Multivariate Behavioral Research, 27 (2), 175-204

The Irrelevance of Factor Analysis for the Study of Group Differences

(A Continuing Commentary on A.R. Jensen's target article, *The nature of the black-white difference on various psychometric tests: Spearman's hypothesis.*)

Louis Guttman
The Hebrew University of Jerusalem and
The Israel Institute of Applied Social Research

Jensen's target article (Jensen, 1985) discusses three hypotheses. The first is the basic one of the late Charles Spearman (1932) on the factor structure of intelligence test scores (Spearman's *g* hypothesis). The second is one mentioned only in passing by Spearman, but pursued vigorously here by Jensen, on a relation of *g* to racial differences on intelligence test scores. The third is Jensen's own, on the biological *nature* of *g*. The main thrust of the target article is to try to establish the last two hypotheses. The purpose of the present commentary is to show how Jensen has failed in his efforts, the failure being more fundamental than brought out in the 29 peer commentaries published alongside the target article.

For the convenience of the reader, each of the three hypotheses will be restated in full when its turn comes for in-depth discussion below. The second and third hypotheses depend directly on the first, namely, that there exists a single common factor for intelligence test scores. Hence the last two hypotheses can hardly be discussed coherently, in the form presented, if the first is incorrect. Accordingly, the present commentary will open with an analysis in some depth of Jensen's (1985)(and other factor analysts') exposition of the first hypothesis — an exposition that we shall find to be an inaccurate and misleading account of Spearman's works. An extended critique seems to be in order here, since almost all the peer commentaries accompanying the target article appear to accept Jensen's faulty formulations concerning *g* at face value, and some even give them a hearty amen.

The second hypothesis concerns the relationship between group differences in test scores and the factor structure of those scores. It is surprising to find that neither Jensen's target article (1985), nor any of the 29 peer commentaries on it, contains any mention of the algebraic analyses of a closely related problem published decades ago by two of the giants of factor analysis: Godfrey Thomson (1939) and L.L. Thurstone (1947). Had Jensen realized the *algebraic*

nature of the problem, he might have been dissuaded from his *empirical* attempts to *substantiate* the hypothesis, and from writing the target article at all. We shall see how the second hypothesis is actually a *missing theorem* for the algebra of factor analysis. Proof of the theorem is provided towards the end of this commentary, in the section entitled *The Missing Theorem on Group Differences.*

Jensen (1985) and all the 29 commentators are similarly at fault with respect to the third hypothesis, this time by ignoring Spearman (1932) himself. Not one of them cites Spearman's own writings on the *nature* of g — despite the title of the target article; and so none of them has pointed out that Spearman specifically rejected Jensen's hypothesis of reaction-time in this regard. The discussion to follow may reinforce and modernize Spearman's cogent anti-Jensen arguments.

The present commentary, then, is to show how Jensen's (1985) conceptualization and treatment of each of the three hypotheses require basic revision, over and beyond what the 29 commentaries may have indicated. But before going into details, let me put in a word in defense of Spearman (1932), to whom the target article does great wrong.

Disservice to Spearman: Distortion of Basics

It turns out that the very title of the target article is in error. Contrary to what is implied by that title, *neither* of the first two hypotheses of Spearman (1932) has anything to do with the *nature* of black-white differences. It is Jensen (1985) himself who proposes and puts great effort here into trying to substantiate the third hypotheses, about the biological nature of g; and he neglects to point out that Spearman had rejected this at the outset. We shall analyze Jensen's valiant attempt to read physiology (via reaction-time) into the picture of the *nature* problem, and shall show how it fits into the general theory of mental tests without any regard to Spearman's factor analytic hypotheses or to biology.

The main body of the target article is based on — and uncritically expounds — many of the conceptual and computational distortions of factor analysis that have grown up over the years, especially with respect to the hypothesis of Spearman's g, (1932) its (falsifiable) generalizations, and its correlates. To unscramble this jumble, it will be necessary to go back a bit into history and to first principles. In a way, Jensen (1985) may have done a useful service in bringing together such a catalogue of malpractices and peccadillos of factor analysis, so that it can be dissected in the context of a single scientific problem (black-white differences) for which the first hypothesis was coopted. The

deficiencies to be detailed in the following are quite general for the practice of factor analysis, and are not at all limited to the present substantive problem.

Unintentionally, Jensen (1985) has done a particularly useful service in raising the question of group differences in the form of the second hypothesis involving factor analysis. Despite his protestations (p. 248, first column), this question shows that there is a chapter missing in the algebraic literature of factor analysis. We shall fill in the essentials, from which it will become apparent that seeking biology via factor analysis may be just tilting at a windmill. Spearman himself was wisely skeptical of such a biological quest (Spearman, 1932, p. 384; this edition is hardly changed from the first edition of 1927), even though he was unaware that the second hypothesis was essentially but an algebraic consequence of the first.

Being an algebraic proposition, the second hypothesis is not to be tested by tortuous empirical manipulations such as those employed by Jensen (1985). Indeed, Jensen himself points out that there are two algebraic features that need to be specified in advance if the hypothesis is to be well-defined. But without even attempting an algebraic analysis of the problem, Jensen simply makes two guesses, one for each of these algebraic features. Our *missing theorem*, to follow, proves that one of Jensen's guesses is wrong, while the other happens to be right. But the need to discuss the *right* feature at all raises doubts, from the point of view of empirical science, as to the very foundations of factor analysis. To lay bare what is involved, we shall in effect have to give a short course on the basics. This may be particularly helpful to the reader not very conversant with technicalities of factor analysis. Some of the basics may be news even to veteran factor analysts.

Mistreatment of Spearman's g

Jensen (1985) is fully aware of the fact that Spearman's (1932) famous first hypothesis — of a unidimensional g factor — is false, yet he tries to read the non-existent g into a multiple factor framework. He accepts the multidimensionality, so it is not clear why he does not frame the second hypothesis accordingly — without the ghost of g — and study the response surface of the race differences over the multidimensional space. We shall show here inconsistencies generated by the attempts of Jensen and others to save g — including the inconsistencies of the multiple common factor proposals — and shall suggest a proper way of studying group differences over the universe of tests.

Spearman (1932) himself was among the first to disprove his own hypothesis of a single common factor for intelligence test scores. Therefore,

he and others developed varieties of multiple common factor theories (whose correctness is irrelevant, for the moment, to the present discussion — their mere publication implies rejecting the *g* hypothesis). One of the curious chapters in the history of science is how the terminology of a false hypothesis — Spearman's *g* — has been so attractive that it has been carried over into other contexts, giving the impression that the concept was still viable but only being studied under a different dress.

An example of a basic conceptual distortion in Jensen's article is in his assertion: "The *g* factor is Spearman's label for the single largest independent source of individual differences that is common to all mental tests ..." (1985, p. 194). Were this assertion true, then all the efforts by Spearman (1932) and other researchers to establish the *existence* of *g* have been unnecessary; *g* would exist *by definition* as the first principle component of an infinite matrix. Spearman would certainly have objected to Jensen's attributing to him such a non-falsifiable hypothesis. According to Spearman, "the very definition of *g*" is the factor common to mental tests whose correlation matrix satisfied his "tetrad equation" (p. 161). For the reader's convenience, the tetrad condition is stated below, in the form of Equation 5, along with a sketch of its proof. It is the failure of mental tests to satisfy Spearman's equation in practice that destroyed his hypothesis.

It is hard to reconcile Jensen's (1985) assertion with Spearman's definition (1932), since satisfying Spearman's tetrad equation has nothing to do with a "largest independent source of individual differences". Common factor theory per se — unidimensional or multidimensional — is concerned directly only with reproducing observed covariances; paradoxically, this can be perfectly achieved (as when Spearman's equation is satisfied) no matter how little the variances (individual differences) are *accounted for* thereby. We shall shortly restate Spearman's algebra — and also the algebra for $m > 1$ common factors — in a way that the reader can see for himself the irrelevance of size-of-variance considerations to the covariance analysis.

It is even more difficult to reconcile Jensen's (1985) assertion (not to speak of Spearman's, 1932, definition) with Jensen's later statement that *g* can be "extracted as a first principle factor or as a hierarchical second-order factor" (p. 196). Why this liberty as how to "extract"? Perhaps because one is dealing only with a ghost of a departed hypothesis. Jensen later apparently recants and says that "the largest common factor ... may often be interpreted as a general factor, or *g*" (p. 198). Why only "may"? Why only "often"? When may it and when may it not? Is there a testable hypothesis here after all? If so, what is the hypothesis and what is the test? Further conflicting and amorphous statements of these kinds abound in the article, in contrast to Spearman's original clear formulation of his falsifiable (and unfortunately false) *g* hypothesis.

One does not have to know anything about factor analysis to sense that here is a rather cavalier treatment of definitions. Jensen's (1985) main excuse given for his confusing confounding of concepts is *practical*: Everything correlates positively with everything. But such loose treatment of concepts and data is hardly the stuff of which science is made. Let us take a closer look at Jensen's excuse.

The Phenomenon of Positive Covariance

Jensen (1985) properly points out that one of the best established phenomena in all of science is that, for usual populations, scores on any two mental tests will correlate (covary) positively with each other. It was this phenomenon of positive covariance that stimulated Spearman (1932) into thinking about *g*. Jensen could have gone on to cite something essential that Spearman lacked in this regard, namely, a formal definition of what belongs in an intelligence test. Such a definition was first published in 1973, and has made possible restatement of the positiveness of covariance as the "First Law of Intelligence". The 1973 definition of the universe of content of intelligence tests is:

> Intelligence Test Definition: An item belongs to the universe of intelligence tests items if and only if its domain asks about an objective rule, and its range is ordered from "very right" to "very wrong" with respect to that rule. (Guttman in Gratch, 1973, p. 37; cf. also Levy, 1985, p. 60).

Two basic facets for classifying the domain, that yield empirical lawfulness, were published earlier (Guttman, 1964), and have since been extended (cf. Guttman, 1980; Levy, 1985; Schlesinger & Guttman, 1969). These provide examples of how defining a universe of content makes it possible to state hypotheses – that prove to be viable and cumulative – concerning empirical consequences. The first, most universal hypothesis, for mental tests does not require subclassification of the domain, but implicitly focusses only on the range:

> The Positive Monotonicity (First) Law of Intelligence: If any two items are selected from the universe of intelligence items, and if the population observed is not selected artificially, then the population regressions between those two items will be monotone and with positive or zero sign. (Guttman in Gratch, 1973, p. 37; cf. also Levy, 1985, p. 62).

A crucial feature for the 1973 Intelligence Test Definition (Gratch, 1973) is the *common range* of being objectively right or wrong. A regression curve is obtained by plotting the range of one variable against the range of the other. Thus, the definitional commonality of range provides part of the rationale for the positiveness of regression slope. Positiveness of slope accounts for the well-replicated phenomenon of positive covariance.

Note that the First Law specifies only monotonicity, and not strict linearity of slope. The numerator of monotonicity coefficient μ_2 is actually the same covariance as for the numerator of Pearson's r (cf. Guttman, 1986; Raveh, 1978); both coefficients have the same sign, namely that of the covariance. The LAW goes more deeply into the problem of shape of slope — and hence of covariance — than done previously, and does not depend on factor analysis or any other kind of algebraic analysis.

Had Jensen (1985) paid proper attention to the 1973 Intelligence Test Definition (Gratch, 1973), he might have avoided the pitfall of the reaction-time hypothesis he introduces at the end of the target article, a pitfall analyzed below. Similarly, had he paid proper attention to the First Law (Gratch, 1973), he might have avoided the three *traps* and other misinterpretations analyzed to follow, and saved writing the bulk of his article that is devoted to factor analysis.

Spearman's g Hypothesis and Its Failure

Spearman (1932) did not give a sharp definition of intelligence tests, but was impressed by the positive covariance phenomenon. Instead of focussing on the *common range*, he proposed his g common factor hypothesis as an algebraic rationale for the phenomenon. His emphasis was more on algebra than on content. A succinct way of stating Spearman's hypothesis is in terms of partial correlation (Spearman, 1932, Appendix p. iii), as follows. (Linearity of regressions is assumed throughout here — a typical assumption of factor analysis.)

Spearman's g Hypothesis

There exists a variable, to be denoted by g, on which every individual in an ordinary population can be scored, and which satisfies the following two conditions for that population:

1. *Sign Condition* — If x is any empirical test item on which the individuals have been scored, then the correlation between x and g will be positive (or zero):

$$r_{xg} \geq 0. \tag{1}$$

2. Common factor *Condition* — If x and y are any two empirical test items on which the individuals have been scored, then the correlation between the x and y scores will vanish when the g scores are held constant:

(2) $$r_{xy.g} = 0 \ (x \neq y).$$

Should part 2 of the hypothesis be true, it would follow that

(3) $$r_{xy} = r_{xg} r_{yg} \ (x \neq y).$$

In consequence, were also part 1 true, then the positiveness of covariance must hold:

(4) $$r_{xy} \geq 0.$$

Thus, Spearman's *g* hypothesis (1932), were it true, would suffice to account for the positiveness phenomenon.

Spearman (1932) was quick to realize that, unfortunately, his hypothesis accounted for too much: it says something very restrictive about the *sizes* of the r_{xy}, and not merely about their sign.

In the right of Equation 3 are correlations with *g*, which will need to be calculated if *g* exists. However, Equation 3 in turn leads to a further consequence that does not involve *g* directly, namely Spearman's (1932) famous "tetrad condition". Let *x*, *y*, *u*, and *v* be any four distinct tests in the battery. If Equation 3 is true, then also the following must be true (Spearman's Tetrad Condition):

(5) $$r_{xy} r_{uv} = r_{xv} r_{uy} \ (x, y, u, v \text{ all distinct}).$$

How well equality in Equation 5 is satisfied by the empirical data can be checked directly from the observed correlations, without having to estimate factor loadings.

Equation 5 as is provides only a necessary condition for a *g* to exist for the battery. If the proviso is added that no calculated *loading* exceeds 1, then the condition becomes also sufficient. For convenience in the present discussion, we shall specify that the term "tetrad condition" includes this proviso, so as to save separate discussion of sufficiency.

Modern treatment of the algebra of factor analysis is largely in terms of matrix algebra. In matrix language, Equation 5 is the condition for the correlation matrix — to be of *rank 1*. Spearman's (1932) "tetrad" terminology has gone out of fashion. However, for the present discussion, it will be edifying to return to Spearman's language. Jensen (1985) claims that his target article is based on Spearman's work because of "Spearman's excellent track record in psychometrics" (p. 194). Using Spearman's own words can help us check whether or not Jensen has stayed on the right track.

The tetrad condition is equivalent to saying that any two rows (columns) of the observed correlation matrix must be proportional to each other (excluding the main diagonal elements). Examination of many matrices, by Spearman (1932) himself and by others, showed that only few — unless tampered with by throwing out tests — had some fair approximation to the proportionality property. Any reader of these lines can himself easily disprove *g* by looking at almost any mental test correlation matrix at his disposal and checking for the proportionality.

A rapid way to do the checking is merely to scan the matrix according to the following.

Monotonicity Check for g

Let *x*, *y*, and *z* be any three distinct variables of the matrix. If *g* exists for the matrix in the sense of conditions 1 and 2 listed previously with Spearman's *g* Hypothesis, and if $r_{xz} \geq r_{yz}$ for one *z*, then this inequality is true for all *z*.

Thus, the *g* hypothesis has the robust property that it can be checked without any arithmetical calculations at all. Just look at the observed matrix two rows at a time and see if the inequality check, just described, holds consistently.

The Monotonicity Check can be facilitated by rearranging the rows and columns of the correlation matrix so that the largest correlations are in the upper left-hand corner, and the smallest in the lower right. Spearman (1932) knew that this should yield a clear gradient if his hypothesis were true. He apparently was also aware that this rearranging would make obvious any systematic departures from the gradients, so he preferred algebraic checks that would be correct in principle (like his tetrad equation), but less revealing in practice. Thus, in his *The Abilities of Man* (Spearman, 1932), he largely refrained from rearranging the rows and columns of the few small and selected matrices he claimed supported *g*; this would have weakened the impression given by his arithmetical checks.

Spearman (1932), 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. He tried to save his hypothesis by *explaining away* the recalcitrant cases. One interesting variety of explanation that he suggested was that of "overlap" (Spearman, 1932, pp. 150-153). In effect, he pointed out facet designs for batteries of mental tests that would lead to hypotheses different from *g*. In doing this, he was anticipating facet theory (cf. Canter, 1985). Unfortunately, he did not go on to try to systematize hypotheses based on domain facets. His interest was largely to eliminate such facets in order to save his *g*. Thus, he

acknowledged that there is no single common factor for the universe of intelligence tests.

Spearman did not really resuscitate g, despite the last chapters devoted to this effort in his *The Abilities of Man* (Spearman, 1932). He gave no necessary nor sufficient conditions — neither algebraically nor contentwise — for doing so. Instead, Spearman was guilty of setting the example — followed by so many later factor analysts — of not being able to give up the terminology of the non-existent g, and of sacrificing falsifiable algebraic formulations and clarity of conceptualization in trying to save g. Like most other ventures into multiple common factor formulations — that continue to this day — Spearman forgot to check whether or not his extended formulations account for the phenomenon of positive covariance. It is remarkable that, with possibly but one exception — to be described next — no extant multiple common factor formulation attempts to account for, or even relates to, the positiveness phenomenon which originally motivated factor analysis.

Thurstone's Positive Manifold Hypothesis and Its Failure

The failure of the g hypothesis led to a proliferation of multiple common factor hypotheses. The one staying closest in spirit to g is perhaps Thurstone's (1935) hypothesis of the *positive manifold.* Here again is a peculiar chapter in science contributed by devotees of factor analysis. The term *positive manifold* has been mangled almost as much as has been g. Jensen joins many other writers in misusing Thurstone's term, as when he writes, "... the positive manifold phenomenon; that is, the existence of positive correlations between all tests ..." (Jensen, 1985, p. 195). Now, the positive manifold is a technical term in geometry. Thurstone introduced it into factor analysis as an *hypothesis* to account for the positiveness phenomenon — not as an unnecessary new name for the phenomenon itself. Indeed, Thurstone was careful to admonish that, "even if all of the original intercorrelations are positive or zero, it does not follow that the trait configuration can be inscribed in a positive orthogonal manifold" (Thurstone, 1935, p. 202).

Were the positive manifold hypothesis — or its special case of g — correct, this would account algebraically for the positiveness of covariance. The converse is not true: the positiveness phenomenon — or the First Law of Intelligence — does not automatically imply a positive manifold, and certainly not g.

Thurstone himself disproved his own hypothesis empirically for the case of intelligence tests. It may be useful to restate Thurstone's hypothesis here in a manner which will show how it is a most immediate generalization of Spearman's g hypothesis.

Thurstone's Positive (Orthogonal) Manifold Hypothesis

There exist m variables — to be denoted by $c_1, c_2, ..., c_m$ — on which every individual in an ordinary population can be scored, and which satisfy the following three conditions.

1. *Sign Condition* — If x is any empirical test item on which the individuals have been scored, then the correlation between x and each of the m variables will be positive (or zero):

$$r_{xc_j} \geq 0 \; (j = 1, 2, ..., m). \tag{5}$$

2. Common factor *Condition* — If x and y are any two empirical test items on which the individuals have been scored, then the correlation between the x and y scores will vanish when the m variables are held constant:

$$r_{xy.c_1c_2...c_m} = 0 \; (x \neq y). \tag{6}$$

3. *Orthogonality Condition* — The correlations among the m variables themselves are zero:

$$r_{c_jc_k} = 0 \; (j \neq k; \; j, k = 1, 2, ..., m). \tag{7}$$

The subscript m has been chosen for the conditions here, and the subscript 1 for Spearman's conditions (1932), in order to bring out how Spearman's is the special case where $m = 1$. (Condition 3 is not relevant to Spearman's case.)

Should conditions 2 and 3 of Thurstone's (1935) hypothesis be true, it would follow that

$$r_{xy} = r_{xc_1}r_{yc_1} + r_{xc_2}r_{yc_2} + ... + r_{xcm}r_{ycm} \; (x = y). \tag{8}$$

In consequence, were Thurstone's (1935) condition 1 also true, then the positiveness of covariance must hold, since r_{xy} would then be the sum of m positive (or zero) terms as in the right of Equation 8.

In his classic textbook, *The Vectors of Mind*, Thurstone devotes an entire chapter — entitled "The Positive Manifold" — to this hypothesis (Thurstone, 1935, Chapter VIII). The chapter explores many algebraic aspects of the hypothesis (including an interesting genetic excursion). How the practitioners of factor analysis have come to twist Thurstone's terms, despite this chapter, is no compliment to the rigor of their practice. Thurstone himself, of course, became disenchanted with the hypothesis; in the second — greatly enlarged —

edition of his textbook, "The Positive Manifold" no longer commands a chapter, but is relegated to two or three pages (Thurstone, 1947, pp. 341-343).

Thurstone's "Simple Structure" Hypothesis and Its Failure

In casting about for an alternative falsifiable hypothesis of multiple common factors, Thurstone (1935) latched onto the concept he called "simple structure". It is especially relevant to the present discussion to note that this new proposal had nothing to do with the positiveness of covariance of mental tests, or with mental tests at all. This is tacitly acknowledged by Thurstone by his omitting "mind" from the title of his second edition. Thurstone's motivation now was more technical than psychological. He posited that "One of the important restrictions that must be satisfied by any acceptable solution to the factor problem is that the factorial description of a trait or test must be invariant when it is moved from one battery to another ... This is the reason why I have discarded one of my earlier solutions, namely the principal axes of the configuration ..." (Thurstone, 1935, p. viii). Contrary to Jensen and others, Thurstone at the outset rejected the "proportion of variance" approach of principal axes or components as being unscientific for determining common factors — whether or not this approach is diverted into an attempt to revive *g*.

The simple structure hypothesis, too, has its problems. We shall not go into technical details here, but state two major failings. The first is the frailty of the computing techniques used in practice: *no computer program rejects the hypothesis in practice*. This never-fail property reflects the fragility of the hypothesis itself: the hypothesis calls for a sharp difference between zero and non-zero "loadings", which is almost mission impossible since "loadings" can vary continuously around zero. The factor analyst of the Thurstone school typically reports that he used such-and-such a computer program (often the Varimax) and "rotated to simple structure." For him, the concept is not a falsifiable hypothesis, but is something always there to be "rotated to." Most factor analysts would be hard pressed if asked to state a clear criterion for rejecting the simple structure hypothesis; and they have no computer program to rely on for this assessment. In practice, the rotation usually yields only gradients in sizes of factor loadings, and not the sharp jump between zero and non-zero loadings desired *for each and every* common factor.

In the next paragraphs we shall see that there are basic psychological reasons why the hypothesis should be false. But theoretical and empirical considerations do not deter the devotees: no matter what the (Varimax) rotation shows, the factor loadings are always "interpreted." Like Spearman, Thurstone departed from clear algebra to vague calculations which make it

difficult to reject a wrong hypothesis. Non-existent *simple structure* has the same ghostly persistence as non-existent *g*.

A second, more fundamental, frailty of Thurstone's (1935) hypothesis is its view of how mental tests are constructed. Like Spearman (1932), Thurstone did not offer a definition of what belongs in intelligence tests. Spearman did suggest many facets for classifying mental tests (but – as already noted previously – did not go on to a real facet theory coordinated with data analysis). Guilford (1956, 1967) later used Spearman's facets for classifying proposed common factors; Anastasi has commented on the distinction between using facets for classifying tests and for classifying common factors (Anastasi, 1983). Thurstone blurred the classification problem by attempting to "name" his common factors by how they relate to tests with various contents, with only intuitive and unsystematic a posteriori categories for characterizing the tests. The *simple structure* hypothesis puts the cart before the horse, and asserts that it is *impossible* to construct certain kinds of tests: if there are *m* common factors enjoying a simple structure, then it is *impossible* to construct a further test which will have a non-zero (positive or negative) loading on all *m* factors.

Thurstone (1935) originally gave no psychological rationale for such an *impossibility* hypothesis. Simple structure was proposed largely to meet the criterion of invariance across test batteries (Thurstone, 1935, p. vii). A second consideration was also technical: parsimony in factor loadings (Thurstone, 1935, p. 150). Only in his second edition does Thurstone attempt to bring in psychology, but then only in a general fashion which de facto contradicts his hypothesis of a small number of common factors (Thurstone, 1947, p. 58). He apparently did not realize that his brief discussion there was not relevant to the problems of batteries of tests with which the book was concerned, including the problem of invariance across batteries.

Jensen's (1985) laissez-faire attitude towards basic concepts and their technical terms manifests itself even with respect to the *simple structure* hypothesis. He speaks of "primary abilities independent of *g*" (p. 195). Thurstone's (1935) "primary abilities" in principle *contradict g* by the "impossibility" feature of the *simple structure* hypothesis, and hence cannot exist simultaneously with *g* – independently or not.

There are actual psychological hypotheses – which have been verified – which also meet the invariance criterion, but contradict Thurstone's (1935) "impossibility." These are *regional* hypotheses, based on a priori facet design of the content *domain* of test batteries. If there is a common factor space at all for mental tests, there is no reason why tests cannot be constructed to involve all the common factors effectively. There is no real psychological basis for the *simple structure* hypothesis. More rigorous treatment of the data, by regional analysis, shows the hypothesis to be false for the universe of intelligence tests.

Regionality turns out to be invariant across test batteries, not "factors." (Regionality also meets Thurstone's desire for parsimony, and involves even less parameters than do most factor analytic models -- including Thurstone's).

Murky Algebra

Several of the commentaries on Jensen's (1985) article — especially that of Lyle Jones (1985, p. 233) — point out some of the confusion we have just discussed. Others only add to the confusion by still trying to save *g*. For example, Cattell's commentary (1985, p. 227) asserts the "Spearman's *g* actually factors into *two* main factors, 'fluid' intelligence and 'crystallized' intelligence." It is a tribute to the vagueness of factor-analytic calculations that a factor that is supposed to satisfy Spearman's (1932) condition 2 can be divided into two factors which do not satisfy the condition, neither separately nor in some joint sense.

Even more incredible is Jensen's (1985) pronouncement that "it has proved possible to devise tests that measure *g* and little or nothing else" (p. 195, second column). If this pronouncement is true, each of such tests can be checked directly to show how it satisfies Spearman's conditions 1 and 2; and all controversy about factor analysis will disappear. Jensen makes this pronouncement (among many others on the same level) "as background for the present study" of Spearman's (1932) second hypothesis (p. 195, first column). But going on to the "Evidence for the Spearman [second] hypothesis" (beginning p. 201), one can find no mention of such pure tests. Instead, data are presented from eleven researches which use batteries like the WISC, and which are "factor analyzed" multidimensionally in quest of *g* with all the problematics outlined previously (and more, for which we shall not take the space here to discuss). It so happens that the WISC battery has been shown to have a cross-culturally invariant cylindrical regional structure (cf. Guttman, 1980; Levy, 1985). It is a pity that Jensen did not refer to this lawfulness, because it enables a straightforward comparison of racial or other group differences in means, without the distraction of the ghost of *g*.

If Spearman's (1932) tetrad condition for *g* is to be disregarded, then it would seem to be in order to state alternative falsifiable conditions — as Thurstone (1935) did. It would also be in order to drop the misleading name *g*. Furthermore, to have a psychological rationale for invariance across batteries, it is essential that the facet design of the content domain of the batteries be made explicit. Neither Cattell, nor other proposers of multiple common factor hypotheses do these things, making their hypotheses virtually non-rejectable: the hypotheses can always be read into the data. Newer computing techniques, like the "hierarchical factor analysis" that Jensen

(1985) has adopted in the target article, don't even pretend to have a substantive and falsifiable hypothesis – they always give "results," no matter what. And they all fail to account for – or even try to address – the positive covariance phenomenon that gave rise to factor analysis in the first place. Thus, there is a contradiction between Jensen's using the Schmid-Leiman (1957) supposed factor-analytic calculations for g – which do *not* account for the positiveness phenomena – and his lengthy introduction on how g supposedly underlies the positiveness phenomenon.

Three Algebraic Traps

No mental test, to our knowledge, has ever been shown to satisfy Spearman's conditions 1 and 2 directly. Had one been constructed, it would be widely in use, either by itself or as a member of a test battery. The amorphous *g*s Jensen (1985) and others talk about usually satisfy condition 1 – it is hard not to because of the algebraic theorem that might be called.

Trap 1

Given the positive covariance phenomenon among the observed tests, *any* positively weighted linear function of the observed test scores must correlate positively with each of those scores.

The mean battery score, the first principal component, and other ghosts of g are each positively weighted functions, and hence – as a mathematical consequence – must satisfy condition 1. But these ghosts hardly satisfy the more difficult condition 2. Just satisfying condition 1 is poor evidence for g.

Trap 2

Similarly, finding that three or more different "methods currently in use for factoring a correlation matrix ... yield such similar results" in estimating a "general factor" (Jensen, 1985, p. 198) is a poor excuse for thinking they are thereby corroborating the existence of g. Given the positiveness of observed covariances, any positively weighted linear functions of the test scores must correlate positively with each other. How different weighting systems for the same variables must yield similar results was examined algebraically over 50 years ago by Wilkes (1938), without any reference to factor analysis. With the weights used by "methods currently in use", the correlations are algebraically unavoidably high, despite the fact that the same data disprove g by failing the satisfy the *monotonicity check*. It is not g that is being reflected by the high

correlations, but the First Law of Intelligence — which doesn't require *g*.

For the more technically inclined reader, a further algebraic trap should be pointed out.

Trap 3

Even should Spearman's (1932) conditions 1 and 2 be satisfied, these would only pin down the common factor *loadings* of the tests (correlations of the *x* with *g*). The *scores* of the individuals on *g* would generally remain indeterminate despite the fixed factor loadings. Widely different *g* scores can be calculated to satisfy exactly the same "loadings" (Guttman, 1955; Steiger & Schönemann, 1978). If *g* exists, then it will be uniquely determinable from the test scores only if its multiple correlation on the observed tests equals one. Generally, an infinite number of tests with the same *g* would be required to achieve such perfection.

The converse is of course not generally true. Any linear function of the tests has its multiple correlation on the tests automatically perfect, and is generally *not g*. To the contrary, if the battery of tests is not extremely large, *no* linear function of it can be very close to a determinate *g* (given the usual sizes of factor loadings). Similar indeterminacy considerations hold for multiple common factors, no matter what "rotation" is used for the factor loadings.

A Mapping Sentence and Cylindrical Structure for the WISC

Of the eleven studies cited by Jensen (1985) as supporting the second hypothesis, four employed the WISC. As remarked above, regional analysis of this battery has shown a remarkably simple picture. Looking at the factor analyses, reported by Jensen in the light of this picture may be very instructive.

A regional analysis requires an a priori facet classification of the content domain of the tests. Such a classification for the WISC has been made in the form of the following mapping sentence:

"The performance of testee (x) through $\left(\begin{array}{l}\text{FACET A}\\ \text{oral}\\ \text{manual manipulations}\\ \text{paper and pencil}\end{array}\right)$

expression on an item presented orally by the tester with aid of

$$\text{FACET B}\left(\begin{array}{l}\text{verbal}\\ \text{numerical}\\ \text{geometrical}\end{array}\right)\text{language, and requiring}\left(\begin{array}{l}\text{FACET C}\\ \text{inference}\\ \text{application}\\ \text{learning}\end{array}\right)\text{of an}$$

$$\text{objective rule} \rightarrow \left(\begin{array}{c}\text{RANGE}\\ \text{very right}\\ \text{to}\\ \text{very wrong}\end{array}\right)\text{performance according to that}$$

rule." (Levy, 1985, p. 76).

Regionality corresponding to the domain facets (A, B, and C in the present case) is sought in the smallest dimensional space that enables reproducing the relative sizes of the observed correlations among the tests in the battery. To generate the space, each test is represented as a point, and the distance between two points is smaller as the correlation between the two tests is larger. Thus if d_{xy} is the distance between tests x and y and r_{xy} is the observed correlation, and if u and v are another pair of tests, then the test points are to be plotted so that

$$d_{xy} < d_{uv} \text{ whenever } r_{xy} > r_{uv}. \tag{9}$$

Using the SSA-I computer program (Guttman, 1968; Lingoes, 1973), a three-dimensional space was found to give good fit to the condition Equation 9 for every age group matrix in the WISC manuals, USA and Israel (Hebrew and Arabic). (Actually, a four-dimensional space was used, within which three-dimensional regionality became apparent, as described next.) More important than the merely technical fit to Equation 9 was the emergence of regionality in the SSA (Smallest Space Analysis) space corresponding to the domain content facets. The placement of points of the 12 subtests of the WISC is sketched in the cylinder of Figure 1, which also shows the three intersecting partitions of the space corresponding to Facets A, B, and C respectively.

SSA is the acronym of "Smallest Space Analysis"; it could better be called "Similarity Structure Analysis". It is often less appropriately called "multidimensional scaling", which distracts from thinking regionwise. This portrayal of data has the pleasing feature that any reader can check it without any calculations, not unlike the Monotonicity Check previously for Spearman's (1932) g hypothesis. All the reader has to do is to compare the distances between points in Figure 1 with the relative sizes of the correlations in any of

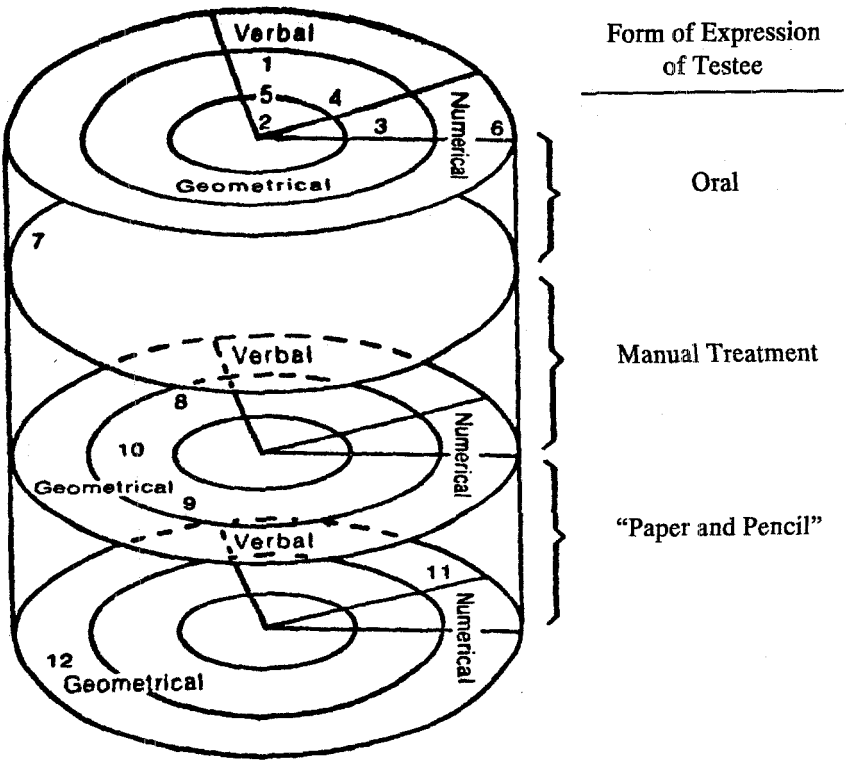

Figure 1
Schematic Representation of the Cylindrical Structure of the Wechsler Intelligence Tests for Children. (Replicated for each age group in the U.S. and in Israel -- Hebrew and Arabic.)

the available WISC correlation matrices, according to monotonicity Equation 9. Pointwise fit is imperfect in each case, but the discrepancies are bounded regionwise as in Figure 1. That is why the single Figure can be used to represent matrices that differ in local details.

Facet B and C are old friends, reappearing for almost all the dozens of article and pencil batteries studied since their regionality prowess was first published (Guttman, 1964). These invariably provide a *radex* partitioning of a two-dimensional space (or two-dimensional projection as in the present case). Some recent examples of radex replication are in Adler and Guttman (1982), Peled (1984), Tziner and Riemer (1984), and Koop (1985). Facet A takes the domain away from being only article and pencil, and turns out to partition a third dimension orthogonal to the radex plane. In reviewing the radex for intelligence tests, Sternberg and Powell (1982) conclude:"We view Guttman's type of theory as a kind of culmination of correlationally based theorizing about the nature of intelligence" (1982, p. 989).

Of special interest for exorcising the ghost of *g* is Facet C. This is a modulating facet, with rule-inference corresponding to the inner circle. This means that, for any battery of paper and pencil tests (or with any other element of Facet A constant), if its content is well stratified according to Facets B and C, then the highest average correlations will be with the rule-inference tests. This may help account for the "finding" that tests "involving greater complexity of mental manipulation are consistently more *g*-loaded than others" (Jensen, 1985, p. 195). The typical calculation aimed at the non-existent *g* is usually a weighted average of the tests, and this will tend to project onto the middle of the radex. Practically any current rule-inference test will tend to have the correlational properties attributed to *g*. Cattell's (1963) attempted — ex post factum — distinction between "fluid" and "crystallized" intelligence can be viewed as an approximation to a priori classification of tests by Facet C.

The radex — and cylindrex — also show how, if the battery is limited to one element of Facet B (say, "verbal"), then the center may generally not be rule-inference. Thus the location of the major principle axis or component of a battery in the test space depends on the regions from which the tests were selected, and cannot be discussed scientifically without knowing the facet design and how it was used. The positiveness phenomenon holds for each region separately, leading to correlational "traps" such as the three pointed out previously. Jensen is aware, of course, of this fact that "different collections of tests will result in somewhat different first principle factors" (1985, p. 200). His advice on how to cope with this is rather incoherent and circular. It can hardly be otherwise in the absence of a clear design of domain facets. Stratified sampling — whether of people or of tests — requires a priori definition of the facets of the stratification.

In another article, Jensen shows that he is aware of the radex of intelligence for article and pencil tests, and of Facet C in particular (Jensen, 1984; esp. pp. 389-390). But even the radex does not escape his propensity for misinterpretation; he erroneously associates regions with "group factors", and even tries to read *g* into the picture. He cites my original article on the radex (Guttman, 1954); but this article explicitly shows — algebraically — how simplexes, circumplexes, etc. *contradict* the hypothesis of a small number of common factors, and the hypothesis of *g* in particular.

Spearman's Objections to Reaction-Time as a Basis of g

A rationale for Jensen's (1985) persistence in invoking the ghost of *g* may lie in his desire to show a biological basis for race differences. If *g* exists, and if it can be shown that *g* has a biological basis, then groups that differ on *g* differ biologically. In Jensen's own words: "If the black-white difference is mainly

a difference in *g*, then a logical first step toward understanding it scientifically would be to understand the nature of *g* itself" (p. 206-208). Jensen's "first step" in this direction is to make studies of reaction-time, discussed in the last two sections of the target article. He concludes: "I believe that Spearman's [*g*] hypothesis has been substantiated in psychometric test data, and that we have made a good beginning to investigating its possible locus in the speed or efficiency of various cognitive processes, as measured by reaction-time techniques" (p. 212). Many of the peer commentaries center on this issue of reaction-time (without remarking on the non-existence of *g*). My own comments will differ from these in two respects. The first will be to cite Spearman's (1932) own position on the matter. The second – presented in the following section – will show how reaction-time can better be regarded as but another variety of mental test item, rather than as a basis for "explaining" psychometric factors of intelligence.

My first comment here goes back to the misleading title of the target article. Spearman (1932) clearly distinguished between his *g* hypothesis and the "nature" of *g*; he made no actual hypothesis about the latter, despite Jensen's (1985) title. And Spearman explicitly rejected mere reaction-time as a possible basis for *g*. To the contrary, Spearman devotes an entire chapter trying to show how "speed of response" is partly a *function of g* (Spearman, 1932, chap. XIV). According to Spearman, Jensen is putting the cart before the horse.

Spearman does devote an earlier chapter to *Proposed Explanations of G* (Spearman, 1932, chap. VII). In particular, he discusses possible physiological bases such as brain energy, plasticity of nervous system, blood supply, and even endocrine glands and respiration. All this he didn't seem to take too seriously, since he concludes that: "from this physiological standpoint, the universal factors would seem to be multipliable almost without limit" (Spearman, 1932, p. 92). It is interesting that here too he explicitly discounts "mere speed" in favor of "plasticity", when discussing the possible role of the nervous system.

Jensen (1985) has hardly been fair to Spearman (1932) by failing to point out that Spearman explicitly rejected reaction-time as a basis for *g*.

Reaction-Time as an Intelligence Item

Spearman early verified that "goodness" and "speed" of response are positively correlated (Spearman, 1932, chap. XIV). Jensen (1985), following many other researchers of reaction-time, devotes the last part of the target article to providing further evidence of this. Accordingly, the Phenomenon of Positive Covariance for mental tests, discussed previously, seems to extend to reaction-time. Does this mean that reaction time itself belongs to the universe

of intelligence tests? Not at all: correlations do not determine content (cf. Guttman, 1981, p. 38) According to the Intelligence Test Definition above, a necessary condition for an item to be classified as an intelligence test item is that its range be from "right" to "wrong". Reaction-time is assessed from "fast" to "slow", and thus fails to satisfy the intelligence condition for the range.

However, the reaction-time items under discussion do satisfy the *domain* condition for an intelligence item: they do ask for performance with respect to an objective rule — for which there is a right answer. It is therefore possible to recast the phrasing of these items to make also the range conform to that of intelligence. For example, a direct phrasing of a reaction-time item might be: "How many seconds does it take subject *s* to give the correct answer to the task?" This can be recast into a sequence of items of the form: "How correct is the answer given by subject *s* within 5 seconds? Within 6 seconds? Within 7 seconds? ..." By scoring "no answer" as more wrong than giving the right answer, the range of each item in such a sequence properly satisfies the intelligence requirement. Since the rephrased items belong to the universe of intelligence test items, the Positive Monotonicity Law should hold for them.

Now, if we count the number of right answers in a sequence such as previously, the total will correlate perfectly with the reaction-time itself. Clearly (cf. Trap 1, previously), if each sequence item satisfies the First Law, so must the sequence sum. Therefore, so must the reaction-time itself — despite its different range. In this sense, reaction-time tests for objective rules are best to be regarded as but a further variety of intelligence test.

The facet of "time" could be added to the domain of the mapping sentence above for the WISC. Should an extended battery of tests be constructed accordingly, it might be hypothesized that the previously cylindrical structure would now become four-dimensional, "time" playing the role of a further axial facet.

In any event, by not paying sufficient heed to a clear definition of what items belong to the universe of intelligence tests, Jensen (1985) has become a cropper. He has merely tested a further variety of intelligence and has not really explored a possible basis for the "nature" of g — even assuming g existed. (Again, extending the Positive Covariance phenomenon to reaction-time holds despite the non-existence of g.

Jensen's (1985) goal would be better served by studying reaction-time of behavior which is *not* assessed to be right or wrong (with respect to an objective rule). One might share Spearman's (1932) skepticism about the prospects of success in this direction. Jensen's intelligence reaction-time correlations are small enough as is. To remove intelligence content completely could be expected to make reaction-time correlations even smaller. This leaves little basis for Jensen's recommendations for "the future of this line of research" (p. 212).

Regionality and the Study of Group Differences

We now come to the heart of Jensen's (1985) motivation for the target article: racial differences. According to Jensen, the question asked by the second "hypothesis per se" is: "which content features or psychometric characteristics of tests are associated with the conspicuous variation in the size of the mean black-white difference on different tests?" (p. 247). The missing chapter of *algebra* of "psychometric characteristics" will be presented in the next sections to follow. Here we shall discuss "content features".

A straightforward way of seeing what kind of test has the greatest difference of means for two given groups is to classify the tests by "kind" and look at the size of difference for each kind. Since SSA shows such a simple regional correspondence with the domain facets of the WISC (and of many other batteries), this can easily be capitalized on to see any further relationship with group differences. Let d_x denote the difference in means on test x for groups I and II (say white and black):

$$d_x = \bar{x}_I - \bar{x}_{II}. \tag{10}$$

For each x in the cylinder of Figure 1, write d_x alongside its point. One can then see which region of the space has the largest differences and which the smallest. Jensen's (1985) discussion would lend one to expect that the largest differences should be along the inner axis of the cylinder, namely with the rule-inference tests. How they should differ within this axis — or with respect to Facet A — is an interesting question. It could hardly have been raised before without SSA and the mapping sentence for the WISC.

It should be an easy matter for Jensen (1985) — and others with data on group difference — to make this plotting of the differences onto the cylinder of the WISC battery. Any lawfulness revealed by such direct data analysis will stand on its own feet, without reference to any supposed factors. And it will be a direct answer to the question posed: what kinds of tests show the largest group differences?

The Irrelevance of the Second Hypothesis

Jensen (1985) attempts to answer the question about "kinds of tests" only in a most convoluted fashion, by bringing in a further hypothesis that he attributes to Spearman. Spearman actually gave it but scant attention (Spearman, 1932, p. 379), and specifically deprecated its utility for the study of biological heredity (Spearman, 1932, p. 380). It is hardly fair to the memory of Spearman to call some minor comments on his part — referring to the work of other

people — his "second" hypothesis, thus giving an entirely false impression of Spearman's order of priorities. For the present discussion, I shall not retain Jensen's labelling, and shall omit Spearman's name in citing the second hypothesis. My analysis is to be taken as a criticism of Jensen and not of Spearman.

The second hypothesis can be stated in the following form.

The Second Hypothesis

Suppose g exists for a given population for a given battery of tests. Suppose the population is divided into two subpopulations, and the differences in means on two tests, x and y, are d_x and d_y (as defined in Equation 10). Then the test with the larger correlation with g will have the larger (standardized) difference. More precisely,

$$\text{if } r_{xg} > r_{yg}, \text{ then } d_x/s_x > d_y/s_y, \tag{11}$$

where $s_x(s_y)$ is a standardizing parameter for $d_x(d_y)$.

The importance that Jensen (1985) attaches to this hypothesis is attested to by his statement: "[the] hypothesis that the magnitudes of black-white mean differences on various mental tests are directly related to the tests g loadings, if fully substantiated, would be an important and unifying discovery in the study of population differences in mental abilities" (p. 197).

In the next section, I shall substantiate this proposition — but not its importance — by showing it to be but a mathematical consequence of the g hypothesis and not at all in need of empirical evidence. I shall prove a purely factor-analytic theorem — the"missing" theorem — of which proposition Equation 11 is an immediate corollary, that gives a much stronger result than Equation 11. Actual proportionality must hold between the loadings and the standardized differences:

$$d_x/s_x = Cr_{xg}, \tag{12}$$

where C is the constant of proportionality given in Equation 22 to follow. Equation 11 is weaker than Equation 12, requiring only monotonicity and not strict proportionality. Consequently, neither Equations 11 nor 12 need the empirical verification to which Jensen devotes so much effort; basically, all that is required to prove Equation 12 is that Spearman's (1932) tetrad condition hold for each of the subpopulations as well as for the total population. (A more general "missing theorem" has consequences for multiple common factors as

well.) Jensen (1985) is mistaken in believing that substantiating Equation 11 would be a new empirical discovery for the study of group differences.

Since Jensen (1985) did not attempt to derive Equation 11 mathematically, he was in no position to say what the s_x should be. He properly points out that this second hypothesis makes no sense without proper standardization. Accordingly, he guessed a formula for s_x (p. 199), and used this guess in his subsequent data analysis. The algebraic proof of the theorem leading to Equation 11 gives a precise formula for s_x, namely Equation 21 to follow. Jensen's guess for the structure of s_x turns out to be well-motivated, but wrong.

As Jensen (1985) points out further (p. 199), there is a certain ambiguity in Equation 11. All told, three populations are under discussion: the total population and its two subpopulations. Each may have different factor loadings on g. To which of the three do the r_{xg} and r_{yg} of proposition Equation 11 belong? Jensen's intuitive answer is: to the subpopulations (which should not differ in their loadings). My algebra confirms this second conjecture of Jensen's. The r_{xg} in Equations 11 and 12 are to be interpreted as the joint values for the two subpopulations.

Not knowing that Equation 11 is but an algebraic corollary, Jensen (1985) treats it only as an approximate empirical statement, and allows for great experimental error around it, leading to imperfect correlation between the loadings and the mean differences. His data show a correlation of .59 (p. 201). But Equation 11 should hold *exactly* if g exists. Jensen's empirical findings are not only superfluous for establishing the second hypothesis, but they actually serve better to *disprove* the existence of g. Jensen tries to explain away the relatively bad fit by considering empirical imperfections; however, the known falsity of the first hypothesis is a more than adequate explanation for the discrepancies from the second hypothesis.

Before going on to the proof of Equation 12, it may be worth making a further comment on Jensen's (1985) data analysis. While Equation 11 only asserts monotonicity, Jensen actually plots straight line regressions. He gives no rationale for such strict linearity. The theorem producing Equation 12 does prove that linearity must hold. Interestingly, Jensen notices in his data that the empirical line goes approximately through the origin, so that the linearity reduces to proportionality (p. 202). This again must be algebraically true, according to Equation 12.

The Missing Theorem on Group Differences

In the history of factor analysis, Godfrey Thomson was perhaps the first to address the problem of the algebraic relations between factor structure of populations and subpopulations (Thomson, 1939). Thurstone does not mention

this problem in his first book, but — following Thomson — devotes a whole chapter to it in his second (Thurstone, 1947, chap XIX). Thomson and Thurstone look at the problem only from the point of view of a certain algebraic way of selecting a subpopulation from the population, and do not arrive at results directly relevant to the second hypothesis that is so central to Jensen's target article. The requisite theorem is missing. Had Thomson or Thurstone gone on to develop this theorem, Jensen might have been deterred from his entire empirical enterprise.

One of the most surprising features of the target article, and of all the commentaries on it, is the complete absence of any mention of the algebraic work of Thomson (1939) and Thurstone (1947) on group differences. True that their pioneering work went in a bit different direction from that directly required for the second hypothesis, but awareness of that work should have stimulated Jensen (1985) into more algebraic thinking that might have saved some of the enormous empirical effort on which the target article is based. It might have prevented him from dismissing out-of-hand any suggestion, made in some peer comments, that the second hypothesis might be but an algebraic consequence of the first (p. 247). Perhaps more important, it might have stimulated doubts as to the utility of factor analysis for the empirical study of group differences at all, and even doubts about the foundations of factor analysis itself.

The missing theorem is a rather immediate consequence of a well-known lemma on covariances for a population that is partitioned into two subpopulations. The lemma can be stated as:

Covariance Lemma

Let p_{I} and p_{II} denote the respective proportions the two subpopulations are of the total population, so that $p_{\mathrm{I}} + p_{\mathrm{II}} = 1$. For any two variables x and y, let $\mathrm{cov}(x,y)$ denote the population covariance; and let $\mathrm{cov}_{\mathrm{I}}$ and $\mathrm{cov}_{\mathrm{II}}$ denote the respective subpopulation covariances. Then,

$$\mathrm{cov}(x,y) = p_{\mathrm{I}}\mathrm{cov}_{\mathrm{I}} + p_{\mathrm{II}}\mathrm{cov}_{\mathrm{II}} + p_{\mathrm{I}}p_{\mathrm{II}}d_x d_y, \tag{13}$$

where $d_x(d_y)$ is as defined in Equation 10.

Equation 13 is essentially a tautology. It is easily verified by expanding each of the three covariances into the form of an expected product minus the product of the expected values.

That the Lemma is directly relevant to our problem is indicated by the explicit appearance of d_x and d_y in the right member of Equation 13. Any

hypothesis as to the factor structures of the three covariance matrices must automatically involve the subpopulation differences in means.

Tautology Equation 13 holds in particular for the special case of variances, when $x = y$:

$$\sigma_x^2 = p_I\sigma_{xI}^2 + p_{II}\sigma_{xII}^2 + p_Ip_{II}d_x^2. \tag{14}$$

where σ is the usual symbol for standard deviation. Note that Equation 14 allows the subpopulation variances to differ, and that the total variance will differ from these if there is a difference in means between the subpopulations.

For present purposes, it will be useful to restate the tautology in terms of correlation coefficients:

$$\sigma_x\sigma_y r_{xy} = p_I\sigma_{xI}\sigma_{yI}r_{xyI} + p_{II}\sigma_{xII}\sigma_{yII}r_{xyII} + p_Ip_{II}d_xd_y. \tag{15}$$

Now, if Spearman's g holds for each of the subpopulations as well as for the whole population, Equation 3 above can be used for rewriting Equation 15, for $x \neq y$, as:

$$\sigma_x\sigma_y R_{xg}R_{yg} = p_I\sigma_{xI}\sigma_{yI}r_{xgI}r_{ygI} + p_{II}\sigma_{xII}\sigma_{yII}r_{xgII}r_{ygII} + p_Ip_{II}d_xd_y \quad (x \neq y), \tag{16}$$

where R_{xg}, r_{xgI}, and r_{xgII} are the respective loadings for the total population and the two subpopulations.

For the next step, let n denote the number of variables in the battery; let $\bar{d}$ denote the mean of the n values of d_x; and let $\bar{R}$, $\bar{r}_I$ and $\bar{r}_{II}$ denote the mean values respectively of $\sigma_x R_{xg}$, $\sigma_{xI}r_{xgI}$, and $\sigma_{xII}r_{xgII}$. Sum both members of Equation 16 over the n - 1 values of index y that are different from x, and divide by n. The result can be written as

$$\sigma_x R_{xg}(\bar{R} - \sigma_x R_{xg}/n) = p_I\sigma_{xI}r_{xgI}\,(\bar{r}_I - \sigma_{xI}r_{xgI}/n) \tag{17}$$

$$+ p_{II}\sigma_{xII}r_{xgII}\,(\bar{r}_{II} - \sigma_{xII}r_{xgII}/n) + p_Ip_{II}d_x(\bar{d} - d_x/n).$$

Now, factor theories are for infinitely large universes of tests; so we can take the limits in Equation 17 as n increases indefinitely, and obtain

$$\bar{R}\sigma_x R_{xg} + p_I\bar{r}_I\sigma_{xI}r_{xgI} + p_{II}\bar{r}_{II}\sigma_{xII}r_{xgII} + p_Ip_{II}\bar{d}d_x. \tag{18}$$

Another equality for the left member of Equation 18 can be obtained by using tautology Equation 15 for the covariance of x with g:

(19) $$\sigma_g \sigma_x R_{xg} + p_I \sigma_{gI} \sigma_{xI} r_{xgI} + p_{II} \sigma_{gII} \sigma_{xII} r_{xgII} + p_I p_{II} d_g d_x.$$

Multiplying Equation 18 through by σ_g and Equation 19 through by $\bar{R}$, and subtracting corresponding members, yield

(20) $$p_I \sigma_{xI} r_{xgI} (\bar{R}\sigma_{gI} - \bar{r}_I \sigma_g) + p_{II} \sigma_{xII} r_{xgII} (\bar{R}\sigma_{gII} - \bar{r}_{II} \sigma_g) + p_I p_{II} d_x (\bar{d}\sigma_g - \bar{R} d_g).$$

Equality Equation 20 still allows for subpopulation differences in correlations with the *g*s. One of the frailties of factor analysis — as Jensen acknowledges — is that the only way it can test for a factor to be invariant across subpopulations is be seeing that its loadings are invariant. Accordingly, for the same *g* to hold for the two subpopulations, it must be that $r_{xgI} = r_{xgII}$ for all x. Let this common value be denoted by r_{xg}. Let s_x be defined as

(21) $$s_x = p_I \sigma_{xI} (\bar{R}\sigma_{gI} - \bar{r}_I \sigma_g) + p_{II} \sigma_{xII} (\bar{R}\sigma_{gII} - \bar{r}_{II} \sigma_g),$$

and let C be defined as

(22) $$C - 1/p_I p_{II} (\bar{d}s_g - \bar{R} d_g).$$

Then Equation 20 can be rewritten as

(23) $$s_x r_{xg} + d_x / C.$$

Equation 12 above follows immediately from Equation 23, so the promised proof is completed.

These results can be stated as:

The Missing Theorem

Suppose a given population is divided into two subpopulations. Suppose, for a given battery of tests, Spearman's (1932) tetrad criterion holds for each of the three observed correlation matrices. If the two subpopulations have the same correlations for the tests with their respective single common factors:

(24) $$r_{xgI} = r_{xgII},$$

then the proportionality Equation 12 holds, where r_{xg} is the common value in Equation 24, s_x is as defined in Equation 21 and C is as defined in Equation 22.

Group Differences and the Foundations of Factor Analysis

As remarked previously with respect to the Second Hypothesis, two features there need pinning down for the hypothesis to be well-defined. One is the formula for s_x, and the other the meaning of "r_{xg}". Jensen's (1985) guess for s_x is simply the *quadratic* mean of the subpopulation standard deviations:

$$(p_I\sigma_{xI}^2 + p_{II}\sigma_{xII}^2)^{1/2}.$$

The correct formula, as given by Equation 21, is a weighted *arithmetic* mean, where the weights could hardly be guessed at in advance of an algebraic analysis. In the present case, there may be little practical difference between the two formulas, but this could not be known until the correct formula was actually produced. More important, not just s_x, but the whole proposition Equation 12 has now been derived algebraically, removing it from the realm of empirical conjecture to which Jensen (1985) consigned it.

The second feature, for which Jensen's (1985) guess *is* correct, is the choice of r_{xg}, namely the common value in Equation 24. The need for making such a choice raises a basic question for all of common factor analysis. The factor loadings for the total population must in general be different from those of its subpopulations. Jensen interprets this as being "contaminated by [sub]population differences on the various tests" (p. 199). But each subpopulation can be further subdivided, say by sex. Then the factor loadings of the blacks and the whites are also "contaminated". Further subsubpopulations can be obtained by dividing according to age, etc., yielding more and more "contamination". An obvious question awaiting an answer is: is there a basic partition into subpopulations that will yield "pure" factor loadings?

My own a priori guess as to the proper choice for r_{xg} in Equations 11 and 12 would have been the wrong one: the loadings R_{xg} for the total population. The rationale for such a guess is as follows. Condition Equation 11 must hold also when x is replaced by g. Since $r_{gg} = 1$, this must be greater than any r_{yg}. Hence, from Equation 11 – rewriting with x for y:

$$d_g/s_g > d_x/s_x \text{ for all } x. \tag{25}$$

The standardized difference in means on g for the two subpopulations must be greater than that for any observed test. Now, calculating factor loadings from a correlation matrix leaves open the question of units for the factor scores. Conventionally, the means are set equal to zero and the standard deviations equal to 1. Hence, if one were to consider only factoring the subpopulation matrices (as Jensen, 1985, intuitively suggests), this would leave open the

units for the *g*s for the two subpopulations — even when the equality of loadings in Equation 24 held. One could set both means of the *g*s conventionally equal to zero, thus violating Equation 25. To make meaningful a difference in means (and in standard deviations for *g* between the two subpopulations, the units must come from some other source than factor analyses of just their own observed matrices. The "obvious" source is the total population. Indeed, if the requisite *g* scores could not be obtained via the total population, it would hardly make sense to think they nevertheless exist in the same people when divided into subpopulations. So one must think of the *g* scores as coming from the "contaminated" total population. Indeed, this is the point of departure of the work of Thomson (1939) and of Thurstone (1947) on the selection of a subpopulation from a population. They start off with the known structure of the total population.

Here, indeed, is a paradox that raises a question as to the sheer empirical existence of factor scores, apart from any question of substantive meaning. The *g* scores of the "missing theorem" on "uncontaminated" subpopulations must themselves be "contaminated".

But there is no need to think here in terms of "contamination". The algebra involved simply says that a population's (or subpopulation's) factor loadings are a function of the differences in test means between more and more refined subpopulations. This gives a perspective on the meaning of factor loadings that extends the scoring paradox to the meaning of factors themselves for empirical science. Can factors be universal? How can one hypothesize in advance that two subpopulations are of the type to satisfy Equation 23? Such issues arise for any common factor theory, whether Spearman's *g* exists or not. It would take us too far afield to do more than raise the questions here. In a real sense, study of the problem of group differences may have more to contribute to the theory of factor analysis than factor analysis has to contribute to the study of group differences.

Summary

It is disheartening to realize that Jensen's (1985) diligent research has succeeded only in building a house of cards. He has distorted the basic concepts of factor analysis, doing special hurt to Spearman (1932). He has labored prodigiously to produce empirical data which are superfluous, since the proposition in question — the second hypothesis — is algebraically true. He has worked equally hard to show that reaction-time may be part of a physiological basis for *g*, whereas his evidence — of positive correlations — merely reflects the well-established First Law for intelligence test scores. He overlooked the known cylindrical portrayal of the WISC correlation matrices

[made possible jointly by SSA and the mapping sentence for defining the WISC], and hence did not realize that this portrayal could be used directly to relate test differences in means of subpopulation to the types of tests. His recommendations for future research on the "nature" of *g* (p. 212) neglect to mention the need for sharp definitions of content, like the 1973 Intelligence Test Definition, in order to avoid falling into the same redundancy as happened with reaction-time. As things stand, the target article has failed in all its main objectives. Analyzing it shows how the failure is a result of the irrelevance of factor analysis to the study of group differences. Jensen's factor-analytic efforts have shown nothing about group differences not already apparent from the well-known empirical differences on the tests themselves.

References

Adler, N. & Guttman, R. (1982). The radex structure of intelligence: A replication. *Educational and Psychological Measurement, 42*, 737-748.

Anastasi, A. (1983). Evolving trait concepts. *American Psychologist, 38*, 175-184.

Canter, D. (Ed.). (1985). *Facet theory: Approaches to social research.* New York: Springer-Verlag.

Catell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Cattell, R. B. (1985). Intelligence and *g*: An imaginative treatment of unimaginative data. *Behavioral and Brain Sciences, 8*, 227.

Gratch, H. (Ed.). (1973). *Twenty five years of social research in Israel.* Jerusalem Academic Press.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53*, 267-293.

Guilford, J. P. (1967). *Nature of human intelligence.* New York: McGraw-Hill.

Guttman, L. (1954). A new approach to factor analysis: The Radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 216-348). Glencoe, IL: Free Press.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common factor theory. *British Journal of Statistical Psychology, 8*, 65-81.

Guttman, L. (1964). The structure of interrelations among intelligence tests. In *Invitational conference on testing problems.* Princeton, NJ: Educational Testing Service.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika, 33*, 469-506.

Guttman, L. (1980). Integration of test design and analysis: Status in 1979. *New Directions for Testing and Measurement, 5*, 93-98.

Guttman, L. (1981). What is not what in theory construction. In I. Borg (Ed.), *Multidimensional data representation: When and why.* Ann Arbor, MI: Mathesis Press.

Guttman, L. (1986). Coefficients of polytonicity and monotonicity. In *Encyclopedia of the statistical sciences.* New York: Wiley.

Jensen, A. R. (1984). The black-white difference on the K-ABC: Implications for future tests. *Journal of Special Education, 18*, 377-408.

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

Jones, L. V. (1985). Golly *g*: Interpreting Spearman's general factor. *Behavioral and Brain Sciences, 8,* 233.

Koop, T. (1985). Replication of Guttman's structure of intelligence. In D. Canter (Ed.), *Facet theory: Approaches to social research.* New York: Springer-Verlag.

Levy, S. (1985). Lawful roles of facets in social theories. In D. Canter (Ed.), *Facet theory: Approaches to social research.* New York: Springer-Verlag.

Lingoes, J. C. (1973). *The Guttman-Lingoes non-metric program series.* Ann Arbor, MI: Mathesis Press.

Peled, Z. (1984). The multidimensional structure of verbal comprehension test items. *Educational and Psychological Measurement, 44,* 67-83.

Raveh, A. (1978). Guttman's regression-free coefficients of monotonicity. In S. Shye (Ed.), Theory construction and data analysis in the behavioral sciences. San Francisco, CA: Jossey-Bass.

Schlesinger, I. M., & Guttman, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin, 71,* 95-100.

Schmid, J, & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61.

Spearman, C. (1932). *The abilities of man.* New York: AMS Press.

Steiger, J. H., & Schönemann, P. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences.* San Francisco, CA: Jossey-Bass.

Sternberg, R. J., & Powell, J. S. (1982). Theories of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence.* Cambridge, MA: Cambridge University Press.

Thomson, G. H. (1939). *The factorial analysis of human ability.* Boston, MA: Houghton Mifflin.

Thurstone, L. L. (1935). *The vectors of mind.* Chicago, IL: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.

Tziner, A., & Riemer, A. (1984). Examination of an extension of Guttman's model of ability tests. *Applied Psychological Measurement, 8,* 59-69.

Wilkes, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23.