his large-scale studies of Southern California university students whose abilities were indeed not always correlated. In 1964, he reported (Guilford 1964) that his research showed that up to a fourth of correlations between diverse intelligence tests were not different from zero. However, this conclusion was based on bad psychometrics. Alliger 1988 reanalyzed Guilford's data and showed that when you correct for artifacts such as range restriction (the subjects were generally university students), the reported correlations are uniformly positive.

British Ability Scales

Psychometricians have not been discouraged by past failures to discover abilities that are independent of the general factor. They keep constructing tests that supposedly take the measurement of intelligence beyond *g*.

For example, the British Ability Scales was carefully developed in the 1970s and 1980s to measure a wide variety of cognitive abilities, but when the published battery was analyzed (Elliott 1986), the results were quite disappointing:

> Considering the relatively large size of the test battery *[…]* the solutions have yielded perhaps a surprisingly small number of common factors. As would be expected from any cognitive test battery, there is a substantial general factor. After that, there does not seem to be much common variance left *[…]*
>
> What, then, are we to make of the results of these analyses? Do they mean that we are back to square one, as it were, and that after 60 years of research we have turned full circle and are back with the theories of Spearman? Certainly, for this sample and range of cognitive measures, there is little evidence that strong primary factors, such as those postulated by many test theorists over the years, have accounted for any substantial proportion of the common variance of the British Ability Scales. This is despite the fact that the scales sample a wide range of psychological functions, and deliberately include tests with purely verbal and purely visual tasks, tests of fluid and crystallized mental abilities, tests of scholastic attainment, tests of complex mental functioning such as in the reasoning scales and tests of lower order abilities as in the Recall of Digits scale.

CAS

An even better example is the CAS battery. It is based on the PASS theory (which draws heavily on the ideas of Soviet psychologist A.R. Luria, a favorite of Shalizi's), which disavows *g* and asserts that intelligence consists of four processes called Planning, Attention-Arousal, Simultaneous, and Successive. The CAS was designed to assess these four processes.

However, Keith el al. 2001 did a joint confirmatory factor analysis of the CAS together with the WJ-III battery, concluding that not only does the CAS not measure the constructs it was designed to measure, but that notwithstanding the test makers' aversion to *g*, the *g* factor derived from the CAS is large and statistically indistinguishable from the *g* factor of the WJ-III. The CAS therefore appears to be the opposite of what it was supposed to be: an excellent test of the "non-existent" *g* and a poor test of the supposedly real non-*g* abilities it was painstakingly designed to measure.

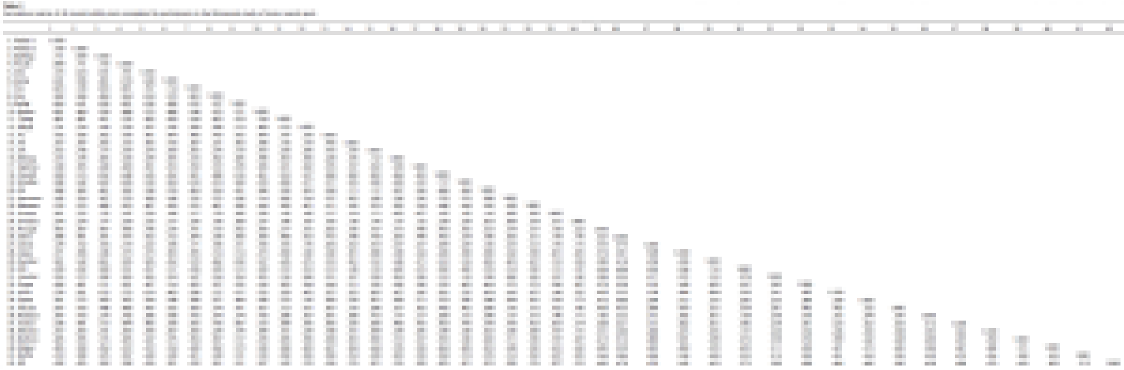
Triarchic intelligence

A particularly amusing confirmation of the positive manifold resulted from Robert Sternberg's attempts at developing measures of non-*g* abilities. Sternberg introduced his "triarchic" theory of intelligence in the 1980s and has tirelessly promoted it ever since while at every turn denigrating the proponents of *g* as troglodytes. He claims that *g* represents a rather narrow domain of analytic or academic intelligence which is more or less uncorrelated with the often much more important creative and practical forms of intelligence. He created a test battery to test these different intellectual domains. It turned out that the three "independent" abilities were highly intercorrelated, which Sternberg absurdly put down to common-method variance.

A reanalysis of Sternberg's data by Nathan Brody (Brody 2003a) showed that not only were the three abilities highly correlated with each other and with Raven's IQ test, but also that the abilities did not exhibit the postulated differential validities (e.g., measures of creative ability and analytic ability were equally good predictors of measures of creativity, and analytic ability was a better predictor of practical outcomes than practical ability), and in general the test had little predictive validity independently of *g*. (Sternberg, true to his style, refused to admit that these results had any implications for the validity of his triarchic theory, prompting the exasperated Brody to publish an acerbic reply called "What Sternberg should have concluded" [Brody 2003b].)

MISTRA

The administration of several different IQ batteries to the same sample of individuals offers another good way to test the generality of the positive manifold. As part of the Minnesota Study of Twins Reared Apart (or MISTRA), three batteries comprising a total of 42 different cognitive tests were taken by the twins studied and also by many of their family members. The three tests were the Comprehensive Ability Battery, the Hawaii Battery, and the Wechsler Adult Intelligence Scale. The tests are highly varied content-wise, with each battery measuring diverse aspects of intelligence. See Johnson & Bouchard 2011 for a description of the tests. Correlations between the 42 tests are presented below (click for larger image):



All 861 correlations are positive. Subtests of each IQ battery correlate positively not only with each other but also with the subtests of the other IQ batteries. This is, of course, something that the developers of the three different batteries could not have planned – and even if they could have, they would not have had any reason to do so, given their different theoretical presuppositions. (Later in this post, I will present some very interesting results from a factor analysis of these data.)

Piagetian tasks

As a final example of the impossibility of doing away with the positive manifold I will discuss a test battery which is rather exotic from a traditional psychometric perspective. The Swiss developmental psychologist Jean Piaget devised a number of cognitive tasks in order to investigate the developmental stages of children. He was not interested in individual differences (a common failing among developmental psychologists) but rather wanted to understand universal human developmental patterns. He never created standardized batteries of his tasks. See here for a description of many of Piaget's tests. Some of them, such as those assessing Logical Operations are quite similar to traditional IQ items, but others, such as Conservation tasks, are unlike anything in IQ tests. Nevertheless, most would agree that all of them measure cognitive abilities.

Humphreys et al. 1985 studied a battery of 27 Piagetian tasks completed by a sample of 150 children. A factor analysis of the Piagetian battery showed that a strong general factor underlies the tasks, with loadings ranging from 0.32 to 0.80:

Table 1
Correlations With Piagetian Composites of the Individual Tasks and Comparison With the General and Group Factor Loadings

Item	27 tasks	22 tasks	13 tasks	General	Factor 2	Factor 3	Factor 4
Conservation of substance	826	845	865	76	44	07	01
One-for-one exchange	697	722	734	66	42	–04	–05
Dissolution (weight)	592	578		51	24	09	21
Dissolution (substance)	701	716	699	65	31	13	03
Dissolution (volume)	369			36	02	–12	18
Conservation of weight	736	757	788	66	50	09	–03
Term-to-term correspondence	744	761	764	67	46	05	00
Class inclusion (animals 3)	554	547		52	19	50	–10
Class inclusion (animals 4)	685	668		66	13	52	–03
Class inclusion (animals 5a)	305			32	–14	39	03
Class inclusion (animals 5b)	408			42	–06	43	06
Conservation of volume 1	716	720	736	64	40	–06	15
Conservation of volume 2	454			43	–11	06	40
Rotation of beads	649	662		63	24	–05	06
Conservation of length	714	732	734	66	34	08	04
Conservation of length (rods)	775	798	838	67	57	05	01
Changing criterion	700	697	693	66	24	03	08
Conservation of liquid	759	783	812	68	54	–02	–01
Class inclusion (beads)	735	733	722	70	24	11	05
Disassociation (weight & volume)	746	708	658	70	04	08	34
Intersection of classes	405			40	–02	09	12
Rotation of squares (1)	630	639		60	14	–09	25
Rotation of squares (2)	506	504		48	10	–06	22
Two–three dimensions	608	614		56	24	–09	15
Changing perspectives (mobile)	688	662		69	–06	11	21
Changing perspectives (fixed)	827	816	771	80	12	08	18
Chemistry	671	660		66	08	05	16

Note. Decimals have been omitted.

But it is possible that the Piagetian general factor is not at all the same as the general factor of IQ batteries or achievement tests. Whether this is the case was tested by having the same sample take Wechsler's IQ test and an achievement test of spelling, arithmetic, and reading. The result was that scores on the Piagetian battery, Wechsler's Performance ("fluid") and Verbal ("crystallized") scales, and the achievement test were highly correlated, clearly indicating that they are measuring the same general factor. (A small caveat here is that the study included an oversample of mildly mentally retarded children in addition to normal children. Such range enhancement tends to inflate correlations between tests, so in a more adequate sample the correlations and *g*loadings would be somewhat lower. On the other hand, the data have not been corrected for measurement error which reduces correlations.) The correlations looked like this:

Table 2
Intercorrelations of Three Piagetian Tests, Wechsler Verbal and Performance IQs, and the Academic Achievement Composite

Test	2	3	4	5	6
1. Piaget–27 items	—[a]	—[a]	800	825	754
2. Piaget–22 items		—[a]	795	825	739
3. Piaget–13 items			763	798	719
4. Verbal IQ				805[b]	840
5. Performance IQ					792
6. Achievement composite					—

[a] Spurious part–whole correlation is not reported.
[b] Humphreys (1980) reported a correlation of .793 between these measures, which was the correlation in the data of Stephens et al. (1972). The difference represents the effect of controlling chronological age statistically as compared to controlling it by means of age norms.

When this correlation matrix of four different measures of general ability is factor analyzed, it can be seen that all of them load very strongly (~0.9) on a single factor:

Table 3
Unrotated and Rotated Factor Loadings From the Intercorrelations in Table 2

	Unrotated			Rotated		
Test	1	2	h^2	General	1	2
Piaget (27-item)	894	−192	836	873	273	005
Verbal IQ	913	126	849	890	054	230
Performance IQ	906	−122	836	884	226	056
Academic Achievement	896	209	846	874	007	285

Note. Decimals have been omitted.

It can be said that a battery of Piagetian tasks is about as good a measure of *g* as Wechsler's test. It does not matter at all that Piagetian and psychometric ideas of intelligence are very different and that the research traditions in which IQ tests and Piagetian tasks were conceived have nothing to do with each other. The *g* factor will emerge regardless of the type of cognitive abilities called for by a test.

Positive manifold as a fact of nature

These examples show that, contrary to Shalizi's claims, all cognitive abilities are intercorrelated. We can be confident about this because the best evidence for it comes not from the proponents of *g* but from numerous competent researchers who were hell-bent on disproving the generality of the positive manifold, only to be refuted by their own work.

Quite contrary to what Shalizi believes, IQ tests are usually constructed to measure several different abilities, not infrequently with the (stubbornly unrealized) objective of measuring abilities that are completely independent of *g*. IQ tests are not devised with the aim of maximizing variance on the first common factor, or *g*; rather, the prominence of the *g* factor is a fact of human nature, and it is impossible to do away with it.

The *g* factor is thus not an artifact of test construction but a genuine *explanandum*, something that any theory of intelligence must account for. The only way to deny this is to redefine intelligence to include skills and talents with little intellectual content. For example, Howard Gardner claims that there is a "bodily-kinesthetic intelligence" which athletes and dancers have plenty of. I don't think such semantic obfuscation contributes anything to the study of intelligence.

III. Shalizi's second error

Towards the end of his piece, Shalizi makes this bizarre claim:

> It is still *conceivable* that those positive correlations are all caused by a general factor of intelligence, but we ought to be long since past the point where supporters of that view were advancing arguments on the basis of evidence *other than* those correlations. So far as I can tell, however, nobody has presented a case for *g* apart from thoroughly invalid arguments from factor analysis; that is, the myth.

One can only conclude that if Shalizi really believes that, he has made no attempt whatsoever to familiarize himself with the arguments of *g* proponents, preferring his own straw man version of *g* theory instead. For example, in 1998 the principal modern *g* theorist, Arthur Jensen, published a book (Jensen 1998) running to nearly 700 pages, most of which consists of arguments and evidence that substantiate the scientific validity and relevance of the *g* factor beyond the mere fact of the positive manifold (which in itself is not a trivial finding, contra Shalizi). The evidence he puts forth encompasses genetics, neurophysiology, mental chronometry, and practical validity, among many other things.

I will next describe some of the most important findings that support the existence of *g* as the central, genetically rooted source of individual differences in cognitive abilities. Together, the different lines of evidence indicate that human behavioral differences cannot be properly understood without reference to *g*.

Evidence from confirmatory factor analyses

Shalizi spends much time castigating intelligence researchers for their reliance on exploratory factor analysis even though more powerful, confirmatory methods are available. This is a curious criticism in light of the fact that confirmatory factor analysis (CFA) was invented for the very purpose of studying the structure of intelligence. The trailblazer was the Swedish statistician Karl Jöreskog who was working at the Educational Testing Service when he wrote his first papers on the topic. There are in fact a large number of published CFAs of IQ tests, some of them discussed above. Shalizi must know this because he refers to John B. Carroll's contribution in the book *Intelligence, Genes, and Success: Scientists Respond to The Bell Curve*. In his article, Carroll discusses classic CFA studies of *g* (e.g., Gustafsson 1984) and reports CFAs of his own which indicate that his three-stratum model (which posits that cognitive abilities constitute a hierarchy topped by the *g* factor) shows good fit to various data sets (Carroll 1995).

Among the many CFA studies showing that *g*-based factor models fit IQ test data well, two published by Wendy Johnson and colleagues are particularly interesting. In Johnson et al. 2004, the MISTRA correlation matrix of three different IQ batteries, discussed above, was analyzed, and it turned out that the *g* factors computed from the three tests were statistically indistinguishable from one another, despite the fact that the tests clearly tapped into partly different sets of abilities. The results of Johnson et al. 2004, which have since been replicated in an another multiple-battery sample (Johnson et al. 2008) are in accord with Spearman and Jensen's argument that any diverse collection of cognitive tests will provide an excellent measure of one and the same *g*; what specific abilities are assessed is not important because they all measure the same *g*. In contrast, these results are not at all what one would have expected based on the theory of intelligence that Shalizi advocates. According to Shalizi's model, *g* factors reflect only the average or sum of the particular abilities called for by a given test battery, with batteries comprising different tests therefore almost always yielding different *g* factors. (I have more to say about Shalizi's preferred theory later in this post.) The omission of Johnson et al.

2004 and other CFA studies of intelligence (such the joint CFA of the PASS and WJ-III tests discussed earlier) from Shalizi's sources is a conspicuous failing.

Behavioral genetic evidence

It has been established beyond any dispute that cognitive abilities are heritable. (Shalizi has some quite wrong ideas on this topic, too, but I will not discuss them in this post.) What is interesting is that the degree of heritability of a given ability test depends on its *g* loading: the higher the *g* loading, the higher the heritability. A meta-analysis of the correlations between *g* loadings and heritabilities even suggested that the true correlation is 1.0, i.e., *g* loadings appear to represent a pure index of the extent of genetic influence on cognitive variation (see Rushton & Jensen 2010).

Moreover, quantitative genetic analyses indicate that *g* is an even stronger genetic variable than it is a phenotypic variable. I quote from Plomin & Spinath 2004:

> Multivariate genetic analysis yields a statistic called genetic correlation, which is an index of the extent to which genetic effects on one trait correlate with genetic effects on another trait independent of the heritability of the two traits. That is, two traits could be highly heritable but the genetic correlation between them could be zero. Conversely, two traits could be only modestly heritable but the genetic correlation between them could be 1.0, indicating that even though genetic effects are not strong (because heritability is modest) the same genetic effects are involved in both traits. In the case of specific cognitive abilities that are moderately heritable, multivariate genetic analyses have consistently found that genetic correlations are very high—close to 1.0 (Petrill 1997). That is, although Spearman's *g* is a phenotypic construct, *g* is even stronger genetically. These multivariate genetic results predict that when genes are found that are associated with one cognitive ability, such as spatial ability, they will also be associated just as strongly with other cognitive abilities, such as verbal ability or memory. Conversely, attempts to find genes for specific cognitive abilities independent of general cognitive ability are unlikely to succeed because what is in common among cognitive abilities is largely genetic and what is independent is largely environmental.

Thus behavior genetic findings support the existence of *g* as a genetically rooted dimension of human differences.

Practical validity

The *sine qua non* of IQ tests is that they reveal and predict current and future real-world capabilities. IQ is the best single predictor of academic and job performance and attainment, and one of the best predictors of a plethora of other outcomes, from income, welfare dependency, and criminality (Gottfredson 1997) to health and mortality and scientific and literary creativity (Robertson et al. 2010), and any number of other things, including even investing success (Grinblatt et al. 2011). If you had to predict the life outcomes of a teenager based on only one fact about them, nothing would be nearly as informative as their IQ.

One interesting thing about the predictive validity of a cognitive test is that it is directly related to the test's *g* loading. The higher the *g* loading, the better the validity. In fact, although the *g* factor generally accounts for less than half of all the variance in a given IQ battery, a lot of research indicates that it accounts for almost all of the predictive validity. The best evidence here are from several large-scale studies of US Air Force personnel. These studies contrasted *g* and a number of more specific abilities as predictors of performance in Air Force training (Ree, & Earles 1991) and jobs (Ree et al. 1994). The results indicated that *g* is the best predictor of training and job performance across all specialties, and that specific ability tests tailored for each specialty provide little or no incremental validity over *g*. Thus if you wanted to predict someone's performance in training or a job, it would be much more useful for you to get their general mental ability score rather than scores on any specific ability tests that are closely matched to the task at hand. This appears to be true in all jobs (Schmidt & Hunter 1998, 2004), although specific ability scores may provide substantial incremental validity in the case of high-IQ individuals (Robertson et al. 2010), which is in accord with Charles Spearman's view that abilities become more differentiated at higher levels of *g*. (This is why it makes sense for selective colleges to use admission tests that assess different abilities.)

For more evidence of how general the predictive validity of *g* is, we can look at the validity of *g* as a predictor of performance in GCSEs, which are academic qualifications awarded in different school subjects at age 14 to 16 in the United Kingdom. Deary et al. 2007 conducted a prospective study with a very large sample where *g* was measured at age 11 and GCSEs were obtained about five years later. The *g* scores correlated positively and substantially with the results of all 25 GCSEs, explaining (to give some examples) about 59 percent of individual differences in math, about 40 to 50 percent in English and foreign languages, and, at the low end, about 18 percent in Art and Design. In contrast, verbal ability, independently of *g*, explained an average of only 2.2 percent (range 0.0-7.2%) of the results in the 25 exams.

Arthur Jensen referred to *g* as the "active ingredient" of IQ tests, because *g* accounts for most if not all of the predictive validity of IQ even though most variance in IQ tests is not *g* variance. From the perspective of predictive validity, non-*g* variance seems to be generally just noise. In other words, if you statistically remove *g* variance from IQ test results, what is left is almost useless for the purposes of predicting behavior (except among high-IQ individuals, as noted above). This is a very surprising finding if you think, like Shalizi, that different mental abilities are actually independent, and *g* is just an uninteresting statistical artifact caused by an occasional recruitment of many uncorrelated abilities for the same task (more on this view of Shalizi's below).

Hollowness of IQ training effects

Another interesting fact about *g* is that there is there is a systematic relation between *g* loadings and practice effects in IQ tests. A meta-analysis of re-testing effects on IQ scores showed that there is a perfect negative correlation between score gains and *g* loadings of tests (te Nijenhuis et al. 2007). It appears that specific abilities are trainable but *g* is generally not (see also Arendasy & Sommer 2013). Similarly, a recent meta-analysis of the effects of working memory training on intelligence showed, in line with many earlier reviews, that cognitive training produces short-term gains in the specific abilities trained, but no "far transfer" to any other abilities (Melby-Lervåg, & Hulme 2013). Jensen called such gains *hollow* because they do not seem to represent actual improvements in real-world intellectual performance. These findings are consistent with the view that *g* is a "central processing unit" that cannot be defined in terms of specific abilities and is not affected by changes in those abilities.

Neurobiology

Chabris 2007 pointed out that findings in neurobiology "establish a biological basis for *g* that is firmer than that of any other human psychological trait". This is a far cry from Shalizi's claim that nothing has been done to investigate *g* beyond the fact of positive correlations between tests. There are a number of well-replicated, small to moderate correlations between *g* and features of brain physiology, including brain size, the volumes of white and grey matter, and nerve conduction velocity (ibid.; Deary et al. 2010). Currently, we do not have a well-validated model of "neuro-*g*", but certainly the findings so far are consistent with a central role for *g* in intelligence.

IV. Shalizi's third error

Besides his misconception that the positive manifold is an artifact of test construction and his disregard for evidence showing that *g* in a central variable in human affairs, there is a third reason why Shalizi believes that the *g* factor is a "myth". It is his conviction that correlations between cognitive tests are best explained in terms of the so-called sampling model. This model holds that there are a large number of uncorrelated abilities (or other "neural elements") and that correlations between tests emerge because all tests measure many different abilities at the same time, with some subset of the abilities being common to all tests in a given battery. Thus, according to Shalizi, there is no general factor of intelligence, but only the appearance of one due to each test tapping into some of the same abilities. Moreover, Shalizi's model suggests that *g* factors from different batteries are dissimilar, reflecting only the particular abilities sampled by each battery. The sampling model is illustrated in the following figure (from Jensen 1998, p. 118):

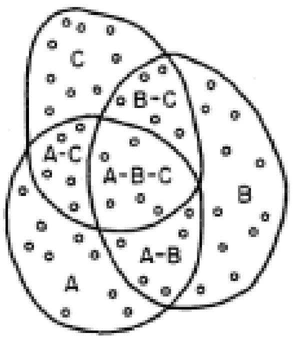


Figure 5.2. Illustration of the sampling theory of ability factors, in which the small circles represent neural elements or bonds and the large circles represent tests that sample different sets of elements (labeled A, B, and C). Correlation between tests is due to the number of elements they sample in common, represented by the areas of overlap. The overlap of A-B-C is the general factor, while the overlaps of A-B, A-C, and B-C are group factors. The non-overlapping areas are the tests' specificities. *Source: Bias in mental testing* by Arthur R. Jensen, Fig. 6.13, p. 238. Copyright © 1980 by Arthur R. Jensen. Reprinted with permission of the Free Press, a Division of Simon & Schuster, and Routledge Ltd.

The sampling model can be contrasted with models based on the idea that *g* is a unitary capacity that contributes to all cognitive efforts, reflecting some general property of the brain. For example, Arthur Jensen hypothesized that *g* is equivalent with mental speed or efficiency. In Jensen's model, there are specific abilities, but all of them depend, to a smaller or greater degree, on the overall speed or efficiency of the brain. In contrast, in the sampling model there are only specific abilities, overlapping samples of which are recruited for each cognitive task. Statistically, both models are equally able to account for empirically observed correlations between cognitive tests (see Bartholomew et al. 2009).

There are many flaws in Shalizi's argument. Firstly, the sampling model has several empirical problems which he ignores. I quote from Jensen 1998, pp. 120–121:

But there are other facts the overlapping elements theory cannot adequately explain. One such question is why a small number of certain kinds of nonverbal tests with minimal informational content, such as the Raven matrices, tend to have the highest *g* loadings, and why they correlate so highly with content-loaded tests such as vocabulary, which surely would seem to tap a largely different pool of neural elements. Another puzzle in terms of sampling theory is that tests such as forward and backward digit span memory, which must tap many common elements, are not as highly correlated as are, for instance, vocabulary and block designs, which would seem to have few elements in common. Of course, one could argue trivially in a circular fashion that a higher correlation means more elements in common, even though the theory can't tell us why seemingly very different tests have many elements in common and seemingly similar tests have relatively few.

[…]

And how would sampling theory explain the finding that choice reaction time is more highly correlated with scores on a nonspeeded vocabulary test than with scores on a test of clerical checking speed?

[…]

Perhaps the most problematic test of overlapping neural elements posited by the sampling theory would be to find two (or more) abilities, say, A and B, that are highly correlated in the general population, and then find some individuals in whom ability A is severely impaired without there being any impairment of ability B. For example, looking back at Figure 5.2 [see above], which illustrates sampling theory, we see a large area of overlap between the elements in Test A and the elements in Test B. But if many of the elements in A are eliminated, some of its elements that are shared with the correlated Test B will also be eliminated, and so performance on Test B (and also on Test C in this diagram) will be diminished accordingly. Yet it has been noted that there are cases of extreme impairment in a particular ability due to brain damage, or sensory deprivation due to blindness or deafness, or a failure in development of a certain ability due to certain chromosomal anomalies, without any sign of

> a corresponding deficit in other highly correlated abilities. On this point, behavioral geneticists Willerman and Bailey comment: "Correlations between phenotypically different mental tests may arise, not because of any causal connection among the mental elements required for correct solutions or because of the physical sharing of neural tissue, but because each test in part requires the same 'qualities' of brain for successful performance. For example, the efficiency of neural conduction or the extent of neuronal arborization may be correlated in different parts of the brain because of a similar epigenetic matrix, not because of concurrent functional overlap." A simple analogy to this would be two independent electric motors (analogous to specific brain functions) that perform different functions both running off the same battery (analogous to *g*). As the battery runs down, both motors slow down at the same rate in performing their functions, which are thus perfectly correlated although the motors themselves have no parts in common. But a malfunction of one machine would have no effect on the other machine, although a sampling theory would have predicted impaired performance for both machines.

But the fact that the sampling model has empirical shortcomings is not the biggest flaw in Shalizi's argument. The most serious problem is that he mistakenly believes that if the sampling model is deemed to be the correct description of the workings of intelligence, it means that there can be no general factor of intelligence. This inference is unwarranted and is based on a confusion of different levels of analysis. The question of whether or not there is a unidimensional scale of intelligence along which individuals can be arranged is independent of the question of what the neurobiological substrate of intelligence is like. Indeed, at a sufficiently basal (neurological, molecular, etc.) level, intelligence necessarily becomes fractionated, but that does not mean that there is no general factor of intelligence at the *behavioral* level. As explained above, many types of evidence show that *g* is indeed a centrally important unidimensional source of behavioral differences between individuals. One can compare this to a phenotype like height, which is simply a linear combination of the lengths of a number of different bones, yet at the same time unmistakably represents a unidimensional phenotype on which individual differ, and which can, among other things, also be a target for natural selection.

While he rejected the sampling model, Arthur Jensen noted that sampling represents an alternative model of *g* rather than a refutation thereof. This is because of the many lines of evidence showing that there is indeed a robust general factor of intellectual behavior. It is undoubtedly possible, with appropriate modifications, to devise a version of the sampling theory to account for all the empirical facts about *g*. However, this would mean that those uncorrelated abilities that are shared between all tests would have to show great invariance and permanence between different test batteries as well as be largely impervious to training effects, and they would also have to explain almost all of the practical validity and heritability of psychometric intelligence. Thus preferring the sampling model to a unitary *g* model is, in many ways, a distinction without a difference. The upshot is that ***regardless of whether "neuro-g" is unitary or the result of sampling, people differ on a highly important, genetically-based dimension of cognition that we may call general intelligence.*** Sampling does not disprove *g*. (The same applies to "mutualism", a third model of *g* introduced in van der Maas et al. 2006, so I will not discuss it in this post.)

V. Conclusions

Shalizi's first error is his assertion that cognitive tests correlate with each other because IQ test makers exclude tests that do not fit the positive manifold. In fact, more or less the opposite is true. Some of the greatest psychometricians have devoted their careers to disproving the positive manifold only to end up with nothing to show for it. Cognitive tests correlate because all of them truly share one or more sources of variance. This is a fact that any theory of intelligence must grapple with.

Shalizi's second error is to disregard the large body of evidence that has been presented in support of *g* as a unidimensional scale of human psychological differences. The *g* factor is not just about the positive manifold. A broad network of findings related to both social and biological variables indicates that people do in fact vary, both phenotypically and genetically, along this continuum that can be revealed by psychometric tests of intelligence and that has has widespread significance in human affairs.

Shalizi's third error is to think that were it shown that *g* is not a unitary variable neurobiologically, it would refute the concept of *g*. However, for most purposes, brain physiology is not the most relevant level of analysis of human intelligence. What matters is that *g* is a remarkably powerful and robust variable that has great explanatory force in understanding human behavior. Thus *g* exists at the behavioral level regardless of what its neurobiological underpinnings are like.

In many ways, criticisms of *g* like Shalizi's amount to "sure, it works in practice, but I don't think it works in theory". Shalizi faults *g* for being a "black box theory" that does not provide a mechanistic explanation of the workings of intelligence, disparaging psychometric measurement of intelligence as a mere "stop-gap" rather than a genuine scientific breakthrough. However, the fact that psychometricians have traditionally been primarily interested in validity and reliability is a feature, not a bug. Intelligence testing, unlike most fields of psychology and social science, is highly practical, being widely applied to diagnose learning problems and medical conditions and to select students and employees. What is important is that IQ tests reliably measure an important human characteristic, not the particular underlying neurobiological mechanisms. Nevertheless, research on general mental ability extends naturally into the life sciences, and continuous progress is being made in understanding *g* in terms of neurobiology (e.g., Lee et al. 2012, Penke et al. 2012, Kievit et al. 2012) and molecular genetics (e.g., Plomin et al., in press, Benyamin et al., in press).

P.S. See some of my further thoughts on these issues here.

References

Alliger, George M. (1988). Do Zero Correlations Really Exist among Measures of Different Intellectual Abilities? Educational and Psychological Measurement, 48, 275–280.

Arendasy, Martin E., & Sommer, Marcus (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. Intelligence, 41, 181–192.

Bartholomew, David J. et al. (2009). A new lease of life for Thomson's bonds model of intelligence. Psychological Review, 116, 567–579.

Benyamin, Beben et al. (in press). Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. Molecular Psychiatry.

Brody, Nathan (2003a). Construct validation of the Sternberg Triarchic Abilities Test. Comment and reanalysis. Intelligence, 31, 319–329.

Brody, Nathan (2003b). What Sternberg should have concluded. Intelligence, 31, 339–342.

Carroll, John B. (1995). Theoretical and Technical Issues in Identifying a Factor of General Intelligence. In Devlin, Bernard et al. (ed.), Intelligence, Genes, and Success: Scientists Respond to The Bell Curve. New York, NY: Springer.

Carroll, John B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In Nyborg, Helmuth (Ed.), The scientific study of general intelligence: Tribute to Arthur R. Jensen. Oxford, UK: Elsevier Science/Pergamon Press.

Chabris, Christopher F. (2007). Cognitive and Neurobiological Mechanisms of the Law of General Intelligence. In Roberts, M. J. (Ed.) Integrating the mind: Domain general versus domain specific processes in higher cognition. Hove, UK: Psychology Press.

Deary, Ian J. et al. (2007). Intelligence and educational achievement. Intelligence, 35, 13–21.

Deary, Ian J. et al. (2010). The neuroscience of human intelligence differences. Nature Reviews Neuroscience, 11, 201–211.

Elliott, Colin D. (1986). The factorial structure and specificity of the British Ability Scales. British Journal of Psychology, 77, 175–185.

Eysenck, Hans J. (1939). Primary Mental Abilities. British Journal of Educational Psychology, 9, 270–275.

Gottfredson, Linda S. (1997). Why *g* matters: The Complexity of Everyday Life. Intelligence, 24, 79–132.

Grinblatt, Mark et al. (2012). IQ, trading behavior, and performance. Journal of Financial Economics, 104, 339–362.

Guilford, Joy P. (1964). Zero correlations among tests of intellectual abilities. Psychological Bulletin, 61, 401–404.

Gustafsson, Jan-Eric (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179–203.

Humphreys, Lloyd G. et al. (1985). A Piagetian Test of General Intelligence. Developmental Psychology, 21, 872–877.

Jensen, Arthur R. (1998). The *g* factor: The science of mental ability. Westport, CT: Praeger.

Johnson, Wendy et al. (2004). Just one *g*: Consistent results from three test batteries. Intelligence, 32, 95–107.

Johnson, Wendy et al. (2008). Still just 1 *g*: Consistent results from five test batteries. Intelligence, 36, 81–95.

Johnson, Wendy, & Bouchard, Thomas J. (2011). The MISTRA data: Forty-two mental ability tests in three batteries. Intelligence, 39, 82–88.

Kamphaus, Randy W. (2005). Clinical Assessment of Child and Adolescent Intelligence. New York, NY: Springer.

Keith, Timothy Z. et al. (2001). What Does the Cognitive Assessment System (CAS) Measure? Joint Confirmatory Factor Analysis of the CAS and the Woodcock-Johnson Tests of Cognitive Ability (3rd Edition). School Psychology Review, 30, 89–119.

Kievit, Rogier A. et al. (2012). Intelligence and the brain: A model-based approach. Cognitive Neuroscience, 3, 89–97.

Lee, Tien-Wen et al. (2012). A smarter brain is associated with stronger neural interaction in healthy young females: A resting EEG coherence study. Intelligence 40, 38–48.

Melby-Lervåg, Monica, & Hulme, Charles. (2013) Is Working Memory Training Effective? A Meta-Analytic Review. Developmental Psychology, 49, 270–291.

Penke, Lars et al. (2012). Brain white matter tract integrity as a neural foundation for general intelligence. Molecular Psychiatry, 17, 1026–1030.

Petrill, Stephen A. (1997). Molarity versus modularity of cognitive functioning? A behavioral genetic perspective. Current Directions in Psychological Science, 6, 96–99.

Plomin, Robert, & Spinath, Frank M. (2004). Intelligence: genetics, genes, and genomics. Journal of Personality and Social Psychology, 86, 112–129.

Plomin, Robert et al. (in press). Common DNA Markers Can Account for More Than Half of the Genetic Influence on Cognitive Abilities. Psychological Science.

Ree, Malcolm J., & Earles, James A. (1991). Predicting training success: Not much more than g. Personnel Psychology, 44, 321–332.

Ree, Malcolm J. et al. (1994). Predicting job performance: Not much more than g. Journal of Applied Psychology, 79, 518–524.

Robertson, Kimberley F. et al. (2010). Beyond the Threshold Hypothesis: Even Among the Gifted and Top Math/Science Graduate Students, Cognitive Abilities, Vocational Interests, and Lifestyle Preferences Matter for Career Choice, Performance, and Persistence. Current Directions in Psychological Science, 19, 346–351.

Rushton, J. Philippe, & Jensen, Arthur R. (2010). The rise and fall of the Flynn Effect as a reason to expect a narrowing of the Black–White IQ gap. Intelligence, 38, 213–219.

Schmidt, Frank L., & Hunter, John E. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. Psychological Bulletin, 124, 262–274.

Schmidt, Frank L., & Hunter, John (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. Journal of Personality and Social Psychology, 86, 162–173.

Spearman, Charles (1939). Thurstone's work re-worked. Journal of Educational Psychology, 30, 1–16.

te Nijenhuis, Jan et al. (2007). Score gains on *g*-loaded tests: No *g*. Intelligence, 35, 283–300.

Thurstone, Edward L. (1938). Primary Mental Abilities. Chicago: Chicago University Press.

van der Maas, Han L. J. et al. (2006). A Dynamical Model of General Intelligence: The Positive Manifold of Intelligence by Mutualism. Psychological Review, 113, 842–861.

Woodcock, Richard W. (1990). Theoretical Foundations of the WJ-R Measures of Cognitive Ability. Journal of Psychoeducational Assessment, 8, 231–258.

Psychometrics

PREVIOUS POST
Cryptic Admixture, Mixed-Race Siblings, & Social Outcomes

NEXT POST
Color Differences: Corrections and Further Analysis. Part 2

71 Comments

B.B.

April 3, 2013 at 10:30 pm

Dalliard said
Shalizi's thesis is that the positive manifold is an artifact of test construction and that full-scale scores from different IQ batteries correlate only because they are designed to do that. It follows from this argument that if a test maker decided to disregard the g factor and construct a battery for assessing several independent abilities, the result would be a test with many zero or negative correlations among its subtests.

Forgive me if I'm missing something hear, but wouldn't Spearman's original work on the g factor already refute this? Presumably the early intelligence tests weren't made with the positive manifold in mind as it was yet to be discovered, yet Spearman was able to deduce a general factor of intelligence from these tests anyway.

REPLY

***Dalliard** (Post author)*

April 4, 2013 at 7:04 am

It's possible that the early intelligence tests did not tap into all cognitive abilities, or that Spearman and other "g men" included in their studies a limited variety of tests, thus guaranteeing the appearance of the positive manifold. However, as I showed above, researchers who have specifically attempted to create tests of uncorrelated abilities have failed, ending up with tests that are not substantially less g-saturated than those made with g in mind.

REPLY

Kiwiguy

April 4, 2013 at 12:03 am

***continuous progress is being made in understanding g in terms of neurobiology (e.g., Lee et al. 2012, Penke et al. 2012, Kievit et al. 2012) and molecular genetics ***

I think Steve Hsu pointed out, anyone who understands factor analysis realises that you can have correlations and a single largest factor even if there are no underlying causal reasons (i.e., it is just an accident). Nonetheless, these models may still be useful.

Prior to the availability of molecular studies the heritability of type II diabetes was estimated at 0.25 using all those methods. Now molecular studies have identified at least 9 loci involved in the disease. There are other examples in relation to height.

REPLY

---

pnard

April 4, 2013 at 7:09 pm

I thought you might have mentioned Gardner a little more. He never actually turned his theory into something testable so 3 researchers tested his intelligences and found intercorrelations and correlations with g.

If anyone's interested, the exchange went like this:
Visser et. al (2006). Beyond g: Putting multiple intelligences theory to the test

Gardner (2006). On failing to grasp the core of MI theory: A response to Visser et al.

Visser et. al (2006). g and the measurement of Multiple Intelligences: A response to Gardner

REPLY

Dalliard *(Post author)*

April 5, 2013 at 7:25 am

Yeah, Gardner's is another one of those failed non-g theories. I've read the Visser et al. articles, but Gardner's theory is really a non-starter because many of his supposedly uncorrelated intelligences are well-known to be correlated, and he does not even try to refute this empirically. Privately, Gardner also admits that relatively high general intelligence is needed for his multiple intelligences to be really operative. Jensen noted this in The g Factor, p. 128:

> As exemplars of each of these "intelligences" Gardner mentions the following famous persons: T. S. Eliot (linguistic), Einstein (logical-mathematical), Picasso (spatial), Stravinsky (musical), Martha Graham (bodily-kinesthetic), Sigmund Freud (intrapersonal), and Mahatma Gandhi (interpersonal). In an interesting book Gardner gives biographical analyses of each of these famous creative geniuses to illustrate his theory of multiple "intelligences" and of the psychological and developmental aspects of socially recognized creativity. **When I personally asked Gardner for his estimate of the lowest IQ one could possibly have and be included in a list of names such as this, he said, "About 120."** This would of course exclude 90 percent of the general population, and it testifies to the threshold nature of g. That is, a fairly high level of g is a necessary but not sufficient condition for achievement of socially significant creativity.

REPLY

Steve Sailer

April 5, 2013 at 4:26 pm

Gardner admitted to me in an email exchange that the existence of multiple intelligences made the existence of racial inequality in intelligence more likely. If only one number is relevant, then it's not that improbable in the abstract that all races could average the same number, just as men and women are pretty similar in overall IQ. But, if seven or eight forms of intelligence are highly important, the odds that all races are the same on all seven or eight is highly unlikely. Gardner agreed.

REPLY

Steve Sailer

April 5, 2013 at 1:58 am

Here's an example Shalizi uses that's worth thinking about because it actually unravels his argument:

"One of the examples in my data-mining class is to take a ten-dimensional data set about the attributes of different models of cars, and boil it down to two factors which, together, describe 83 percent of the variance across automobiles. [6] The leading factor, the automotive equivalent of g, is positively correlated with everything (price, engine size, passengers, length, wheelbase, weight, width, horsepower, turning radius) except gas mileage. It basically says whether the car is bigger or smaller than average. The second factor, which I picked to be uncorrelated with the first, is most positively correlated with price and horsepower, and negatively with the number of passengers — the sports-car/mini-van axis.

"In this case, the analysis makes up some variables which aren't too implausible-sounding, given our background knowledge. Mathematically, however, the first factor is just a weighted sum of the traits, with big positive weights on most variables and a negative weight on gas mileage. That we can make verbal sense of it is, to use a technical term, pure gravy. Really it's all just about redescribing the data."

Actually, I find his factor analysis quite useful. If he simply entered "price" as a negative number, he'd notice that his first factor was essentially Affordability v. Luxury, in which various desirable traits (horsepower, size, etc.) are traded off against price and MPG.

What's really interesting and non-trivial about the g-factor theory is that cognitive traits aren't being traded off the way affordability and luxury are traded off among cars. People who are above average at reading are, typically, also above average on math. That is not something that you would necessarily guess ahead of time. (Presumably, the tradeoff costs for higher g involve things like more difficult births, greater nutrition, poorer balance, more discrete mating, longer immature periods, more investment required in offspring, and so forth.)

REPLY

Steve Sailer

April 5, 2013 at 2:40 am

I think the essence of Shalizi's mistake is conveniently summed up in his first sentence:

"Attention Conservation Notice: About 11,000 words on the triviality of finding that positively correlated variables are all correlated with a linear combination of each other, and why this becomes no more profound when the variables are scores on intelligence tests."

This reminds me of the old joke about the starving economist on the desert island who finds a can of beans: "Assume we have a can opener …"

Shalizi just assumes that all cognitive traits are positively correlated, and then goes on from there with his argument. But the fact that virtually all cognitive traits are positively correlated is astonishing.

Most things in this world involve tradeoffs. Think about automotive engineering. More of one thing (e.g., luxury) means less of another thing (e.g., money left over in your bank account).

Look at Shalizi's example of ten traits regarding automobiles. In terms of desirability, some are positively correlated, some are negatively correlated:

Positive or neutral: passengers, length, wheelbase, weight, width, horsepower, engine size

Negative: price, turning radius, average fuel cost per 15,000 miles (i.e., MPG restated)

The fact that, on average, there aren't tradeoffs between cognitive traits is highly nontrivial.

REPLY

Dalliard *(Post author)*

April 5, 2013 at 7:30 am

Intellectuals may be prone to being skeptical of g because most people they associate with are high on g, which makes specific abilities more salient. For example, in his heritability book, Neven Sesardic gives the following, remarkably wrong-headed quote from the British philosopher Gilbert Ryle:

"Only occasionally is there even a weak inference from a person's possession of a high degree of one species of intelligence to his possession of a high degree of another."

(It's from a 1974 article called Intelligence and the Logic of the Nature–Nurture Issue.)

REPLY

FredRR

April 5, 2013 at 3:26 pm

You should invite him to respond. Would be interesting

REPLY

B.B.

April 6, 2013 at 7:37 am

Looking at Shalizi's last article tagged with "IQ" (dated June 16th, 2009), it looks like he isn't eager about continuing discussions on the matter.

REPLY

Steve Sailer

April 5, 2013 at 4:28 pm

Okay, but am I overlooking something in saying that the root problem with Shalizi's argument, in which he makes up numbers that are all positively related to each other and shows that you often see a high general factor even with random numbers, is that this "positive manifold" in which practically all cognitive tasks are positively correlated is pretty remarkable, since we don't see the kind of trade-offs that we expect in engineering problems?

REPLY

Dalliard *(Post author)*

April 6, 2013 at 11:26 pm

Yes. And thanks for the link, Steve.

REPLY

Jez Wunderin

April 6, 2013 at 6:04 pm

Dalliard, you write very well!

Even though, as Steve Sailer says, it is striking that there are no obvious tradeoffs between the needs of different tasks, we are still left with another question about possible tradeoffs: Why so much variability in g? Is there a Darwinian downside to having too much little g? Is the dumb brute greater in reproductive fitness for some reason? If so, what reason? One can imagine lots of scenarios–is there any way to test them, I wonder?

I guess if it turns out that little g variability reflects mutation load, then there is no need to postulate a tradeoff?

REPLY

Dalliard *(Post author)*

April 6, 2013 at 10:05 pm

Thanks. I don't have a good answer with regard to variability. Mutation load would make the most sense, but it may not be the whole story. It's easy to come up with hypothetical scenarios, as you say. Of course, this is a problem with heritable quantitative traits in general. What I like about heritability analysis is that you don't really need to worry about the ultimate causes of genetic variation.

REPLY

Douglas Knight

April 6, 2013 at 9:08 pm

May I suggest adding a table of contents at the top with internal links to the sections numbered with roman numerals?

REPLY

pnard

April 6, 2013 at 9:35 pm

Yeah. This is a very good article and some extra readability couldn't hurt!

REPLY

Dalliard *(Post author)*

April 6, 2013 at 10:18 pm

Done.

REPLY

teageegeepea

April 6, 2013 at 9:58 pm

Another branch of psychometrics, personality testing, tends to use a five factor model. To what extent can we say those factors are simply what is found "in the data" vs created by psychometricians?

Also, the g factor is first referred to as accounting for 59 percent of common factor variance, and later said to account for less than half the variance in an IQ battery. Is that because of the contribution of non-common factors to the variance?

REPLY

Dalliard *(Post author)*

April 6, 2013 at 11:20 pm

I'd say that the Big Five are much less real than g. There's a good recent paper that compares the Big Five and their facets (sub-traits). They found that most Big Five traits are not "genetically crisp" because genetic effects on the facets are often independent of the genetic effects on the corresponding Big Five traits. Moreover, if you use a Big Five trait to predict something, you will probably forgo substantial validity if you don't analyze data at the facet level, whereas with IQ tests little is gained by going beyond the full-scale score in most cases.

In factor analysis, the total variance is due to common factor variance, test-specific variance, and error variance. (There are no non-common factors, because factors are by definition common to at least two subtests.) g usually accounts for less than 50 percent of the total variance but more than 50 percent of the common factor variance.

REPLY

Steve Sailer

April 6, 2013 at 11:32 pm

So, we can approximate that the g glass is about half full and half empty simultaneously.

I think human beings have problems thinking about things where the glass is both half full and half empty. Yet, we seem to be most interested in arguing about situations that are roughly 50-50.

REPLY

anonymous123

April 7, 2013 at 4:03 pm

A guess as to part of what's happening here:

It stands to reason that multiple areas of the brain are recruited as part of cognition. This makes the "sampling model" intuitively appealing, while making g intuitively difficult to understand as a causal mechanism. However, the question of how cognition works and the question of what underlies individual differences in cognition are two quite separate questions.

The model which seems to fit the data presented here is that g ultimately reflects a collection of features of neuronal cell physiology as well as the physiology of higher-level parts of neuroanatomy that vary between individuals. Genetic effects on cell physiology and brain development tend to have brain-wide impacts, which get reflected in g. In contrast, one might imagine that various non-genetic effects would have more localized impacts on the brain and thus more variegated effects on variation in cognitive abilities. This causes the heritability of g to come close to 100%, while the heritability of composite IQ scores can be much less.

REPLY

Eric Rasmusen

April 7, 2013 at 9:12 pm

(1) Very nice essay. I know I should reread it, and Shalizi. Shalizi's essay is better than you make out. This isn't because it says useful things about IQ, I think, but because it says useful things about factor analysis. Where he goes wrong seems to be in thinking that the deficiencies of factor analysis destroy the concept of g.

(2) Can you write something on the Big Five? I know psychologists like it better than Myers-Briggs, but the main reason seems to be because they like factor analysis. I can see that they may have found the 5 most important factors, and maybe there is a big dropoff to going to six, but I wonder if they can really label the 5 meaningfully (what does Neuroticism really mean?). The nice thing about Myers Briggs is that people see the results and say, "Oh, yes, I see that from my experience," just as with IQ people say say, "Of course, some people are smarter than others, it's just common sense that there exists something we call intelligence."

(3) Is the multiple-intelligences theory, and in particular Shalizi's example of 100s of independent abilities, really just saying that we call somebody smart if they are high in their sum of high abilities rather than just being high in one ability? Is there a real difference between the two things? (that's a serious question)

REPLY

Dalliard *(Post author)*

April 8, 2013 at 6:52 am

A problem with the Big Five is indeed that it relies so heavily on (exploratory) factor analysis (whereas g theory is based on a wide range of evidence aside from factor analysis). See also the article I linked to above in reply to teageegeepea.

The problem with Myers-Briggs is that it lacks predictive validity, i.e., it does not seem to tell anything important about people.

> Is the multiple-intelligences theory, and in particular Shalizi's example of 100s of independent abilities, really just saying that we call somebody smart if they are high in their sum of high abilities rather than just being high in one ability? Is there a real difference between the two things? (that's a serious question)

Yes, that's the basic idea. Shalizi's argument is that it's arbitrary to use this sum of abilities, while my argument is that this supposed sum of abilities looks suspiciously like one single ability or capacity which represents the most important, and often the only important, dimension of cognitive differences.

REPLY

Macrobius

April 7, 2013 at 10:10 pm

I believe the above critique doesn't hit the mark, at least as regards 'errors' 1 and 3 (the reference to work on confirming factor analysis is much more direct).

Part 1

Shalizi: this hypothesis is not falsifiable, and here is a simulation experiment that demonstrates that fact.
Dalliard: here are lots of studies showing the hypothesis is true.

Part 3

@Dalliard: I believe you are mistaking the simulation experiment and its role as a 'null hypothesis' in the overall framework of Shalizi's article, with something else you know all about. He is not advocating the sampling model (and in fact is using random numbers) in his simulation experiment. This section is entirely a stawman argument.

There is very little evidence in the above blog to show that Dalliard has understood and engaged Shalizi's argument.

A more reasonable conclusion would be that the article, written in 2007, is now dated. Whether it was valid in 2007 depends a lot on your assessment of Jensen's 1998 work — both those topics would make for very constructive further explanation, I think.

REPLY

Pincher Martin

April 8, 2013 at 2:39 am

Your summary of Part 1 is inaccurate.

Shalizi claims that intelligence tests are made to positively correlate with each other.

Dalliard counters by arguing that if a test maker decided to ignore *g*, it would still pop up in any test he made because the positive correlations are not constructs of tests, but an empirical reality. He then cites evidence that supports his argument.

Shalizi's simulation, therefore, is nothing more than a GIGO model showing that randomly *positive* correlations also demonstrate a general factor similar to what is found in IQ tests. This is true, but uninteresting; it still doesn't explain how the uniformity of positive correlations in tests exist in the first place.

REPLY

***Dalliard** (Post author)*

April 8, 2013 at 7:03 am

You're right as far as "Part 1" is concerned, but just to be clear, the abilities in Shalizi's toy model are genuinely uncorrelated. Correlations between tests emerge because all of them call on some of the same abilities, and g corresponds to average individual differences across those shared abilities.

REPLY

Pincher Martin

April 9, 2013 at 1:10 am

@Dalliard 7:03 AM,

"You're right as far as "Part 1" is concerned, but just to be clear, the abilities in Shalizi's toy model are genuinely uncorrelated."

Shalizi is pretty clear that the seemingly random variables in his simulation are supposed to be positively correlated – i.e., they're not really random at all. That simulation would not work at showing a *g* factor if those random factors were genuinely uncorrelated.

Shalizi writes:

"If I take any group of variables which are positively correlated, there will, as a matter of algebraic necessity, be a single dominant general factor, which describes more of the variance than any other, and all of them will be "positively loaded" on this factor, i.e., positively correlated with it. Similarly, if you do hierarchical factor analysis, you will always be able to find a single higher-order factor which loads positively onto the lower-order factors and, through them, the actual observables. What psychologists sometimes call the "positive manifold" condition is enough, in and of itself, to

guarantee that there will appear to be a general factor. Since intelligence tests are made to correlate with each other, it follows trivially that there must appear to be a general factor of intelligence. This is true whether or not there really is a single variable which explains test scores or not."

Everything in Shalizi's argument in the above quote hinges on his assumption that IQ tests, and the various subtests within them, are meant to be positively correlated with each other. His simulation works from that assumption, which as you point out is an incorrect assumption.

Shalizi later writes: *"If I take an arbitrary set of positive correlations, provided there are not too many variables and the individual correlations are not too weak, then the apparent general factor will, typically, seem to describe a large chunk of the variance in the individual scores."*

So Shalizi starts off by assuming a *g* factor in his simulation and then wonders why psychologists are so impressed with finding a *g* factor in their tests.

The answer is, of course, that there is no earthly reason why psychologists should have necessarily found a *g* factor in their tests. The abilities measured in them – unlike Shalizi's simulation – could have very well been uncorrelated or even negatively correlated.

REPLY

Dalliard *(Post author)*

April 9, 2013 at 1:59 pm

In Shalizi's model, the *abilities* are based on random numbers and are therefore (approximately) uncorrelated, while the *tests* are positively correlated. Each test taps into many abilities, and correlations between tests are due to overlap between the abilities that the tests call on. If each test in Shalizi's model called on just one ability or on non-overlapping samples of abilities, then the tests would also be uncorrelated.

REPLY

Pincher Martin

April 9, 2013 at 11:43 pm

@Dalliard 1:59 PM,

"In Shalizi's model, the abilities are based on random numbers and are therefore (approximately) uncorrelated, while the tests are positively correlated. Each test taps into many abilities, and correlations between tests are due to overlap between the abilities that the tests call on. If each test in Shalizi's model called on just one ability or on non-overlapping samples of abilities, then the tests would also be uncorrelated."

Thanks for the clarification.

So is Shalizi's error in not realizing that a person's *g* is fairly consistent when measured and compared across several IQ tests? That's how I read this passage you wrote to Macrobius:

"Various kinds of evidence have been proffered in support of the notion that the same g is measured by all diverse IQ batteries, ***but the best evidence comes from confirmatory factor analyses showing that g factors are statistically invariant across batteries****. This, of course, directly contradicts the predictions of g critics like Thurstone, Horn, and Schonemann."*

I assume this evidence contradicts the critics because random numbers – similar to those in Shalizi's simulation – would not produce a consistent *g* across several batteries. Is that correct?

REPLY

Dalliard *(Post author)*
April 10, 2013 at 2:03 pm

The random numbers aren't that important, they're just a way to introduce individual differences to the model. Shalizi's mistake is to think that the fact that correlations between tests can be generated by a model without a unitary general factor has any serious implications for the reality of g. Any sampling model must be capable of explaining the known facts about g, including its invariance across batteries, which means that sampling, if real, is just about explaining the operation of a unidimensional g at a lower level of analysis.

REPLY

Dalliard *(Post author)*
April 8, 2013 at 7:14 am

> Shalizi: this hypothesis is not falsifiable, and here is a simulation experiment that demonstrates that fact.
> Dalliard: here are lots of studies showing the hypothesis is true.

As Pincher Martin pointed out, the simulation experiment is not related to Shalizi's first error. The first error is the assertion that there are cognitive tests that are uncorrelated or negatively correlated with tests included in traditional IQ batteries. There is no evidence that this is the case, and there are tons of evidence to the contrary (perhaps the closest is face recognition ability, which is relatively independent, but even it has a small g loading in studies I've seen). It's conceivable that there are "black swan" tests of abilities that do not fit the pattern of positive correlations, but even then it's clear that a very wide range of cognitive abilities, including all that our educational institutions regard as important, are positively correlated.

> I believe you are mistaking the simulation experiment and its role as a 'null hypothesis' in the overall framework of Shalizi's article, with something else you know all about. He is not advocating the sampling model (and in fact is using random numbers) in his simulation experiment. This section is entirely a stawman argument.

Nope. The simulation represents an extreme version of sampling, and Shalizi doesn't claim that it's a realistic model, but he nevertheless thinks that g is most likely explained by the recruitment of many different neural elements for the same intellectual task, with some of these elements overlapping across different tasks.

This is how he puts it: "[T]here are lots of mental modules, which are highly specialized in their information-processing, and that almost any meaningful task calls on many of them, their pattern of interaction shifting from task to task." My counter-point is that even if sampling is true, it does not invalidate g. Any model of intelligence must account for the empirical facts about g, which in the case of sampling means that there must be a hierarchy of intelligence-related neural elements, some of them central and others much less important, with g corresponding to the former.

> There is very little evidence in the above blog to show that Dalliard has understood and engaged Shalizi's argument.

I understand Shalizi's argument, whereas most people who regard it as a cogent refutation of general intelligence do not. I also engage his argument at length.

> A more reasonable conclusion would be that the article, written in 2007, is now dated. Whether it was valid in 2007 depends a lot on your assessment of Jensen's 1998 work — both those topics would make for very constructive further explanation, I think.

My post could as well have been written in 2007. I cited some more recent studies, but they are not central to my argument. All the relevant evidence was available to Shalizi in 2007, but he didn't know about it or decided to ignore it.

REPLY

Macrobius

April 8, 2013 at 10:06 pm

I thank both @Dalliard and @Pincher Martin for their incisive replies that helped me understand what is being claimed, esp. as regards to 'error 1'.

I do indeed see evidence that Shalizi believes as you say. I will argue, however, that does not harm his argument in the way claimed. Before I do that, though, allow me to comment about what I take to be his point in the post. Unfortunately, most of his substantive points are actually in the footnotes. I take Shalizi to be largely recapitulating the paper of Shoenemann he references in n.2 ('Factorial Definitions of Intelligence: Dubious Legacy of Data Analysis'). My evidence is that hardly anything he says not in that paper, in greater detail, and the tone of the polemic and its aim is quite similar. In fact, I would describe the post as a pedagogical exposition of Shoenemann's views — with *one* extension.

Allow me to explain: Shoenemann is quite clear he regards Spearman's g and related factor analysis, and Jensen's definition of g in terms of PCA, to be changing the definition of g on the fly. In this context, he recapitulates that history of Spearman's g and Thurstone's views, giving all the critiques that Shalizi raises, then shifts gears and gives his opinion of Jensen.

One of the problems with Shalizi's post is he minimizes the transition from Factor Analysis to PCA, which Shoenemann treats as models having very different properties and hence critiques. None the less, it is clear that Shalizi is influenced by Shoenemann, and besides his primary exposition is trying to recapitulate Thomson's construction (pp. 334-6, op. cit.).

Specifically, with regard to 'error 1' I think what Shalizi is trying to do is precisely replicate this passage of Thomson, only for Jensen's PCA based g:

‘Hierarchical order [i.e. ideal rank one] will arise among correlation coefficients unless we take pains to suppress it. It does not point to the presence of a general factor, nor can it be made the touchstone for any particular form of hypothesis, for it occurs even if we make only the negative assumption that *we do not know* how the correlations are caused, if we assume only that the connexions are random’

This passage is immediately followed by a mathematical analogy along the lines of Shalizi’s squares and primes, though different. The honest thing to do is to ask Shalizi at this point if he had this passage in mind, when he devised his toy simulation, though I don’t doubt the answer myself.

One reason to pay attention to this background, from a polemic standpoint, is that even if you ‘take down Shalizi’ and neutralize his post, you leave yourself open to a very simple rejoinder: that Shalizi was just a flawed version of Shoenemann, and what about that? However, I don’t think we’ve yet reached the point we can say Shalizi is flawed, and I will explain that next.

Let’s start with the basic facts of how Frequentist inference works: you have an unrestricted model (H1, estimated by your data), a restricted model (the null hypothesis, usually estimated by some assumptions the restrictions affords you), and a ‘metric’ — say Wald distance, Likelihood Ratio, or Lagrange Multiplier. For definite, since it is most appropriate to this context, let’s take Likelihood Ratio. Next, one notes that the likelihood of the *restricted* model is always less than the unrestricted, so that the LR is bounded 0 <= LR < 1. That is, you *must* put the restricted model in the numerator — if you do not, then you don't get a compelling inference. The restricted model is always *trivially* less likely than the unrestricted, so you are refuted if the trivially less likely model ends up being more likely than H1 given your data. This is what gives Frequentist inference it's force — the fact that it can be trivially falsified, in case it happens to be uninformative.

So here's the puzzler: Shalizi is a statistician, and yet he's chosen to randomize the *unrestricted model*. Did he just make a howling blunder? If so, Dalliard is straining at a speck of an error here, when he should be putting a beam through Shalizi's eye (and maybe Thomson's as well).

Of course, I believe Shalizi has done no such thing, being a competent statistician — beyond perhaps making the form of his inference here explicit. If my hunch that he is following Thomson's logic exactly turns out to be correct, then what he must be doing is some sort of dominance argument, by constructing a likelihood that is *greater* than the 'positive manifold' of the restricted model. It would be really productive if someone involved in this spat were to spell out the form of inference — if any — Shalizi and Thomson are trying to use! Because it certainly doesn't follow the *normal* template of Frequentist reasoning, people may be assuming. It's bass ackwards.

Now, does 'error 1' have any force? I don't think so, even if Shalizi holds the proposition and is wrong about it. Nothing hinges, in the form of argument — assuming again it is not just sheer blunder, which I doubt — on the question of whether the restricted model is enforced by empirics or by design. Frequentist inference is about correlation, and just doesn't give a damn about that sort of thing. I may be in error here, and would be happy to have my error explained to me.

For @Dalliard, a simple question: do you believe Thomson's argument was persuasive against the Two Factor Model? And secondly, if you believe it was, do you believe a similar argument could succeed in principle against Jensen's PCA version of g? If not, why not?

REPLY

Dalliard *(Post author)*

April 9, 2013 at 2:15 pm

> I do indeed see evidence that Shalizi believes as you say. I will argue, however, that does not harm his argument in the way claimed.

What I termed Shalizi's first error is simply the claim that if a sample of people takes a bunch of intelligence tests, the results of those tests will NOT be uniformly positively correlated if you include tests that are different from those used in traditional batteries like the Wechsler. I showed that all the evidence we have indicates that this claim is false.

It appears that you confuse the question *whether* tests correlate with the question *why* they correlate. But these are separate questions. Whether tests correlate because there's some unitary general factor or because all tests call on the same abilities, the correlations are there.

Thomson's model is about *why* the particular pattern of correlations exists. He showed that it would arise even if there are only uncorrelated abilities provided that they are shared between tests to some extent. He didn't claim that his model falsified Spearman's, only that Spearman's explanation wasn't the only possible one. Of course, it later became apparent that both Spearman's two-factor model and Thomson's original model are false, because they cannot account for group factors.

The modern g theory posits that there's a hierarchy of abilities, with g at the apex. As Shalizi points out, such multiple-factor models are unfortunately not as readily falsifiable as Spearman's two-factor model was. Various kinds of evidence have been proffered in support of the notion that the same g is measured by all diverse IQ batteries, but the best evidence comes from confirmatory factor analyses showing that g factors are statistically invariant across batteries. This, of course, directly contradicts the predictions of g critics like Thurstone, Horn, and Schonemann.

When you add to this g's intimate associations with genetic variables, practical outcomes, practice effects, etc., as explained in my post, it becomes clear that it's difficult to explain human cognitive differences without reference to something very much like general intelligence. This is the case regardless of how unitary or not the neurophysiology of intelligence is. If you want to argue that there's no general intelligence, you must show how all these facts fit into an alternative model, something Shalizi doesn't do.

(A note on Jensen and PCA: He actually regarded the Schmid-Leiman procedure as the best method to extract g, although he also showed that the choice of method makes little difference.)

REPLY

Macrobius

April 9, 2013 at 11:24 pm

Thanks for your further clarification. I should mention, before leaving off the topic, that Shalizi has covered the same material in his lecture notes for a course, some two years after the post referenced (see lectures 10-13).

http://vserver1.cscs.lsa.umich.edu/~crshalizi/weblog/617.html

I don't think this will change any of the discussion, but it has more formalism and clarity.

I look forward to you addressing the heritability part of the article, if you get a chance.

REPLY

Dalliard (*Post author*)

April 10, 2013 at 2:10 pm

Thanks, I'll take a look at those notes.

I don't feel like delving into Shalizi's claims about heritability at the moment, but perhaps I will later. In general, GCTA has been a methodological weapon of mass destruction with regard to arguments seeking to minimize the role that genes play in causing intelligence differences (although I don't think those arguments were very strong to begin with).

REPLY

Noah Smith

April 10, 2013 at 5:30 pm

There is an alternative hypothesis to the g hypothesis: Multiple general factors.

It's possible that simple mental tasks (of the kind used in all psychometric tests) can be performed by a number of different substitutable mental systems.

For example, suppose the performances of subject i on tests m and n are given by:

$P_{mi} = a + b_m * X_i + c_m * Y_i + e_{mi}$
$P_{ni} = a + b_n * X_i + c_n * Y_i + e_{ni}$

Here, X and Y are two different cognitive abilities. b and c are positive constants. Assume X_i and Y_i are uncorrelated, and assume e, the error term, is uncorrelated across tests and across individuals.

In this case, assessing the covariance of performances across two tests m and n, we will have:
$Cov(P_{mi}, P_{ni}) = b_m * b_n * Var(X_i) + c_m * c_n * Var(Y_i) > 0$

So even though the two cognitive abilities are uncorrelated (i.e. there is no true "g"), all tests are positively correlated (the "positive manifold" holds), and thus a "g"-type factor can be extracted for any set of tests.

To make this example concrete, suppose that there are two statistically independent mental abilities, spatial modeling and symbolic modeling (I just made those up). And suppose that any simple information-processing task can be solved using spatial modeling, or solved using symbolic modeling, or solved using some combination of the two. That would result in a positive correlation between all simple information-processing tasks, without any dependence between the two mental abilities.

Of course, the functional form I chose has the two abilities be *perfect* substitutes, but that is not necessary for the result to hold.

This has long been my intuitive working hypothesis about mental ability. I have noticed that I tend to solve most math and physics problems symbolically (by writing down equations), while some of my peers seem to solve them all graphically (by drawing pictures). That led me to believe that some people are visual thinkers, while others are symbolic thinkers. That intuition was reinforced by my own high performance on the "mathematical" and "linguistic" parts of IQ tests (which I used to take online for fun), and my average performance on the "visual" or "spatial" parts of the tests.

Of course, my intuition could easily be wrong. But the math above seems to show that g-like factors can emerge when there are in fact many general intelligence factors present.

REPLY

Dalliard *(Post author)*

April 11, 2013 at 9:12 am

Your idea seems to be a version of the sampling model discussed in my post. My point is that even if one general intelligence becomes many at some level, at the behavioral level it is unitary. I discussed research showing that while people have relative cognitive strengths and weaknesses these contribute little to the prediction of educational and job outcomes net of the general level of ability. Moreover, different test batteries appear to tap into one and the same g, and there are many other indications of the generality and unidimensionality of g. While the positive manifold may be explained by reference to different kinds of mechanisms, all of them must be able to account for the empirical facts about g which go far beoynd the mere existence of positive correlations between tests. Those who posit sampling models almost never consider how their models fit together with what we know about g beyond the positive correlations, or whether their models really falsify g or just describe it at another level of analysis.

You may think that there's a big difference between "visual thinkers" and "symbolic thinkers" or whatever, but research does not support this learning styles paradigm. In their review Pashler et al. concluded:

> Our review of the literature disclosed ample evidence that children and adults will, if asked, express preferences about how they prefer information to be presented to them. There is also plentiful evidence arguing that people differ in the degree to which they have some fairly specific aptitudes for different kinds of thinking and for processing different types of information. However, we found virtually no evidence for the interaction pattern mentioned above [i.e., positive interactions between similar instructional and self-reported learning styles, or "meshing"], which was judged to be a precondition for validating the educational applications of learning styles. Although the literature on learning styles is enormous, very few studies have even

> used an experimental methodology capable of testing the validity of learning styles applied to education. Moreover, of those that did use an appropriate method, several found results that flatly contradict the popular meshing hypothesis.

This indicates that learning is a highly general capacity. A caveat here is, as discussed in my post, that at high levels of IQ, specific abilities are more independent. However, you shouldn't use observations based on exceptional, high-ability individuals to make general conclusions.

I corrected the covariance equation.

REPLY

erasmuse

April 11, 2013 at 9:58 am

This is a digression, but an interesting one. I read Pashler's abstract, and what it says is not "evidence shows learning styles don't matter" but "the research on learning styles is done too badly to show whether learning styles matter".

This is a different question than whether there are multiple abilities. The Pashler question is whether you can sort out people using personality tests (or suchlike) and then use, e.g. lectures for some and books for others to teach them things better than if you used lectures for all or books for all. The question is still just as relevant if every student's g intelligence is identical. As I understand it, people suggest that as a hypothesis but nobody's done good experiments on it. Is that right?

Actually, one can ask a similar question based on sorting by g. Common sense says that if you have to teach, for example, Bayes's Rule to a heterogeneous group, you should have the low-g one memorize it and the high-g ones learn how to rederive it. Is that experimentally confirmed?

REPLY

Steve Sailer

April 16, 2013 at 12:21 am

"I have noticed that I tend to solve most math and physics problems symbolically (by writing down equations), while some of my peers seem to solve them all graphically (by drawing pictures)."

Dear Professor Smith:

Your peers at Carnegie-Mellon are all above average in intelligence. Moreover, they typically tend to be above average on all forms of intelligence.

What's really interesting about the g factor is that people who are above average in spatial ability are not, on average, below average on symbolic ability, and vice-veras. Your fellow professors who are geniuses at spatial reasoning don't confine their reading to, say, the comments section on Youtube videos.

In general, those who are above average on one trait tend to be above average on another. It's not like, say, with cars where acceleration and gas mileage tend to be inversely correlated. The positive manifold of cognitive skills is a strange and important fact of nature that Dr. Shalizi tried to assume away in classic "Assume we have a can opener" style.

REPLY

猛虎

April 11, 2013 at 9:46 am

Noah, my english is horrific but anyway your idea is not new. More or less the same thing has been stated elsewhere, in a recent and over cited study.
http://www.thestar.com/life/2012/12/19/iq_a_myth_study_says.html

The referenced paper can be found here :
Fractionating Human Intelligence (Hampshire et al. 2012)

Apart from the small sample size (n=16), it also fails to understand the nature of g. Here's a passage of Jensen's 1998 book, The g Factor (here), pages 130-132, about the unity of g and the concept of modular abilities, which is what your are referring to.

> The g factor, which is needed theoretically to account for the positive correlations between all tests, is necessarily unitary only within the domain of factor analysis. But the brain mechanisms or processes responsible for the fact that individual differences in a variety of abilities are positively correlated, giving rise to g, need not be unitary. … Some modules may be reflected in the primary factors; but there are other modules that do not show up as factors, such as the ability to acquire language, quick recognition memory for human faces, and three-dimensional space perception, because individual differences among normal persons are too slight for these virtually universal abilities to emerge as factors, or sources of variance. This makes them no less real or important. Modules are distinct, innate brain structures that have developed in the course of human evolution. They are especially characterized by the various ways that information or knowledge is represented by the neural activity of the brain. The main modules thus are linguistic (verbal/auditory/lexical/semantic), visuospatial, object recognition, numerical-mathematical, musical, and kinesthetic. …
>
> In contrast, there are persons whose tested general level of ability is within the normal range, yet who, because of a localized brain lesion, show a severe deficiency in some particular ability, such as face recognition, receptive or expressive language dysfunctions (aphasia), or inability to form long-term memories of events. Again, modularity is evidenced by the fact that these functional deficiencies are quite isolated from the person's total repertoire of abilities. Even in persons with a normally intact brain, a module's efficiency can be narrowly enhanced through extensive experience and practice in the particular domain served by the module.

Elsewhere, he notes, pages 259-261 :

> But at some level of analysis of the processes correlated with g it will certainly be found that more than a single process is responsible for g, whether these processes are at the level of the processes measured by elementary cognitive tasks, or at the level of neurophysiological processes, or even at the molecular level of neural activity. If successful performance on every complex mental test involves, let us say, two distinct, uncorrelated processes, A and B (which are distinguishable and measurable at some less complex level than that of the said tests) in addition to any other processes that are specific to each test or common only to certain groups of tests, then in a factor analysis all tests containing A and B will be loaded on a general factor. At this level of analysis, this general factor will forever appear unitary, although it is actually the result of two separate processes, A and B. … However, the fact that g has all the characteristics of a polygenic trait (with a substantial component of nongenetic variance) and is correlated with a number of complexly determined aspects of brain anatomy and physiology, as indicated in Chapter 6, makes it highly probable that g, though unitary at a psychometric level of analysis, is not unitary at a biological level.

By the way,
http://noahpinionblog.blogspot.co.uk/2013/04/nuthin-but-g-thang.html

REPLY

Macrobius

April 12, 2013 at 7:25 pm

Exceedingly interesting. That does make me think that the claim for unitary g must be analogous to what, in Economics, is called GARP — that is, the existence of a unitary utility function that can rationalise the test data, in the sense of Discrete Choice theory — a sort of Generalised Axiom of Revealed Intelligence.

Has there been any work along those lines? Maybe it's time to get Hal Varian involved. I can see that if we ask 'what is the economic value of a question on an IQ test to the individual, that theory heterogeneous variability allows them to 'punch above their weight' on, might be an excellent way to find distributional evidence in the data for the theory. Such situations are not only extremely rare — they are Generalised Extremely *Valuable* to the individual in question, given the economic value of the test!

REPLY

Macrobius

April 13, 2013 at 8:09 am

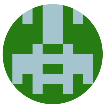
Fleshing out my though a bit — it seems to me that 'IQ' is just voting theory turned on its side, so to speak. Suppose we have an island with 2000 people and 7 policy alternatives. We form a matrix with 7 rows and 2000 columns. Each column contains the preference (a rank from 1 to 7) of each 'voter'. Along the right hand side, we have a social welfare function that computes the social (global island) utility of that alternative. Under suitable assumptions, there is a single voter whose preferences must mirror the social welfare function. Choosing him for dictator is superior to any voting scheme, so far as directly selecting the highest social welfare for the island is concerned.

Now, instead of making the people the columns, make them the rows — that is we have 7 people and give them a battery of IQ tests consisting of 2000 questions total. Along the right hand column, write down their true IQ score — to the 'global intelligence' of each of the 7 test takers in fact known, but we wish to rank the individuals without knowing it using some computation based on their answers or perhaps additional information.

For each question on each test, we can under suitable assumptions rank the value of that question, ordinal-ly, for each of the 7 individuals — and assuredly each question has economic value to them, since the higher their 'bundle of scores' subject to their 'intelligence constraint' the better adapted to life they are, which is a sort of utility. Therefore we have columns that are permutations of the numbers 1-7.

Again, under suitable assumptions, there is a single question on a single test the value of answering which exactly mimics the IQ ranking. Call it the Dictator Question. That is, IQ would seem to be a mirror of 'ordinal social welfare', when we think of people as 'adaptive policy alternatives' in an evolutionary situation, and the situations they face — modelled in tests — as the 'individuals'. The Dictator Question plays the role of a 'representative agent' I guess — in the sense that a test replicating it many times would correctly model the expectations of administering a battery of tests to a population.

REPLY



Macrobius

April 13, 2013 at 12:33 pm

And, my 'final, final thought' is that Shalizi — to return to the topic of the OP — has written a review of the Flynn Effect, which may at least have the advantage of clarifying what he thought, before he swore off 'IQ debates' entirely and it would seem irrevocably:

http://vserver1.cscs.lsa.umich.edu/~crshalizi/reviews/flynn-beyond/

'As data reduction, factor analysis is harmless, but there has always been a temptation to "reify" the factors, to suppose that factor analysis discovers the hidden causal structure which generates the observations. This is a temptation which many psychologists, especially IQ-testers, have failed to resist, even eagerly embraced. Flynn protests the "conceptual imperialism" of g. He correctly insists that factor analysis (and related techniques, like item response theory) at most finds patterns of correlation, and these arise from a complicated mixture of our current social arrangements and priorities and actual functional or causal relationships between mental abilities. Factor analysis is helpless to separate these components, and gives no reason to expect that "factor loadings" will persist. Indeed, the pattern of Flynn-effect gains on different types of IQ test is basically unrelated to the results of factor analysis.

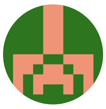
'But really the whole enterprise rests on circularities. It's mathematically necessary that any group of positively-correlated variables has a "positively loaded" general factor. (This follows from the Perron-Frobenius theorem of linear algebra.) A sub-test is "highly g loaded" if and only if it is comparatively strongly correlated with all the other tests; or, to adapt a slogan, positive correlation does not imply common causation. (Saying "Jack solved all the Raven's problems because he had high scores on many other tests which are positively correlated with scores on Raven's" is even more defective as an attempted explanation than attributing sleep to a dormitive power.) Since IQ test questions are selected to be positively correlated, the appearance of g in factor analyses just means that none of the calculations was botched. The only part of the enterprise which isn't either a mathematical tautology or true by construction are the facts that (1) it is possible to assemble large batteries of positively-correlated questions, and (2) the test scores correlate with non-test variables, though more weakly than one is often led to believe. Flynn does not make this argument, and some of his remarks suggest he still attributes too much inferential power to factor analysis, though he correctly says that it has contributed little to our understanding of the brain or cognition.

'After a century of IQ testing, there is still no theory which says which questions belongs on an intelligence test, just correlational analyses and tradition. This is no help in deciding whether IQ tests do measure intelligence, and so whether the Flynn effect means we are becoming smarter. If we accept Flynn's idea that intelligence is how well and how quickly we learn, an IQ test is an odd way to measure it. None of the tests, for instance, set standardized learning tasks and measure the performance achieved within a fixed time. At best they gauge the success of past learning, which could indirectly measure how well and how quickly people learn if we presume that the test-takers had similar opportunities to learn the material they're being tested on. Even then it would be confounded with things like executive function and current and past motivation. For instance, in 1998 Lovaglia et al. (American Journal of Sociology 104: 195–228) did an experiment where they took groups of college students and spent fifteen minutes creating a situation in which either the right- or left- handed students could expect to be better-rewarded for their efforts and abilities; the favored hand was randomly varied by the experimenters. This consistently made students in the favored group score about 7 IQ points higher on Raven's Matrices than those in the disfavored group. That is, a quarter of an hour of motivational priming can be worth a decade or more of the Flynn effect.'

Also: It would seem very few people in this latest spate of blogging have read the paper on causal vs. correlation effects [Glymour's paper on the Bell Curve — http://www.hss.cmu.edu/philosophy/glymour/glymour1998.pdf (PDF)] that he twice mentions in the blog. Skip to section 8 and read the political conclusions (anti-family, for starters), if you want your conservative blood to boil.

It's a critique of _The Bell Curve_ in terms of causal diagrams and factor graphs, and rather devastating I think. If you don't know what a Factor Graph is, Chris Bishop's book, ch. 8, will explain: research.microsoft.com/~cmbishop/prml/Bishop-PRML-sample.pdf [free online sample]

REPLY



johnfuerst

April 13, 2013 at 7:48 pm

Macrobius.

Cosma Shalizi is dishonest in his critiques. You should compare what he says — or more often implies — that g men and hereditarians say with what they actually say. Below is a list of books discussing intelligence and heritability. Nearly all of the positions in the books are granted by g men and hereditarians and nearly all claims are mutually consistent.

Making Sense of Heritability (2005) by Neven Sesardic; Measuring Intelligence: Facts and Fallacies (2004) by David Bartholomew; The g Factor: The Science of Mental Ability (Human Evolution, Behavior, and Intelligence) (1998) by Arthur Jensen

I have put these books online so you can google around and find them. As for your point above:

"For each question on each test, we can under suitable assumptions rank the value of that question, ordinal-ly, for each of the 7 individuals — and assuredly each question has economic value to them, since the higher their 'bundle of scores' subject to their 'intelligence constraint' the better adapted to life they are, which is a sort of utility. Therefore we have columns that are permutations of the numbers 1-7"

I don't find this to be an illuminating analogy. With regards to IQ testing, you can rank individuals by g scores because g represents a common property between subtest scores — g, unlike IQ subtest scores, is uni-dimensional.

"Also: It would seem very few people in this latest spate of blogging have read the paper on causal vs. correlation effects [Glymour's paper on the Bell Curve"

More likely is that we are familiar with the larger body of research:

http://www.gnxp.com/new/2007/01/02/iq-8594-academic-achievement/
http://www.sciencedirect.com/science/article/pii/S0160289699000082

But you would have to make a specific point. What issue, specifically, do you think is being overlooked?

"Skip to section 8 and read the political conclusions (anti-family, for starters), if you want your conservative blood to boil….It's a critique of _The Bell Curve_ in terms of causal diagrams and factor graphs, and rather devastating I think"

First, I don't think anyone here is a self identifying conservative. And second, the basic conclusions of the Bell Curve have been demonstrated repeatedly. It seems that your not familiar with the larger body of research. Read up on the background literature to get a sense of the issue:

Bock, Gregory; Goode, Jamie; Webb, Kate, eds. (2000). The Nature of Intelligence.
Ones, D. S., & Viswesvaran, C. (2002). Introduction to the special issue: Role of general mental ability in industrial, work, and organizational psychology.

REPLY

Macrobius

April 14, 2013 at 12:27 am

'I don't find this to be an illuminating analogy.' It's not really an analogy. It's intended to show that logically unitary g is equivalent to the IIA assumption per Ken Arrow's Impossibility Theorem — or if you like, equivalent to doing logistic regression on 1s and 0s if your test is scored in a binary fashion (which is a convenience not a necessity).

That is, it is intended to demonstrate something, not make an analogy or model to anything not already in the concept.

REPLY

Macrobius

April 14, 2013 at 12:03 pm

To make the construction a bit more clear, the matrix is just the answer sheets turned in by the 7 participants with 2000 questions — scored 0 for incorrect and 1 for correct answer. However, it is perfectly reasonable (if expensive) to have a panel of judges give ordinal scores to all participants on each question, and to use somewhat more complex questions that are more discriminating of total ability. In that case, the ordinal ranks in the columns are just the test scores.

If you think how algorithm evaluation is done for search engines you will see the point — in that case, the 'questions' are query terms and the answers are indeed scored by judges, for each of the algorithms tested, on a numerical scale that has more than 'correct' and 'incorrect' as choices — usually via a 'Mechanical Turk' sort of arrangement, with the human judges solving the HITs ('human intelligence tasks') of scoring the machines. I've simply inverted the machine testing and the humans here.

I'm not sure our commitment to what in search engine land and NLP land to 'binary relevance' of the answers, is such an essential element of intelligence test design.

REPLY

johnfuerst

April 14, 2013 at 3:52 pm

"Along the right hand column, write down their true IQ score — to the 'global intelligence' of each of the 7 test takers in fact known, but we wish to rank the individuals without knowing it using some computation based on their answers or perhaps additional information."

I don't know what you mean by "true IQ scores". You can derive g-scores with factor analysis. And then you can create a regression formula to predict g-scores e.g., g = .34a + .21b + .34c, where a, b, and c are subtest scores. If this is what you are saying — ok. But this isn't always how full IQ scores are calculated. As a result, this makes your statement "'IQ is just" confusing — to me, at least.

Also, the main discussion here concerns the computation of and meaningfulness of "true IQ scores" — or in your comparison "true voter preference scores" — from which to build the regression equation. You could try to calculate "true scores" in a number of different ways (e.g., averaging subtest scores). The real debate here is whether you can create uni-dimensional scores by which you can rank people in something called "intelligence" — or "voter preference (e.g., for text book defined neoliberal policies)". The g-men's point is 'yes' — because all subests share a common factor and score differences on this factor can be compared. Using your example, this would be similar to if voter preference scores positively correlated (e.g., people who supported neoliberal policy 1 tended to support neoliberal policy 2), allowing one to extract a p -factor and rank people in terms of neoliberalism. (Obviously, if voter preferences did not correlate — you could still rank people by averaging their scores and ranking the averages — or by selecting a prototypical question and then ranking individuals on that — but such a ranking involves more arbitrariness.)

REPLY

Steve Sailer

April 16, 2013 at 12:25 am

Dr. Shalizi is a little too impressed with his own IQ. His acquaintance with the field of psychometrics is mediocre at best, and thus he makes amateur mistakes motivated by his ignorance, animus, ideology, and arrogance.

REPLY

Dalliard *(Post author)*

April 17, 2013 at 11:41 am

I just reread Glymour's article after many years, and was again struck by the fact that he first presents a scorched earth critique of social science methods, and then goes on to argue that America's social problems can be solved by pouring tons of money into schools, social programs, etc. Glymour is not at all bothered by the fact that the policies he advocates cannot be supported *even* by those methods he disdains, like regression. Is he being satirical or does he just completely lack self-awareness?

Glymour's critique is unimpressive if you know the wider research which Herrnstein and Murray draw on. His alternative causal models are often highly implausible. For example, Glymour writes that shared family influences like mother's character, attention to small children, the presence of two parents, a scholarly tradition, a strong parental positive attitude towards learning, and where parents went to school may influence both IQ and economic/social outcomes, invalidating Murray and Herrnstein's causal model. However, we know from behavior genetic research that, firstly, all behaviors are heritable, often to a high degree, and, secondly, that environmental influences on IQ and other behavioral variables are generally not shared between adult siblings. See this classic article, for example. Murray also showed later, using the same data from the Bell Curve, how IQ predicts outcome differences between siblings. And what John said above.

Both Glymour and Shalizi think that they can invalidate entire research traditions by using clever conceptual arguments, without knowing anything about empirical findings in the field. You will be impressed by their arguments if you know nothing about the research they're attacking, but if you do, their erudite exercises in straw man slaying are rather tiresome.

REPLY

Macrobius

April 17, 2013 at 10:27 pm

I think you are on to something there, but you need to tune it. Pearl, Glymour, and by implication Shalizi not only overturn unitary g — they overturn 90% of the conclusions of Social Science. Is such a 'scorched earth' critique necessary? I would say yes — Science advance one funeral at a time. But in the mean time, it is well within the purvey of the victims of time's ever rolling stream to say 'I'm not dead yet' — but then they should say what they don't like about Glymour, Spirtes, Pearl etc — it *has* been done.

James M. Robins and Larry Wasserman 'Rejoinder to Glymour and Spirtes', say.

REPLY

Macrobius

April 18, 2013 at 11:31 pm

One (like Columbo) last point: Glymour, Spirtes, and Schenes say this, which is why I asked you if you believed Thomson's critique was devastating against Spearman or not:

'Spearman's inference to common causes from vanishing tetrad differences was challenged by Godfrey Thomson in a series of papers between 1916 and 1935.
In our terms, Thomson's models all violated linear faithfulness.' (p. 200, section 6.13)

That is, even in terms of Pearl's SCM and Glymour's critique, the Thomson model violates Glymour's 'faithfulness' premise, as he says in his own book. I can't help but think that this logical nit is worth following up.

REPLY

Johnathan cannon

September 19, 2013 at 1:29 pm

The Woodcock Johnson III is probably the best indicator of g, although the best available, culture-free, means of estimating g – is the RPM. In contrast to what you state here, even the older WJ-R is a better estimator of g, then the WAIS.

REPLY

Dalliard *(Post author)*

September 20, 2013 at 9:30 am

As indicators of g, all widely used multiple-ability batteries seem to be very similar. That is, their underlying g factors are very highly correlated with each other and the g loadings of their global/full-scale scores are similar. The RPM is not terribly "culture-free", and is probably a poorer indicator of g than tests with more diverse content.

REPLY

Johnathan cannon

August 15, 2014 at 12:17 pm

On second thought, I agree. The RPM seems rely heavily on a Working memory component, which is enough, to automatically disregard it as a 'culture free' measure of intelligence (perhaps, in a limited sense, 'context free', but certainly not 'culture free'). The Quantitative factor (Equation Balancing and Applied Problems) ,is interesting, as it seems to tap into some sub-component of g, that goes beyond what fluid intelligence measures – from my investigations, I believe some efficiency in searching LTM (associative heuristic). Those types of math problems require selecting from a broad set in LTM, and (time) ordering elements to form a meaningful relationship – and probably doing it very very fast, so that it doesn't consume limited resources in working memory capacity. This is in contrast with RPM problems, where the set has already been selected and arranged, but you have to discover the 'operator' that relates the elements. Although, some may be quick to attribute Gq with crystallized (learned) ability, an analysis of the chart above, argues against that idea, that Gq is significantly crystal loaded. First, it's pretty obvious that equation balancing and applied problem sub-tests, do not involve 'complex math', but rather elementary operations – basic 8th grade type math that most any subject can do. But more importantly the analysis-synthesis test almost correlates as well, with Applied Problems, as it does with Concept Formation……and actually according to the chart Applied Problems also correlates better with pretty much all other , non-crystallized measures (Woodcock Johnson Cog-g Scale), than Concept formation. So whatever small crystallized component Gq may be measuring, it is more than offset by it's ability to tap into the g factor, in ways that even the mighty fluid intelligence cannot. Long ago, the philosopher Immanual Kant wrote about the syntheticity of mathematics. I think it's time they listen to what he had to say.

REPLY

Johnathan cannon

August 16, 2014 at 12:57 pm

But with that being said why does the WJ IV standard battery measure Working memory twice, with story recall and numbers reversed, measures fluid intelligence twice, with number series and concept formation, and moved visual-auditory processing to extended battery, in light of it's relatively high g-loading, ability to tap glr, and greater relevance with respect to newer models of dual processing? Also there is no inclusion of anything mathy, like applied problems (which merely consists of very basic calculations), which by more recent accounts correlates significantly better with the general factor, and taps reasoning in a way that can't be assessed by fluid intelligence tasks (number series is a measure of linear reasoning, but not the same as measuring non-linear reasoning with LTM assocaitive/heuristics). What a shame, the new battery is clearly biased towards fluid processing

REPLY

jpz

November 24, 2013 at 5:17 pm

What's interesting, (as mentioned in this post) is that a few studies, indeed, have revealed that quantitative concepts (including applied problems and quantitative concepts) highly correlate with g, even more-so than fluid intelligence tasks – which has long been considered the best proxy of g. On second thought, this shouldn't be too surprising, as Gq problems are, firstly, more complex – requiring the handling of more variables in working memory (and thus underlying features of g), and secondly Gq problems are a better reflection of 'abstraction'. Gq performance relies heavily on associating information to elements outside of the problem. The particular step of relating information via LTM, based on subtle (implicit) similarities, may be a long-ignored hallmark of intelligence. Perhaps related to gq perfromance is the fairly high correlation of associative memory tasks with general intelligence. In fact, the WJIII suggests that the glr cluster can be used as a proxy for general intelligence.

REPLY

C. Bronze

November 25, 2013 at 10:35 pm

Yes, the math factor is separate from the fluid factor, and more loaded, too. This is confirmed by coGAT (although coGAT has validity issues, itself) designer David Lohman and WJIII Radex analysis. The WJ is a fine measure of intelligence based on the dominant CHC paradigm – as you mention it is one of the only tests to measure long term memory functions , which comprise a highly g-weighty, glr cluster. In contrast, to the abundance of claims, there is no test which taps into g directly. Every IQ sub-test is a measure of some degree of g + error . The WJ-III measures 9 broad factors, in order to get a solid estimate of g. Tests, like the Ravens are not as good at measuring g (though the raven's is valid as a proxy), simply because they are estimates based on a fewer number of sub-factors – the Raven's, for example, is restricted to a spatial and fluid factor (and perhaps Short term memory). And the lack of validity on certain tests, such as for the coGAT – suggests they may be largely measures of random skills, which may even inter-correlate to a large extent (because of their sub-tests sharing specific sub-factors), but nonetheless, are not good reflections of g.

REPLY

Joël Cuerrier

March 29, 2014 at 11:11 pm

Reblogged this on joelcuerrier.

REPLY

Neeraj

November 3, 2014 at 5:18 pm

Very nice blog post indeed. My question is somewhat related to it. DO you know any work that correlates "personality traits" with "g-factor"….? I know there is some correlation with openness…but is there any elaborate thinking on this line?

REPLY

Dalliard *(Post author)*
November 3, 2014 at 6:13 pm

Judge et al. meta-analyzed correlations between g and the Big Five personality traits, with the following results:

Conscientiousness -0.04
Agreeableness 0.00
Extraversion +0.02
Emotional stability +0.09
Openness +0.22

Openness has the most substantial association, partly because many openness items assess self-estimated intelligence. The slightly negative correlation between g and conscientiousness has often been found, but it may be an artifact of sampling bias.

In general, there appear to be all sorts of personalities across the entire IQ range, at least in terms of the Big Five, suggesting that the etiologies of intelligence and personality are mostly distinct. Multivariate behavior genetic analyses could shed more light on this, but I'm not aware of any.

REPLY

Dalliard *(Post author)*
August 22, 2015 at 7:37 am

I skimmed his/her comments, but I don't see any coherent argument that would be worth addressing. Most of the relevant issues have been discussed at length in this blog.

REPLY

Dalliard *(Post author)*
August 24, 2015 at 9:17 am

Kan's argument seems to be mostly about the shortcomings of the method of correlated vectors and the fact that non-g models can explain Jensen effects, too. Viewed in isolation, as a methodological critique, I don't really disagree with his argument. But when you consider the totality of evidence and the power of different theories to explain the facts at hand, I view the g theory as clearly superior. While it's possible to explain MCV findings in non-g terms, g theory explains the various findings very straightforwardly, something which cannot be said of competing theories. Moreover, SEM studies are also consistent with g theory, as I discussed here, so the critique of MCV is ultimately moot.

> "Swank"'s attack on Twin Studies, beginning following this sub-thread:

On the validity of the classical twin method, see these two recent papers: [1] and [2].

> "explained in statistical parlance is not the same as caused by.
> "However, since high heritability is simply a correlation between traits and genes, it does not describe the causes of heritability which in humans can be either genetic or environmental."

Heritability is the proportion of phenotypic differences between individuals caused by genetic differences between them. If heritability is estimated accurately, then it does have this straightforward causal interpretation. By definition, the causes of heritability cannot be environmental.

To argue that theoretical constructs like "additive genetic variance" cannot contain causal information because they (supposedly) aren't concrete enough, is to engage in arbitrary theoretical "legislation" of what is allowed in science. There is no one "correct" level of analysis of genetic causation. Certainly, genetic variance components must ultimately be reducible to molecular genetic mechanisms — as is being done in GWAS research — but the current lack of a more reductive account does not in any way invalidate behavioral/quantitative genetics.

> "Part of what is at stake between followers of Lewontin and Sesardic is whether or not VGxE and other components of variance are negligible or significant."

Yes, and theory and data point to those components being small in humans, just like in animals and plants. For example, GCTA research where the genetic and phenotypic similarity of unrelated individuals is compared shows that most of IQ heritability can be explained by the additive effects of common genetic variants. This indicates that interaction terms cannot be large because unrelated invididuals don't share environments except due to genetic reasons.

The mainly additive etiology of individual differences in complex traits is also theoretically and mathematically well nigh necessary in natural populations. This is because when there are lots of genetic and environmental influences, each with small effects on population variance — which is definitely the case with phenotypes like IQ — interactions cannot contribute much to population variance. This is true even when there are strong interactions at the individual level. That is, interactions at the level of gene action in an individual generally contribute to additive rather than nonadditive variance at the population level. This logic is explained in the context of epistasis in this paper. Large interaction components are generally found only in model organisms whose allele frequency spectra have been artificially reduced.

REPLY

Dalliard *(Post author)*

August 24, 2015 at 9:28 am

There are several genetic variants that are replicably associated with IQ and related variables. With sample sizes growing larger, many more genome-wide significant polymorphisms will be known shortly. Polygenic scores for IQ and educational attainment, computed for individuals based on GWAS results using relaxed significance thresholds, also replicate across samples.

REPLY

Dalliard (*Post author*)

August 24, 2015 at 1:57 pm

As to the Flynn effect, different eras and cultural contexts favor the development of different skills. If you want to call them intelligence, then the Flynn effect has increased intelligence, sure. However, if you want to compare the intelligence of different groups, you must have invariant indicators of intelligence, which are uncommon between generations but common between contemporaneous groups.

REPLY

Dalliard (*Post author*)

August 25, 2015 at 2:17 pm

1) Wealth and race differences

To statistically adjust for wealth differences between races is not a form of causal analysis. Because wealth differences between whites and blacks are vast, the net worth of the average white family is similar to that of black families that are exceptional, in the right tail of the black wealth distribution. From a hereditarian perspective, it's not meaningful to compare ordinary white families, with average genetic propensities for their race, to exceptional black ones, with exceptional genetic propensities for their race.

I am not familiar with Conley's research on this issue, but in her analysis of math test score differences in the CNLSY sample, Amy

Orr found that the race coefficient remained significant after parental wealth was included as a predictor, indicating that wealth did not statistically explain all of the test score differences between blacks and whites.

2) Flynn effect

If you want to think of skills at solving cognitive test items as intelligence, then the Flynn effect has increased people's intelligence. However, from my perspective the purpose of cognitive tests is not to test your vocabulary size, or skills at arithmetic problems or Raven's matrices, or whatever other kinds of items an IQ test may have. Rather, the purpose of tests is to measure latent abilities, such as general intelligence, of which test scores are just unreliable, epiphenomenal, culture-bound indicators. The extent, if any, to which those latent abilities have improved across generations is unclear, but it's certainly much less than the increases in observed test scores.

3) Neural plasticity

Neural plasticity is a buzzword that doesn't mean much of anything. The fact that the brain is malleable is not news. We have always known that humans are able to *learn* things, which changes your brain. Reading this comment changes your brain. Everything changes it. None of this means that we have, say, the means of making a stupid person smart. Instead, individual differences in intelligence remain highly stable across the life span, mainly due to the life-long persistence of genetic effects.

4) Hegel on Africa

I don't know what your question is.

REPLY

Dalliard *(Post author)*

August 25, 2015 at 3:27 pm

The shift from hunting and gathering to agriculture led to a lower standard of living for the average man, yes, but most of black

Africa had adopted agriculture long before significant contacts with Europe.

REPLY

Dalliard *(Post author)*

August 26, 2015 at 3:34 pm

It's not a topic I know much about, but John Reader's "Africa: A Biography of the Continent" has some good discussion of the peculiarities of agriculture in Africa.

REPLY

Dalliard *(Post author)*

September 9, 2015 at 5:59 am

Ben, I didn't publish your recent comments because they consist just of long quotations from miscellaneous sources. Please stay on topic and be concise.

REPLY

Dalliard *(Post author)*

September 12, 2015 at 8:05 am

Ben, this is the comments section for my article on the g factor. If you want to promote your views on eugenics, start a blog of your own.

REPLY

werkat

June 5, 2022 at 5:58 am

I have three more examples. The Differential Aptitude Test (DAT), The General Aptitude Test Battery (GATB), and The Armed Services Vocational Aptitude Battery (ASVAB) were developed in response to the influence of multiple abilities theorists like Thurstone in order to measure specific abilities (s1… sN) and to make specific predictions about job or training performance. All ended up with a strong general factor and with s having miniscule non-g validity for predicting job performance and training success: https://sci-hub.ru/https://doi.org/10.1111/j.1744-6570.1991.tb00961.x

REPLY

9 Pingbacks

1. Facets of the Big Five personality factors | Entitled to an Opinion
2. linkfest – 04/07/13 | hbd* chick
3. What is AGI? | Machine Intelligence Research Institute
4. g | Technology as Nature
5. In a variable world, are averages just epiphenomena? | Dynamic Ecology
6. Against Unidimensionality – spottedtoad
7. Hunter Avallone is Wrong About Race and Intelligence: A Response to 'The Truth About Black People'
8. Response to Richardson and Norgate (2015): "Does IQ Really Predict Job Performance?"
9. Taleb: IQ is largely a pseudoscientific swindle (Wrong again!) | greyenlightenment.com

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment *

Name *

Email *

Website

Save my name, email, and website in this browser for the next time I comment.

Post Comment