CPC, 1997, 16 (6), 788-812

AUTHOR'S RESPONSE

The rise and fall of Spearman's hypothesis

Peter H. Schönemann

*National Taiwan University**

To start off, I wish to thank all commentators for taking the time to reflect on the issues I raised, and for agreeing to comment on them. In this reply, I shall first address technical issues raised by Horn, Kadlec, Maraun, Millsap, and Steiger, and then move on to the broader concerns of the other commentators.

Technical issues

Maraun: « It is evident from the geometry of the 2-component mixture distribution that ... the orientation of d, w_1, and z_1 in R^p are not restricted in *any* way » (Maraun, 1997, emphasis in the original).

Professor Maraun's Figure 1 consists of 3 panels, each depicting two bivariate normal equidensity contours. The matrices defining them, Σ_1 and Σ_2, and the mean vector μ_2, are kept fixed. The mean vector μ_1 increases from panel 1a to panel 1c, so that the length of the difference vector d also increases. Maraun then finds that the dominant eigenvector for the pooled group approaches collinearity with d as d increases, in accordance with my Level I argument in Schönemann (1985). His pictures are an improvement over my own Figure 1, in that they stress that the Level I case, in contrast to Level II, in no way depends on the positivity of the two within covariance matrices.

* On leave from Purdue. Please address correspondence to Peter H. Schönemann, Department of Psychological Sciences, Purdue University, West Lafayette IN 47907, U.S.A. (e-mail: phs@psych.purdue.edu).

Maraun then turns to Level II and asks: « Does there exist such a dependency of the case II of the Spearman 2 correlations? » His laconic answer: « I think it is evident from the geometry of the 2-component mixture distribution that it does not » (Maraun, 1997). This leap of faith does not suffice to invalidate my Level II banana argument.

The problem with Maraun's ellipses is that only Σ_1 is positive (has positive elements throughout), but not Σ_2. What does this imply for Black/White IQ data? It means that the structure of 'intelligence' tests would have to be drastically different for Blacks versus Whites. However, this is not what the data say. Nor is it what Jensen would want them to say, because it would ruin his chances of comparing Whites and Blacks on 'intelligence'. As he has repeatedly emphasized: « One of the most striking and solidly established phenomena in all of psychology is the fact of ubiquitous positive correlations of all mental ability. This is simply a fact of nature » (Jensen, 1980, p. 249). This means that such correlation matrices are positive. So far as I know, no-one but Maraun has taken issue with Jensen on this point.

In view of this fact of nature, Maraun's counterexample, which includes a non-positive covariance matrix, falls short of providing « an appropriate analytic framework for the consideration of an empirical issue » (Maraun, 1997).

Steiger: « the 'Spearman hypothesis correlation' is definitely not an inevitable 'artifact' if it occurs in a multivariate population that is a mixture of two multivariate normals. » (Steiger, 1997).

Professor Steiger acknowledges his debt to Loehlin (1992), who had previously questioned whether my Level II explanation is « a very felicitous modeling of racial differences in the United States » (p. 262). Following up on a 'thought experiment' by Lewontin (1970), he arrived at a mixture of two normal distributions: « Now, to our immediate concerns. Ask yourself, what will be the distribution of the plant heights, in Lewontin's (1970) experiment ...? Will it be normal? No. ... Within each group, the distribution may well be normal ... but the combined population of heights will be bimodal, with each subpopulation normally distributed around its own mean. » (p. 261).

Loehlin did not adduce any data in support of his bimodality thesis, and perhaps this may not be necessary. Let us simply ask whether a mixture of two normals will always be bimodal. Common sense suggests that this should depend on the size of the mean difference. For Black and White IQ data, the literature shows that the mean difference is roughly one standard deviation.

Table 1 shows the average of two standard normals one standard deviation apart. The pooled distribution is nowhere bimodal. It is virtually indistinguishable from a normal distribution, except that it is slightly more platycurtic. For weighted mixtures, one expects some skewness (negative if the higher group is larger), which would translate into pear shaped rather than exactly elliptical bivariate plots. None of this invalidates the Level II banana argument. In short, in the context of Black/White IQ data, Loehlin's bimodality specter is a red herring.

Table 1
Equally weighted mixture of two normal distributions one standard deviation apart

Midpoint	-3.0	-2.5	-2.0	-1.5	-1.0	-.5	.00	.05	1.0	1.5	2.0	2.5	3.0
n(0,1)	.00	.02	.05	.13	.24	.35	.40	.35	.24	.13	.05	.02	.00
Mixture	.01	.03	.07	.15	.24	.32	.35	.32	.24	.15	.07	.03	.01
Difference	-.01	-.01	-.02	-.02	.00	.03	.05	.03	.00	-.02	-.02	-.01	-.01

Turning now to Steiger's simulation, one finds that both his ellipses, in contrast to Maraun's, agree with Jensen's fact of nature that IQ correlations tend to be positive. In the planar correlation case this means that both dominant eigenvectors within groups are collinear with (1, 1) on the main diagonal of the positive quadrant, and hence are collinear with each other. He then imposes orthogonality on PC1 and d by constructing a mean difference vector which is orthogonal to the dominant eigenvector of one of the two subpopulations. In the planar case this is $d' := (1, -1)$. Regardless of the number of dimensions, such an orthogonal mean difference vector would always have to contain roughly as many positive as negative elements, since it is known that the dominant eigenvector of a positive matrix can always be chosen positive (see below). However, this is not what Black/White IQ data show. For example, in Jensen's (1985) Table 11, all 11 mean difference vectors d are positive. This means, concretely, that on average, Whites outscore Blacks on most 'intelligence tests' at the present time – for whatever reason. This fact never was in doubt. What is in doubt is whether it teaches us anything about the existence and genetics of g.

As a result of Steiger's choice of the mean difference vector, it further becomes impossible to distinguish between High and Low groups. Group A outscores B on some tests, and Group B outscores Group A on other tests. This is rarely a problem with real Black/White IQ data.

Since neither Loehlin, Steiger, nor Maraun adduced any empirical evidence in support of their conjectures, it may be instructive to include at least one 'real' data set (Table 2). All marginals, while not *exactly* normal, are unimodal. Most are positively skewed, because the lower scoring group is larger. While the contours are not *exactly* bivariate normal in any of the three populations, they are approximately so, tending towards pearshape as a consequence of the uneven split. Most importantly, all three correlations are positive, and so is the mean difference vector. This, I claim, is the typical case, and it is all that is needed to make my Level II banana argument go through.

Before moving on, I would like to touch briefly on two minor issues Steiger has injected: One concerns the exact nature of the boundary dividing the High and Low groups. Kadlec has taken care of this point decisively. Steiger has also raised doubts about my computer generation of « positive manifolds » devoid of g:

« A pxp 'positive manifold' for the pooled sample was constructed as C =3D TT' + sI, where the elements t_{ij} of T were uniformly [0,1] distributed random numbers, I is the pxp identity matrix and s is a scalar 0 » (target article).

Steiger worries that this « method of Monte Carlo generation of data 'devoid of g' is questionable, because it can produce a correlation matrix that is very similar to one where there is a strong g » (Steiger, 1997).

He fears that g (in Spearman's sense of hierarchy, not Jensen's PC1 trivialization) may have crept in through the back door held open, so to speak, by Thomson's Sampling Theory. Thomson (1916) had pointed out that a pxp correlation matrix constructed from a very large number of factors m, with 0, 1 loadings throughout, will tend towards hierarchy as the ratio q:= m/p goes to infinity. In Schönemann (1981, p. 364), I gave a concrete numerical illustration and some simulation results: For a sample size N = 200, 9 tests, and 20 factors (i.e., q = 2.2), a maximum likelihood factor analysis program was unable to reject the null hypothesis of only one common factor in 3 out of 4 cases, for 1,000 factors (q = 111.1) in 4 out of 6 cases, and for 2,000 factors (q = 222.2) in 5 out of 5 cases.

Table 2
Within Group and Pooled Distributions for Head Start Data (MRT)
(Variables: Vocabulary vs. Arithmetic. Groups: Black vs. White)

BLACK:	10	15	20	25	30	>30	sum
>30							
30		1	1				2
25		1	7		1		9
20		12	8	4			24
15	2	72	14	2			90
10		1					1
sum	2	87	30	6	1		126

	mean	st. dev.
voc:	15.1	3.4
arithm:	14.3	3.3
correl:		.43

WHITE:	10	15	20	25	30	>30	sum
>30				3	3		6
30			2	4	2		8
25			2	4	6	1	13
20			4	6	1		11
15		7	9	2			18
10		1					1
sum	0	8	17	19	12	1	57

	mean	st. dev.
voc:	21.1	8.1
arithm:	21.1	4.9
correl:		.61

POOLED:	10	15	20	25	30	>30	sum
>30				3	3		6
30		1	3	4	2		10
25		1	9	4	7	1	22
20		12	12	10	1		35
15	2	79	23	4			108
10		2					2
sum	2	95	47	25	13	1	183

	mean	st. dev.
voc:	17.0	6.0
arithm:	16.4	5.0
correl:		.66

However, in my target article simulation, q was nowhere near infinity, it was 1. Since Steiger had brought it up, I looked into the matter. Table 3 summarizes the results. The hypothesis of only one common

factor was uniformly rejected even for relatively small sample sizes (N = 100 and N = 200). I ran two replications per parameter configuration. In short, Steiger's 'caution' is groundless.

Table 3
Thomson's Sampling Theory

Sample Size:				N = 100		N = 200	
Population Correlation	No. of Variables	No. of Factors	df	Chi-Square	Prob	Chi-Square	Prob
.5	10	1	35	65.7	.00	137.3	.00
				82.0	.00	116.9	.00
		2	26	44.5	.01	75.9	.00
				40.6	.03	75.2	.00
.6	10	1	35	190.7	.00	284.1	.00
				213.7	.00	239.2	.00
		2	26	118.2	.00	146.6	.00
				121.7	.00	150.0	.00

In conclusion, it should be noted that none of the critics so far questioned the stringency of my mathematical argument in Schönemann (1992). What they did question is the empirical validity of assuming multinormality for the pooled group at Level II. I remain unimpressed by the challenges mounted by Maraun and Steiger via Loehlin against this assumption. The problem seems to be that none of these critics have carefully read Jensen's (1985) seminal BBS article, or looked at Black/White IQ data.

Kadlec: « Are the demonstrations still not realistic enough for people of Jensen's ilk? » (Kadlec, 1997).

Professor Kadlec's commentary picks up where Steiger's left off: Steiger had raised questions about the effects of violations of multivariate normality and the nature of the boundary between the high and low group. Kadlec – though presumably unaware of Steiger's commentary at the time she drew up her(s – answered them both in a straightforward, decisive manner.

The upshot is, yes, there is a difference between cutting a multivariate normal into two pieces, as assumed in my formal proof of Level II, and selection, which geometrically amounts to constructing an envelope of two multinormals within groups: If there is overlap, a cutting plane will misclassify some members of both subpopulations, as also Steiger noticed with some concern. Kadlec showed that this concern is unwarranted, because the only difference is that cutting compared to selecting will exaggerate the length of the mean difference vector. This in no way invalidates the Level II banana argument.

What remains intact under this change of perspective are not just the positive Spearman hypothesis correlations, but also, more importantly, the unequal splits effect, which Maraun and Steiger completely ignored. As already noted, this problem would have to be faced by anyone wanting to challenge my Level II argument.

Kadlec also studied the effect of substituting discrete variables for the normal variables assumed in my proof. The effect is again negligible. It should be noted, in this connection, that the truism that mixtures of normals are no longer normal could easily have been circumvented by simply replacing normal variables with binomial variables. Under suitable conditions, mixtures of multivariate binomials remain multivariate binomial, and normals and binomials are very difficult to tell apart in practice as the number of underlying Bernoullis increases. The reason one prefers to assume normality in formal proofs is simply that the normal distribution is more tractable than the binomial.

I thank Professor Kadlec for her straightforward treatment of two side issues which also intrigued other commentators.

Interlude 1: The empirical background of Spearman's hypothesis

Before any discussion of Spearman's hypotheses can become productive, at least three basic empirical facts about Black/White IQ data must be acknowledged: (a) both within covariance matrices tend to be positive, and the pooled matrix almost always is, (b) the mean difference vector is almost always positive, and as a result, (c) the pooled distribution usually is sufficiently elongated to render the banana argument applicable.

However, at this point, we face one more empirical fact: (d) The degree of collinearity as measured by the cosines varies monotonically with the sample sizes of the two subgroups.

Interlude 2: Positive matrices and Perron's theorem

« Schönemann (1992, p.223) invokes Perron's theorem in support of part b of his theorem. Perron's theorem ... gives certain properties of the eigenvalue decomposition of a square matrix that contains only positive elements. The relevance of this to the second part is hard to see » (Dolan, 1997, p. 323).

If it were not for this fourth empirical fact about unequal splits, multinormality would not be needed at all. To explain the remaining three facts (a)-(c), all one needs are some elementary results from the theory of non-negative matrices:

« SOME DEFINITIONS

Let A = (a_{ij}) be a matrix with non-negative elements a_{ij}. We call A *positive* and write A > 0 if each of the elements is strictly positive. Otherwise we write A ≥ 0 and call A *non-negative.* » (Cox & Miller, 1968, p. 119).

« For non-negative square matrices, the main result is the theorem of Perron and Frobenius which states, broadly speaking, that a non-negative square matrix has a maximal non-negative eigenvalue which is not exceeded in absolute value by any other eigenvalue and corresponding to which there is a non-negative eigenvector. » (Cox & Miller, 1968, p. 118). For positive matrices, this vector can be chosen positive.

This means that the dominant latent vectors of all « positive manifolds » lie in the positive quadrant, octant, and in general, 2^D-tant, together with the positive mean difference vector. This case corresponds to the three well-known empirical facts (a)-(c). Hence, the cosine (PC1, *d*) is apt to be positive.

The important point is that approximate collinearity of the 3 PC1s and, to a lesser extent, positive Spearman Hypothesis correlations, are induced by the positivity of the two within-group covariance matrices, whatever the exact shape of their equidensity contours may be. The positive mean difference vector further stretches the roughly ellipsoidal envelope. This is the salient fact behind the Spearman hypothesis artefact, not the nature of the boundary between both groups, which fascinated some critics. It is also the gist of my qualitative banana argument.

Approximate multinormality explains empirical fact (d).

In contrast to Loehlin, Steiger, and Maraun, Dolan (1997) did question the stringency of my mathematical proof in Schönemann (1992). He correctly pointed out that, for the proof to go through, the partitioning plane must be orthogonal to the PC1, not the centroid, as mistakenly stated in the proof (p. 221). That is, the equation for the cutting plane is $\Sigma_k v_{1k} y_k = 0$, not $\Sigma_k y_k = 0$. While the PC1 and the first centroid coincide in the planar correlation case, they no longer do so in the higher dimensional case. However, both the graph and the proof left little doubt about what was intended. As Dolan himself noted, once the vector of 1's defining the centroid is replaced by the largest eigenvector defining PC1, then « the required collinearity » (p. 322) follows.

While this flaw certainly bears correction in a formal proof, it also bears mention that the difference between the PC1 and the first centroid is minute in practice. Hence this point in no way invalidates the main conclusion, as Dolan makes it sound. For example, in a simulation with positive gramian random matrices of order 20, the median cosine was .992, corresponding to an angular separation between 7° and 8° (see Table 4B). Thus, for positive matrices, the PC1 and the first centroid are virtually collinear – as was widely appreciated in the days of Thurstone. This also has implications for the congruence coefficient lore, to which I shall return in my reply to Garriga-Trillo.

To answer Dolan's question about the role of Perron's theorem in the sufficiency part of my proof: « Since d is an eigenvector of Σ and dd', it is also an eigenvector of Σ_W. In particular, if Σ_W is positive, as assumed, then d is the unique eigenvector associated with the largest latent roots of Σ, Σ_H, and Σ_L, (Perron's Theorem), which proves (b). » (Schönemann, 1992, p. 222f). That is, Perron's theorem is needed to ensure a *unique* solution.

Millsap: « I must agree with Schönemann's fundamental point, and I find it surprising that the point is still controversial » (Millsap, 1997).

Professor Millsap extends Guttman's (1992) argument from one common factor to more than one common factor. This is useful, at least in theory, because eventually even Spearman lost faith in his tetrad difference condition meant to ensure the existence of exactly one common factor, since it was virtually never met by the data. As Guttman (1992) has stressed, to lend credence to Spearman's Two Factor Theory, it would have to hold for *all* 'intelligence' data: « Spearman ..., of course, was taken aback by the fact that there is no g that will satisfy conditions 1 and 2 *for the entire universe of mental tests*" (p. 182, emphasis mine). I underline this point because it is often glossed over. If hierarchy were satisfied only occasionally, it would prove nothing about the existence

of g, because there would be no basis for claiming that the 'g"s derived from different batteries « are the same » (see also Horn's commentary on this point). The same problem persists in Jensen's trivialization of Spearman's theory, which substitutes PC1 for g. Instead of appealing to universal hierarchy as evidence that the g's from different batteries must all be the same (Spearman's Two Factor Theory), Jensen's disciples now point to 'congruence coefficients' (cosines) as evidence that PC1's from different batteries are all the same. I shall return to this issue in my reply to Garriga-Trillo.

Spearman, in his time, tried to fix his factor theory by introducing group factors. As Guttman (1992) observed, this was actually a veiled concession that g does not exist:

« Spearman did not really resuscitate g, despite the last chapters in his *The Abilities of Man* (Spearman, 1932). He gave no necessary nor sufficient conditions – neither algebraically or contentwise – for doing so. Instead, Spearman was guilty of setting the example – followed by so many factor analysts – of not being able to give up the terminology of the nonexistent g, and of sacrificing falsifiable algebraic formulations and clarity of conceptualization in trying to save g » (p. 183).

Since the multiple factor generalization makes no falsifiable prediction of any kind, it cannot prove anything about the existence of g. To strengthen it, one might single out one salient 'first factor' and demand that it affect all tests positively. Perron's theorem guarantees that such a factor always exists if the correlation matrix is positive. This is essentially the paradigm Jensen advocates for principal components. He takes out m components, leaves the first intact and then rotates the remaining m-1 to simple structure. Since positivity of the test correlations can be trivially ensured by test construction, this amended paradigm is still not falsifiable. To strengthen it further, Millsap adds the strong factorial invariance requirement, that the covariance structure remains invariant across subpopulations for the same battery. This is a falsifiable condition, but it is much weaker than Spearman's demand that hierarchy persist across all *batteries*.

To achieve his generalization from one to m common factors, Millsap invokes six conditions which vary in intuitive appeal: (a) postulates that the group differences in the means are largely a function of the mean differences on the first factor, (b) that the 'first factor' accounts for some portion of the variance in each variable, and (c) that the loadings on the 'first factor' are variable. The other three conditions are: (i) the factor loadings are invariant across groups, (ii) any group differences in means are determined solely by differences in the common factor means, and (iii) that all loadings are positive. Thus Millsap's conditions ensure that the correlation matrices in both groups are the same and positive, which is what one usually finds for IQ data (approximately, not exactly). The purely technical condition (c) is needed if one wants to assess collinearity with correlations. Since this is not possible in the planar correlation case, and more generally, when all elements of the dominant eigenvector are equal, I suggested that collinearity be assessed with cosines. As it turns out, condition (c) also helps explain departures from perfect collinearity (see below).

Since Millsap's 'first factor' is a close relative of the PC1 in 'positive manifolds', it should be possible to reduce the number of his conditions by direct appeal to Perron's theorem about positive matrices. In particular, (a) the mean difference vector is positive, (b) both within group matrices are positive, and (c) both are equal, should suffice to do most of the work.

They are only slightly stronger than my conditions for the banana argument. Hence it is not surprising that Millsap arrives at similar results for factors as I did for components. However, it is not clear to me how this approach can be amended to account for the unequal splits prediction.

Horn: « It is neither good science nor good ethics to proclaim that an outcome that is a result of mathematical constraint is entirely empirical evidence in support [of] an hypothesis stipulating that Black persons ... are inferior in respect to a highly valued attribute, intelligence (Horn, 1997).

Professor Horn's comments partially overlap with Millsap's, since both authors treat the Spearman hypothesis in the context of multiple factor analysis and selection, Horn's being the more informal version. Since Horn and I agree on the basic point that Jensen's positive correlations are *not* an empirical result, but rather an inevitable outcome of mathematical constraints (viz. positivity induced by test construction), the remaining points of disagreement between us are minor:

(a) It is not correct that I assume 'explicitly' factorial invariance at Level I. Far from it. My Level I explanation does not even require positivity, let alone factorial invariance. The only thing needed for Level I is a sufficiently large mean difference vector. The within group covariance structures are completely unconstrained, as illustrated in Maraun's Figure 1c.

(b) After switching to principal components, Horn states: « When the g-factor model does not fit data ... the correlation between the differences between the means for groups 1 and 2 and the PC1 coefficients will be reduced ». This is incorrect. My claim that, in the exact multinormal case, all three PCs and d are *perfectly collinear*, in no way depends on the factorial structure of any of the three covariance matrices. This is precisely the advantage of treating the problem in terms of components rather than factors. Rather, the correlations will be reduced if the components of the mean difference vector d vary, so that it no longer lies on the main diagonal of the positive 2^D-tant, as it does in the exact multinormal case. Recall that this was one of the conditions Millsap required for his multiple factor model.

(c) I fail to see why my sliced banana argument is « a good deal more complex » than his algebraic demonstrations.

Notwithstanding these minor points of disagreement, I am pleased that Professor Horn, who has extensive research experience in this area, agrees with me on the fundamental issue, that Jensen's Spearman hypothesis correlations are artefacts (aka artifacts).

Broader issues

Hay: « Schönemann ... has contributed to such charlatanism by a highly selective and outdated review of general intelligence and its genetic analysis » (Hay, 1997)

Professor Hay is concerned that the collapse of Jensen's claims about Spearman's hypothesis may jeopardize the g concept more generally, together with all heritability claims that surround it. I agree with him. As Wechsler (1939) observed: « As to the existence of g there seems to be no possibility of doubt. Psychometrics, without it, loses its basic prop » (p. 8).

The reason why Jensen received so much applause when he exhumed Spearman's hypothesis was probably because it seemed he had put the notion of 'intelligence' back onto a firm empirical footing, after it had eroded during Thurstone's tenure. Since my results threaten to remove the g prop once again, it follows that they also threaten all heritability claims surrounding g.

I also agree with Hay that I have « done much more in [my] commentary [presumably meaning: target article, PHS] beyond dealing with a technical issue in psychometrics. He has used this opportunity, especially in the Discussion, to make much more general attacks upon the concept of intelligence, of behavior genetics and even of the peer review process! » (Hay, 1997; exclamation mark in the original).

However, a point of disagreement that remains between us concerns the need for technical arguments in discussions of *g* and heritability claims. As I already indicated, I doubt that such discussions can be fruitful in a technical vacuum.

A case in point is Holzinger's h2. To Hay (in press):

« It seems incomprehensible in 1997 that Schönemann can actually claim his recent discovery that 'one of the most popular 'heritability estimates' is unsound (Schönemann, 1993)'. (p. 23). He fails to note that it was ironically Jensen (1967) himself who raised the question of the inconsistency among the various traditional estimates (including that developed by Holzinger to which Schönemann refers). »

Actually, in Schönemann (1993, p. 60), I did give Jensen credit for having noticed that « the precise nature of the inadequacy of the [h^2] index has remained conceptually obscure » (Jensen, 1967, p. 149).

To go beyond this first step and shed light on the exact nature of the inadequacies, some technical work was needed. Jensen had pointed out that h^2 and Nichols' HR often do not agree (usually, HR $>$ h^2). This does not prove that there is anything wrong with h^2: The two answers, '4' and '5' to the riddle $2+2 = ?$ do not agree either, yet one of them, 4, is perfectly correct.

By attending to technical details which Jensen ignored, I was able to show that Holzinger's h^2 cannot possibly estimate the narrow heritability ratio ($=: v(a)/[v(a)+v(e)]$, where $v(a) :=$ additive genetic variance, and $v(e) :=$ environmental variance), for the simple reason that h^2 is not a function of $v(e)$, but only of $v(a)$ and measurement error variance. So far as I know, Jensen has never claimed to have shown that. Jensen (1987) tends to dismiss such technical demonstrations as « mathematical machinations (p. 387) ». In contrast, I believe they are indispensable if elementary statistical problems are not to remain conceptually obscure forever.

The widespread sentiment that technical competence is, at best, a dispensable luxury, and, at worst, a nuisance, has probably contributed to the delay in uncovering this specific inadequacy of h^2 (since 1929). I fail to see how this can be cause for celebration in the behavior genetics community.

Hay also takes me to task for confounding "intelligence" and scholastic aptitude: « I thought that the low correlations of intelligence test performance with University [sic] success was one of the standard first year exercises, used to show the limitations of correlations » (Hay, 1997). I confess that I did not know this. Apparently neither did Estes (1992), who still believes that « Intelligence tests ... are excellent predictors, ranging from school to a wide variety of occupations » (p. 278).

As long as we lack a generally agreed upon definition of 'intelligence', it remains a matter of taste which tests we call 'intelligence tests'. The SAT, for example, is a direct descendent of the Army Alpha test, which usually rates a bona fide 'intelligence test'. No less an authority than a former president of ETS, Chauncey, stubbornly clung to the belief that they are essentially the same kind of test: « Intelligence tests and scholastic aptitude tests have the same purpose: to estimate the capacity of the student for school learning ... For all practical purposes, and in all their school uses, they are the same kind of test. » (Chauncey & Dobbin, 1963, cited in Owen, 1985, p. 200). The reason why I focus on college entrance tests is simply that we have more solid evidence about their predictive validities than for the commercial IQ tests.

Humphreys: « Professor Schönemann's rejection of the significance of Jensen's confirmation of the Spearman hypothesis is well justified. ... However, Schönemann's failure to distinguish between this limited finding about differences between blacks and whites and Spearman's conception of a general factor of intelligence is serious error » (Humphreys, 1997).

I am grateful to Professor Humphreys, who is known for his stern standards, that he considers me at least half right. However, I want to emphasize that I never questioned the *significance* of Spearman's hypothesis. Given its social implications, one might regard it as one of the greatest discoveries of this century, or, alternatively, as one of the greatest hoaxes of this century. Either way, it does not lack significance.

Nagoshi: « I am in complete agreement with his procedures and conclusions in formally testing this assertion [i.e., Spearman's hypothesis, PHS]. »

Nagoshi enjoys the distinction of having been one of the few people who early on suspected that Jensen's Spearman hypothesis results may be artefactual.

The empirical results he cites point in the opposite direction of the conventional interpretation of Jensen's correlations: DeFries et al. (1982) found that « generational group differences across the cognitive abilities tests were highly correlated with the *g*-loadings of the tests » (Nagoshi, 1997). As Nagoshi points out, if these findings were not artefacts, then they « would argue that those abilities that are particularly affected by environmental influences are, in fact, the most *g*-loaded » (ibid.) – the exact opposite of what Jensen and his followers believe.

This suggests, at the very least, that things are not quite as simple as we once were led to believe. It also suggests that some of the far reaching claims about the genetic inferiority of certain ethnic groups and related concerns about the nation's economic welfare may have been premature.

Hirsch: « ... there are no genetic reasons to predict that, if one group has lower test scores than another in one generation, a similar disparity will necessarily be manifested by their children in the next generation, irrespective of the conditions of development » (Hirsch, 1997).

Professor Hirsch has been an indefatigable critic of Jensen's genetic claims ever since Jensen's (1969) « How Much » article first appeared. For this he deserves credit, for at least two reasons: (a), as one of the founding fathers of the behavior genetics movement in the 60's, he could be forgiven if he had been more lenient towards exaggerated heritability claims, and (b), as a geneticist who actually *does* know something about genes, he could have had a much more tranquil life had he chosen to chime in with Jensen's Eugenics Revival choir.

Perhaps it should be noted that, by now, the term 'Spearman's hypothesis' has acquired three different meanings: (a) the Level I interpretation of Jensen's loose definitions, (b) their Level II interpretation, and (c) the Two Factor Theory which, of course, is also a hypothesis due to Spearman. Ideally, all three should be carefully distinguished.

Sternberg and Grigorenko: « Ability tests have continued to exist in roughly the same form since the beginning of the twentieth century. The lack of progress seems to give great pride to those who create the tests » (Sternberg & Grigorenko, 1997).

Most conventional IQ tests are descendents of Binet's test which was aimed explicitly at school children. Binet was appalled at the misuse to which his tests were put in the United States. He never bought into the heritability muth with which it quickly became intertwined during the heydays of the immigration debate (see, e.g., Chase, 1980; Allen, 1997). One consequence of this shift in purpose was that IQ tests tend to overemphasize elementary problem solving skills under time pressure. These are at best only a minor part of the cognitive demands adults face in their daily lives which, presumably, is one of the reasons why the predictive validities of IQ and college entrance tests are so low.

Some independent minds have begun to rally against this legacy of the eugenics era, stressing that, if we are ever to make any real progress, then a new start may have to be made, even if one of the casualties is Spearman's *g* and the heritability lore surrounding it. Professor Sternberg is one of the leaders of this reform movement. It is refreshing to hear him air his doubts about the present state of the art in mental testing.

Sternberg and Grigorenko single out one facet of the stagnation in this field: No problem, no matter how trivial, ever gets settled. There are probably several reasons for this. Sternberg and Grigorenko point to confirmatory bias as one possible factor, and to the popularity of least squares techniques as another. It is well known that these can been used to confirm virtually anything, as long as the number of estimated parameters is large enough. Procrustes procedures are one example. I would add LISREL type programs, MDS, and, of course, exploratory factor analysis. All these procedures invite mindless abuse and, as a consequence, have rarely produced non-trivial empirical discoveries. Having contributed to the development of some of these procedures myself, I hasten to add that it is not the methods that produce junk science, but the people who misuse them.

This trend has gathered momentum since computers have now made it possible to simulate deep thought. More than 25 years ago, Andreski (1972) forecast the eventual course of events with uncanny perspicacity. He noticed that « The chief advantage of the mechanical application of routine techniques is that it permits a massive production of printed matter without much mental effort » (p. 109). As a result, the creators of such routine techniques rise to become cult figures held in high esteem: « The hordes of devotees, whose lack of brains would have debarred them from intellectual occupations in more civilized times, naturally adore the apostle who has enabled them to make an easy living by posing as scientists » (idem, p. 152).

The same dynamics, Andreski reasoned, also encourage muddle-headedness and discourage critical thought, « because clear and logical thinking leads to a cumulation of knowledge ... and the advance of knowledge sooner or later undercuts the traditional order. Confused thinking, on the other hand, leads nowhere in particular and can be indulged indefinitely without producing any impact on the world » (idem, p. 90).

Garriga-Trillo: « ... a final reason for not accepting the clear cut mathematical deductions presented in Schönemann's article might be found in a specific aspect of social differentiation: racial prejudice » (Garriga-Trillo, 1997).

In her commentary, Professor Garriga-Trillo touches a very sensitive nerve: Does anyone seriously believe that sheer incompetence and bad luck suffice to explain why it took 25 years before someone (Wilson, 1928) noticed that the factor model is afflicted with a serious defect which defeats the very purpose it was supposed to serve, or why it took several decades for someone (Kamin, 1974) to notice that there was something strange about Burt's MZA data, or why it should have taken me 12 years to lay the Spearman hypothesis myth to rest, albeit in foreign soil? Might there be other artefacts in the closets of our mainstream journals? And if so, might racism be a latent factor?

As to the first part of this question, I believe the unequivocal answer is Yes. I know of several "controversial" issues that could have been resolved long ago, had they not been quietly shelved and had subsequent efforts to raise them again not been systematically suppressed. Curiously, most them concern elementary statistical problems, such as factor indeterminacy, the power problem in ANOVA (see, e.g., Wahlsten, 1990), Holzinger's h^2, or the 'file drawer problem' of meta analysis (cf. Schönemann, 1991). As to the second part, I am not so sure. In some cases racism may have been a factor, but numerous other plausible explanations come to mind: Technical incompetence, vested interests, indifference, arrogance – the list goes on.

To illustrate the depth of the problem, let me return once more to the congruence coefficients which the Jensen school routinely holds up as proof of the ubiquity of *g*. The underlying myth that PC1s can validate Spearman's *g* has been around for decades.

Tests in different batteries often differ widely in content. For example, simple inspection of the Wechsler, say, shows that it consists of at least two distinct clusters of tests. It would be most surprising if the PC1 of its verbal tests were the same as the PC1 of its performance tests. In this particular instance we know that they are not the same, because we can compute their correlation, roughly .7. For different batteries given to different subpopulations, this simple check is no longer possible. In this case, we are routinely shown congruence coefficients as evidence of the ubiquity of *g*. This fallacious argument has been around for so long that it no longer seems possible to dislodge it by rational argument.

Let us therefore try a different tack and revisit Perron's theorem: We know from Perron that any positive correlation matrix whatever will always have a PC1 which explains more observed variance than any other component, and that the defining eigenvector can always be chosen positive. These positive eigenvectors serve as input into Jensen's Magnificent *g*-Machine. It seems a fair question to ask: If we just take any two pxp positive matrices whatever, what would be the distribution of the cosines between their largest eigenvectors?

This question can be approached indirectly by first asking for the distribution of cosines between a vector of 1s and the dominant eigenvectors of positive random matrices. The results of such simulations are summarized in Table 4. Two different parent distributions were used: (a) the Chi-square distribution with 1 df, and (b) uniform (0,1) random deviates. The construction of the positive, gramian random correlation matrices followed the recipe given in the target article, except that the average off-diagonal correlations were controlled exactly.

One finds that the cosines ('congruence coefficients') are somewhat lower for the severely skewed chi-square distribution, and smaller p, than for the uniform distribution and larger p. The median cosines do not vary much across average correlations. All of them are in the high nineties. For the chi-square case and $p = 20$, the median is .992, with associated angles between 7° and 9°. For the uniform case the cosines are still higher (and the angles smaller). To obtain upper bounds on the angular separations between the dominant eigenvectors of two different positive random matrices, the angles in the Table 4 would have to be doubled. That puts them well below 20°, with a cosine of .94.

Table 4
Angular separation between vectors of ones and dominant eigenvectors of random positive matrices

Order:		10x10					20x20				
Average correlation		min	Q1	Med	Q3	max	min	Q1	Med	Q3	max
A. Parent Distribution Uniform (0,1)											
.2	cos	995	997	999	999	999	998	999	999	999	999
	angle	5.9	3.4	2.9	2.6	2.6	3.2	2.4	2.4	2.4	2.4
.4	cos	997	998	999	999	999	998	999	999	999	999
	angle	4.8	3.4	2.9	2.4	2.4	3.4	2.6	2.3	2.3	2.3
.6	cos	995	998	999	999	999	999	999	999	999	999
	angle	6.0	3.4	2.8	2.3	2.3	3.0	2.4	2.3	2.3	2.1
B. Parent Distribution Chi-Square 1											
.2	cos	968	980	987	991	999	980	989	992	993	996
	angle	14.4	11.6	9.2	7.7	2.9	11.6	8.3	7.5	6.7	4.8
.4	cos	966	982	988	991	998	981	990	992	994	998
	angle	15.0	10.9	8.9	7.7	4.1	11.2	8.0	7.3	6.5	3.6
.6	cos	949	981	986	991	997	980	989	992	994	996
	angle	18.4	11.2	9.5	7.5	4.1	11.4	8.6	7.4	6.5	4.8

Legend: cos: = cosine; Q1, Q3 : = lower and upper quartile; Med: = median; min : = minimum, max : = maximum. Decimal point omitted for cosines.

Thus, as a direct consequence of Perron's theorem, and of test construction procedures which ensure that it applies to IQ data, one arrives at stunning confirmations of the myth that *g* is ubiquitous, if one adopts the logic of the Jensen school.

It seems besides the point to worry whether some people follow this logic because they are racists, or are racists because they follow this

logic. The important point is that nobody seemed to mind as these transparent fallacies percolated through our mainstream journals for decades. During all this time, I never heard of a single statistician protesting against this patent charade.

This, then, would be another example of a significant artefact.

Turkheimer: « A psychometric left would recognize that human ability, individual differences in human ability, measures of human ability, and genetic influences on human ability are all real but profoundly complex, too complex for the imposition of biogenetic or political schemata. It would assert that the most important difference between the races is racism, with its origins in the horrific institutions of slavery only a few generations ago. » (Turkheimer, 1997).

There is much in Professor Turkheimer's commentary I resonate with. As a guiding principle I could endorse almost every word in his eloquent statement, with the possible exception of 'psychometric left'. I do not believe it is helpful to invoke political categories to characterize positions on the IQ and nature/nurture issue. As Allen (1997) has stressed, many of the early advocates of eugenics considered themselves social reformers. Karl Pearson, for example, fancied himself a socialist. He expended considerable effort in studying the dirt under the fingernails of Jewish immigrant children in an effort to persuade the British government to follow the U.S. example and tighten the immigration laws (Pearson & Mohl, 1925/1926).

Conversely, many critics of Herrnstein's and Murray's (1994) *The bell curve*, while undoubtedly sincere and well-intentioned, lack the technical background needed to mount an effective challenge to the many erroneous claims supposedly backed up by data in the forbidding Appendix. To give but one example, take Herrnstein's and Murray's claim, that « ... one of the smallest R^2 in the following analysis, only .17, is for white men out of the labor force for four weeks or more in 1989 » (p. 594). None of the 50 odd critics in Jacobi and Glauberman's (1995) *The bell curve debate* noticed that this claim is quite simply false. If one ranks the R^2s in the Appendix, one finds that their median is somewhere near .07, not .17, and that 80% of the values are smaller than « one of the smallest R^2 ». Taking the root of .07 leads right back to the familiar low validities in the 20s, which erode the very foundation of Herrnstein and Murray's house of cards. The point is, once again, that strong convictions alone – on either side of the aisle and no matter how sincere – are no substitute for minimal technical skills. It is more difficult to explain why Jacobi and Glauberman's long list of critics of *The bell curve* does not include a number of statisticians who would be capable to refute Herrnstein and Murray decisively.

I am also prepared to endorse Turkheimer's conjectures about human abilities and their possible genetic basis, at least as plausible *working hypotheses*, together with other, perhaps contradictory assertions I might consider equally plausible. Where I draw the line is when it comes to passing any one of them off as scientific fact. Here, again, the devil is in the details.

Take 'intelligence'. As soon as we get down to specifics, we find experts still arguing over whether a general human ability, e.g., *g*, exists. One might think they would have gotten beyond the plausible working hypothesis stage by now. Since they did not, they should at least admit that they are unable to measure "intelligence", and to say how much of it is inherited, simply because they cannot agree on a definition of 'intelligence'.

Thus, I cannot agree that elementary logic should be negotiable in a science just for the sake of harmony. If psychologists persist in their indifference towards violations of plain common sense, then they may eventually lose the public trust they need to underwrite their research.

Turkheimer takes me to task for what he sees as an unfortunate and counterproductive stance of radicalism:

« Intelligence is a myth (Schönemann, 1997), its factor structure an illusion, heritability overestimated and probably nonexistent ... the opponents are either fools or charlatans (Hirsch, 1981) ». While my « target paper largely avoids this unfortunate tone », my Discussion « lapses into ad hominem argument [and] conspiracy theories about a peer-review system ... » (Turkheimer, 1997).

I did not exactly say « intelligence is a myth ». I said it is 'undefined'. I may have said « general intelligence is a myth ». As long as 'intelligence' is undefined, we cannot estimate its heritability. This much seems obvious. How can we find out whether an Uidu has two horns or one, if we are not told what an 'Uidu' is? On the other hand, if we want to define 'intelligence' as *g*, as Jensen (1969) once did, then we have to make sure that *g* exists, before heritability estimates can make sense. Whether it exists is a matter of looking at its definition (Spearman's, not Jensen's) and the data. If one does this, then one finds that g does not exist, any more than round triangles or unicorns exist. May we

call someone a 'charlatan » who claims that an Uidu, just like intelligence, is easier to measure than to define?

If Turkheimer doubts that h^2 contains no environmental variance (so that it cannot possibly estimate narrow heritability), why not simply retrace the steps in (Holzinger, 1929)? Similarly, if he doubts that the 'correct' estimate, HR, is frequently larger than 1, why not simply read Osborne (1980)? He gives numerous tables of such absurd heritability estimates without comment (e.g., Table VII-G, p. 68, Table IX-H, p. 98, Table XII-I, p. 136), and the reader does not have to compute a thing.

In short, I do not agree that the issue boils down to a choice between left and right, or radical and conservative. So far as I can see, it boils down to a choice between good and bad science.

Wahlsten: « Participants in the heritability enterprise will continue to thrive as long as millions of dollars of grant money continue to flow their way and they are placed in positions of power to hire and fire young scholars, not to mention accept and reject papers for influential journals » (Wahlsten, 1997b).

Professor Wahlsten is well known for his courageous stand against abuses of behavior genetics (e.g., Wahlsten, 1990, 1997a). I agree with him that the American peer review system is in dire need of reform.

One benefit of my Odyssey was a rich harvest of evidence that the American peer review system no longer seems to be able to deal with controversial issues in a rational manner if they impinge on vested interests. This includes mainstream journals both from psychology and statistics.[1]

ACKNOWLEDGEMENTS

Support of the Taiwanese Foundation for the Advancement of Outstanding Scholarship (FAOS) is gratefully acknowledged.

I would like to thank my wife, Roberta, for her unstinting loyalty in hard times, and the managing editor of CPC, Françoise Joubaud, for her encouragement and patience.

1. A timeline of attempts to refute the Spearman hypothesis artefact is available from the author upon request.

REFERENCES

Allen, G. E. (1997). The social and economic origins of genetic determinism: a case history of the American Eugenics Movement, 1900-1940, and its lessons for today. *Genetica, 99,* 77-88.

Andreski, S (1972). *Social science as sorcery.* New York: St. Martins's Press.

Chase, A. (1975). *The legacy of Malthus: The social costs of the new scientific racism.* New York: Knopf.

Chauncey, H., & Dobbin, J. E. (1963). *Testing: Its place in education today.* New York: Harper and Row.

Cox, D., & Miller, H. D. (1968). *The theory of stochastic processes.* New York: Wiley.

DeFries, J. C., Corely, R. P., Johnson, R. C., Vandenberg, S. G., & Wilson, J. R. (1982). Sex-by-generation and ethnic group-by-generation interactions in the Hawaii Family Study of Cognition. *Behavior Genetics, 12,* 223-230.

Dolan, C. V. (1997). A note on Schönemann's refutation of Spearman's Hypothesis. *Multivariate Behavioral Research, 32,* 319-325.

Estes, W. K. E. (1992). Ability testing: Postscript on ability tests, testing and public policy. *Cognitive Science, 5,* 278.

Garriga-Trillo, A. (1997). Are there other famous artefacts? *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 695-701.

Guttman, L. (1992). The irrelevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27,* 175-204.

Hay, D. (1997). Genetic research on intelligence is more than just psychometrics. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 702-708.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve.* New York: Free Press.

Hirsch, J. (1981). To "unfrock the charlatans". *Sage Race Relations Abstracts, 6,* 1-67.

Hirsch, J. (1997). The triumph of wishfull thinking over genetic irrelevance. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 709-718.

Holzinger, (1929). The relative effect of nature and nurture influences on twin differences. *Journal of Educational Psychology, 20,* 241-248.

Horn, J. (1997). On the mathematical relationship between factor or component coefficients and differences between means. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 719-726.

Humphreys,. L. G. (1997). Professor Schönemann is both right and wrong. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 727-730.

Jacobi, R., & Glauberman, N. (1995). *The bell curve debate.* New York: Random House.

Jensen, A. (1967). Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proceedings of the National Academy of Sciences, 56,* 149-156.

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1-123.

Jensen, A. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences, 8,* 193-263.

Jensen, A. (1987). Differential psychology: Towards consensus. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 353-399). New York: Falmer Press.

Kadlec, H. (1997). The correlation between the mean difference vector and first principal component holds even with discrete variables containing no *g*. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 731-737.

Kamin, L. (1974). *The science and politics of IQ.* Potomac, MD: Erlbaum.

Lewontin, R. C. (1970). Race and intelligence. *Bulletin of the Atomic Scientists, 26,* 2-8.

Loehlin, J. C. (1992). On Schönemann on Guttman on Jensen, via Lewontin. *Multivariate Behavioral Research, 27,* 261-263.

Maraun, M. (1997). Exactly what is an artefact? *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 738-747.

Millsap, R. E. (1997). The investigation of Spearman's Hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 748-755.

Nagoshi, C. T. (1997). *g*-loadings and the nature and salience of intelligence in the Hawaii Family Study of Cognition. *Cahiers de psychologie Cognitive Current Psychology of Cognition, 16,* 756-759.

Osborne, R. T. (1980). *Twins: Black and White.* Athens Georgia: Foundation for Human Understanding.

Owen, D. (1985). *None of the above. Behind the myth of scholastic aptitude.* Boston, MA: Houghton Mifflin.

Pearson, K., & Mohl, M. (1925/1926). The problem of alien immigration into Great Britain, illustrated by an examination of Russian and Polish Jewish children. *Annals of Eugenics, 1,* 5-127.

Schönemann, P. H. (1981). Factorial definitions of intelligence: Dubious legacy of dogma in data analysis. In I. Borg (Ed.), *Multidimensional data representations: When and why* (pp. 325-374). Ann Arbor, MI: Mathesis.

Schönemann, P. H. (1985). On artificial intelligence. *Behavioral and Brain Sciences, 8,* 241-242.

Schönemann, P. H. (1991). In praise of randomness. *Behavioral and Brain Sciences, 14,* 162-163.

Schönemann, P. H. (1992). Extension of Guttman's result from g to PC1. *Multivariate Behavioral Research, 27,* 219-224.

Schönemann, P. H. (1993). A note on Holzinger's heritability coefficient h^2. *Chinese Journal of Psychology, 35,* 59-65.

Shockley, W. (1987). Jensen's data on Spearman's hypothesis: No artifact. *Behavioral and Brain Sciences, 10,* 512.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201-293.
Steiger, J. H. (1997). Alternate models and the evaluation of social issues. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 760-766.
Sternberg, R. J., & Grigorenko, E. L. (1997). Infamous artefacts in the study of intelligence: Why is there so much support for so many hypotheses. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 767-776.
Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology, 8,* 271-281.
Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.
Turkheimer, E. (1997). The search for a psychometric left. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 777-781.
Wahlsten, D. (1990). Insensitivity of the analysis of variance to heredity-environment interaction. Target Article. *Behavioral and Brain Sciences, 13,* 109-120.
Wahlsten, D. (1997a). Leilani Muir versus philospher king: Eugenics on trial in Alberta. *Genetica, 99,* 185-198.
Wahlsten, D. (1997b). Studied ignorance weakens human behavior genetics. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 16,* 782-785.
Wechsler, D. (1939). *The measurement and appraisal of adult intelligence.* Baltimore, MD: Williams and Wilkins.
Wilson, E. B. (1928). Review of The Abilities of Man, their Nature and Measurement by C. Spearman. *Science, 67,* 244-248.

CPC

Cahiers de Psychologie Cognitive
Current Psychology of Cognition

VOLUME 16 NUMBER 6 DECEMBER 1997

Target article

Commentaries

Author's response

Continuing commentary/Rationality in reasoning

Author's response