te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82*, 675–687.
te Nijenhuis, J., & van der Flier, H. (2001). Group differences in mean intelligence for the Dutch and Third World immigrants. *Journal of Biosocial Science, 33*, 469–475.
te Nijenhuis, J., Tolboom, E., Resing, W., & Bleichrodt, N. (in press). Does cultural background influence the intellectual performance of children from immigrant groups? Validity of the RAKIT intelligence test for immigrant children. *European Journal of Psychological Assessment.*
Todd, T. W. (1923). Cranial capacity and linear dimensions, in white and Negro. *American Journal of Physical Anthropology, 6*, 97–194.
Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences, 10*, 573–576.
Vernon, P. E. (1982). *The abilities and achievements of Orientals in North America.* New York: Academic.
Vint, F. W. (1934). The brain of the Kenya native. *Journal of Anatomy, 48*, 216–223.
Weinberg, R. A., Scarr, S., & Waldman, I. D. (1992). The Minnesota Transracial Adoption Study: A follow-up of IQ test performance at adolescence. *Intelligence, 16*, 117–135.
Zaaiman, H., van der Flier, H., & Thijs, G. D. (2001). Dynamic testing in selection for an educational programme: Assessing South African performance on the Raven Progressive Matrices. *International Journal of Selection and Assessment, 9*, 258–269.
Zindi, F. (1994). Differences in performance. *The Psychologist, 7*, 549–552.

Chapter 10

Sex differences in *g*

Helmuth Nyborg

1. Introduction

Even a quick review of the research literature reveals a fundamental disagreement about the existence of a sex difference in general intelligence. It is imperative in this connection to clearly distinguish between *general intelligence* and *intelligence in general*. The use of these terms will become much clearer later in the chapter. For now it suffices to say that general intelligence can be estimated by the higher-order *g* factor score, that can be obtained by factor analyzing the pattern of correlations among test items. In clear distinction, *intelligence in general* — or total IQ score — can be obtained by summing the standardized item scores.

Empirical evidence abounds both for and against a difference in general intelligence. This chapter tests the hypothesis that there is actually a small male average superiority in general intelligence but it can be seen only if the most sophisticated contemporary methodology is brought into action. It is an interesting twist to this test that Arthur Jensen promotes the advanced tools needed to identify the difference, but at the same time comes to the conclusion that there is no consistent sex difference.

The chapter first lines up the positions, evaluates the methodological and analytic qualities of selected studies, and then comes to the conclusion that there is in fact a small difference in favor of males. It is shown how even such a small average sex difference can take on practical importance at the high end of the general intelligence distribution scale. Finally, some speculations are presented on the likely future of sex difference research.

1.1. The Positions

1.1.1. There is no sex difference in general intelligence! The possibility of sex differences in intelligence has fascinated researchers, philosophers and lay people for millennia, and they have aired their interest in such different places as in a religious

The Scientific Study of General Intelligence: Tribute to Arthur R. Jensen

ISBN: 0-08-043793-1

ancient Sanskrit paper informing us that: "Ten shares of talk were handed down to earth; the nine went to the women", in literally hundreds of contemporary books and thousands of scientific articles, in Ladies' magazines, and in myriads of radio and TV shows.

Often, the conclusion reached is that there indeed are real sex differences in first order group factors like verbal or spatial abilities, but these are not terribly important. These lower order factors usually have only moderate to low validity in predicting sex differences in achievement in school, jobs or life, when compared to the considerable predictive validity of higher order general intelligence. The important point is, they say, that most studies find no real sex difference in a general intelligence (e.g. Brody 1992; Halpern & LaMay 2000; Neisser *et al.* 1996).

The theoretical implications of this widespread view cannot easily be overestimated. No other constructs in psychology come even close in predicting one's final level of education, occupational status and income, one's likely belongingness to the administrative or political elite or, conversely, to predict the risk of finding oneself caught in a wide range of unfavourably economic, social and criminal life circumstances (e.g. Herrnstein & Murray 1994). Researchers of various stripes usually have no difficulties in admitting the male over-representation at most societal top positions. However, given that there are no sex differences in general intelligence, they must explain this male superiority by "old boys network", unsavoury tradition, unfair differences in female opportunity, a lack of female role models, learned helplessness, male oppression, or socially induced differences in motivation or personality. The possibility that genes, hormones, neurobiology or evolutionary history may provide part of the explanation of a sex difference in general intelligence accordingly needs little consideration, or may even call for active resistance from the academic left (Gould 1996, see Chapter 20 in this volume for further details). The subject index for the authoritative *Handbook of Intelligence: Theories, measurements and applications* (Wolman 1985), does not even have an entry to sex differences.

1.1.2. There is a sex difference in general intelligence! Defenders of this opposite view hold that it is, in fact, directly counterintuitive to assume that there is no sex difference in general intelligence. They point to good practical and theoretical reasons to back up their point.

On the practical side, they refer to the vast male over-representation in top positions in education, occupations, and in the social power structures. These areas no doubt call for capacity to deal with high degrees of complexity. Moreover, capacity to deal with complexity is just another way to define general intelligence. It would therefore, according to their view, be downright counterintuitive to assume intellectual equality among the sexes. The male over-representation in most elites will naturally raise the suspicion of a higher general intelligence, everything else equal.

Theoretically speaking we should also expect a male superiority in general intelligence. This idea is based on a paradox, the underlying logic of which cannot easily be dismissed. Thus, most experts agree that general intelligence correlates positively with head size, ranging in size from $r = 0.1$ to 0.45. Aside from measurement error, the differences in correlations depend essentially on whether the measure is based

on simply taping head circumference or on the more exactly measured brain volume by modern imaging techniques. The rule seems to be: the more exact the measurement, the higher the IQ-brain size correlation (see Chapter 6 in this volume for details). Given this fact and given the common if debatable assumption, that males and females do not differ in overall intelligence, one would obviously expect to see on average equal head size or brain volume in males and females. This is not what we see, however. Males have larger heads with more brain tissue, on average of course, than females, quite as expected from a higher general intelligence (Ankney 1992, 1995; Lynn 1994, 1999; Rushton 1992).

This so-called anomaly has elicited contrasting interpretations. Lynn (1994, 1999) argues, for example, that there really is no problem here. Having averaged the IQs of a number of studies, he found that the male lead in general intelligence amounts to 3.8 IQ points. This value corresponds to a male SD advantage of 0.3 in intelligence. Lynn then demonstrated that the well-documented sex difference in brain size actually predicts the observed male average IQ lead closely enough to solve the apparent paradox. The details of Lynn's prediction are presented in Table 10.2. Jensen (1998: 541–543) disagrees with Lynn's interpretation, and suggests that perhaps there is a greater neural "packing density" in the female brain. This interpretation is, in turn, contradicted by an observation by Pakkenberg & Gundersen (1997). Applying a new neuronal counting technique they found equal packing density throughout male and female brains. Moreover, the average female brain contains 15% fewer neurons than the male brain. Lynn takes this to support for his particular interpretation, given the reasonable premise that more neurones are needed for a more efficient brain (even though one should always keep in mind that more is not always better!)

1.2. Diagnosing the Main Problem

As leading scientists disagreed about the existence of a genuine sex difference in general intelligence and also used different methods I began to suspect that the disagreement could be explained by the use of less than optimal methods for studying the sex difference.

On the one side, there was the longstanding tradition of summing standardized scores, an approach that loses important information on its way. Most clinicians and researchers sum standardized subtest scores to reach an overall intelligence score, as for example in the widely used WAIS or WISC IQ tests for adults and children, developed by David Wechsler in the mid-twentieth century. Two subscale scores — Performance and Verbal IQ — can be combined to form a Full Scale IQ score (FSIQ). Wechsler explicitly dismissed test items greatly favouring one sex when constructing the test, and then balanced out the remaining items so as to avoid male or female superiority in overall IQ. Males often lead in Performance IQ and females tend toward superiority in at least some verbal abilities. Considering Wechsler's manipulation with items in the construction, it is in fact a bit surprising to find that the average of recent studies points to a male superiority of 3.8 points in total FSIQ (Lynn 1997). The fabrication of the test and the use of summed scores leave us entirely in the dark about the origin of the observed sex difference in the WISC and WAIS. Is it due to bias in item selection or is it due to a true

sex difference in general intelligence. The only way to find out is to apply analytic techniques that go well beyond summing scores.

On the other side, there was the quite sophisticated factor analytic approach.

Perhaps both types of approach may produce contaminated measures of general intelligence. Perhaps the main problem is that the key measure of general intelligence falls victim to fatal contamination. Perhaps we are looking for a petite sex difference, that will reveal itself only if we methodologically step up from simple summed scores, over the application of analytically speaking quite sophisticated psychometrics, to ultimately reach a position where we attain an uncontaminated and trustworthy higher order factor of general intelligence. Phrased differently, perhaps the fragile sex difference will appear reliably only after successful derivation of an uncontaminated measure of general intelligence.

The grand master of psychometrics, Jensen wrote in 1998 (p. 532) that the study of sex differences in general intelligence is "technically the most difficult to answer . . . the least investigated, the least written about, and, indeed, even the least often asked". Perhaps methodological uncertainty would explain why the field has for so long been beleaguered by confusion, occasional glimpses of clarification, wildly differing interpretations, and the hasty formation of conclusions not rarely based more on what "what ought to be" than on "what is" sexist attitudes.

Given the present discrepancy in opinions and the worries over the methodology, it became mandatory to ask how the methodology can be improved to decide the difficult question of the existence of a sex difference in general intelligence. A step on the way was to develop a simple questionnaire for ranking studies of sex differences in accordance with their analytic capability to avoid making type 1 or type 2 errors. This may enable us to grade studies in accordance with their potential for safely identifying even a subtle sex difference.

However, even having accomplished that, there are further methodological problems that researchers may run into when entering the minefield of sex differences. As they hold the potential to degrade the quality of the studies, I first name and use them to establish criteria for the proper scientific approach to sex difference research on general intelligence. Several studies are then measured up against the criteria, and ranked in accordance with how well they conform to them. This evaluation becomes the basis for deciding how much confidence we can ascribe to their conclusions about the sex difference.

1.3. Further Problems

1.3.1. Ideology gone awry It is a sad fact that scientists finding a sex difference in intelligence too often become woven into an odd struggle, characterized by direct personal attacks or having to deal with tongue-tied politically correct terminological anomalies. The latter involves accusations of believing/postulating/wrongly assuming that he/she has found a gender/sex difference in intelligence/aptitude/achievement/educational challenges rather than sticking to the numbers. The apparently unavoidable and pervasive influences of strong ideological and emotional loadings on the matter may force herself/himself into the process of academic and/or personal survival. The nature of this matter is discussed more fully in Lynn (2001), Segerstråle (2000), and summarized in Chapter 20 in this volume. Suffice it to say that the study of sex differences in general ability has long been hampered by ideology run amok, by academic or personal intimidation of researchers finding a difference, and by the near absence of research specifically capable of solving the problems.

1.3.2. Ambiguous definitions A number of critiques point to the fact that there is little agreement about how to define intelligence. Therefore, they argue, even if a sex difference is found, it cannot be trusted. We would not know what it was all about. It is food for thought to realize that one of the truly great pioneers in psychometrics, Spearman, addressed this problem as far back in time as around 1900 (e.g. Spearman 1904, 1923, 1927; Spearman & Jones 1950). Unfortunately, even if he came of great age, he did not live long enough to see his important points getting generally accepted in the scientific community. This is all too bad because, without a full understanding of his particular reasoning on the matter, it is virtually impossible to see how easy and essential it is to substitute vague concepts of intelligence in general with proper operational definitions, and thereby realize how futile the whole discussion was all the time. Spearman's frustration clearly shines through in his report on reading a paper from a symposium: "Intelligence and Its Measurement" (The Editors 1921). The paper made it obvious that fourteen leading researchers featured fourteen different definitions of intelligence. Spearman reacted with despair: "Chaos itself can go no farther . . . 'intelligence' has become a mere vocal sound, a word with so many meanings that finally it has none" (1927: 14). Sixty-five years later Sternberg & Detterman (1986) convened a symposium with the aim to answer the very same question: "What is Intelligence?" Now it was Jensen's turn to pass judgment on the report of the meeting. His conclusion was as depressing as the one Spearman had reached: "The overall picture remains almost as chaotic as it was in 1921" (1998: 48).

1.3.3. Competing "intelligences" New definitions of intelligence continue to see the day of light. This is without doubt a sign of impatience with prevailing definitions. Unfortunately, Spearman's early ground work and Jensen's (1998) later development of a proper theory for the objective measurement of general intelligence (see below) seem either ignored by the inventers of the new intelligences or are met with persistent attempts to deny their methodological and practical validity (see Brody (in press) and Gottfredson (in press)). Several recent varieties of competing theories feed in part on the disagreement about how to define, measure or explain intelligence. One of these is Howard Gardner's (1983, 1993) model for *Multiple Intelligence*. Another is *Emotional Intelligence* by David Goleman (1995). A third is *Triarchic Mind* by Robert Sternberg (1988). Without going into details, three considerations about the alternatives are relevant here: a sex specific, a general theoretical, and an operational aspect.

With respect to sex, none of these new intelligences tells us anything useful about the question of sex differences in general ability. For that reason alone they can safely be sidestepped in the present chapter. A more general theoretical concern is that many contemporary uses of the term intelligence are so vague as to be of little use in a

scientific study, a point Jensen (1998) discussed at some length in his Chapter 3 entitled: 'The Trouble with "Intelligence"' (but see also Jensen 1987, 1993, 1994a). Most importantly, a precise linguistic definition is not at all needed for a proper operational approach to the question of whether there is a sex difference in general intelligence, as will be illustrated shortly.

1.4. Summary

Previous research on sex differences in intelligence is characterized by exorbitant confusion due to a number of factors. One factor is that widely different sex-based ideologies weigh down the field. Other important points are the disagreement about how to define and measure general intelligence, and the undeniable existence of contrasting and paradoxical findings. It is in such situations a troubled field finds itself in desperate need for a knight in shining armour with the brainpower and vision needed to see what has to be done. Arthur Jensen is just that kind of person. He has the intellectual power and also musters the rugged personality and professional integrity, without which no battle can ever be won when sailing through the troubled waters of sex differences in general intelligence. He not only cut through the emotional parts, but also refined the methods and perspectives, and thereby changed the field radically. We shall discuss in detail how he accomplished this. This chapter focuses in particular on the use of the tools he recommended.

2. Clever But Disengaged

2.1. Jensen as a Slow Starter

Given that Jensen clarified vital aspects of the study of sex differences in intelligence, it is quite surprising to realize in hindsight that he actually entered the area rather late in his professional career. His first publication went to print in 1955 and was on an entirely different matter, followed by a series of works on aggression in fantasy, projective techniques, and learning. Full sixteen years went by before he in 1971 gradually began to close in on the area. It was then in terms of a possible "race X sex X ability" interaction, a finding he later dismissed. Apparently as a side effect, Jensen then wrote theoretical notes on sex linkage and race differences in spatial ability (1975, 1978). His classical book *Bias in mental testing* (Jensen 1980, ch. 13) naturally deals more with sex bias than with sex differences. Jensen states that, like racial and social-class differences, the question of a sex difference in selection rates ". . . has two main aspects: true differences in ability versus artifactual differences due to bias in the tests . . .". Summarizing the outcome of studies between 1966 and the late seventies, he concluded that ". . . a *majority* of the studies find no sex differences large enough to be significant beyond the 0.05 level", and ". . . when significant sex differences *are* found, they never consistently favor males or females for any given ability . . .". He notes that

Maccoby (1966) reached a similar conclusion from her review of pre-1966 studies. Pondering over the reality of sex differences in the ability realm, Jensen (1980: 622) found that they ". . . are a relatively small-magnitude phenomenon as compared with racial and social-class differences . . .". The inconsistencies among studies suggest that they ". . . are complexly determined and are conditional on a number of other factors, such as age of the subject, educational level, regional differences, and secular trends". The first report in which Jensen directly addressed the question of a sex difference on the WISC-R came as late as in 1983, and was followed by a commentary on arithmetic computation and reasoning in pre-pubertal boys and girls (Jensen 1988). Again seemingly in passing, Jensen later commented on the previously mentioned sex differences in head size and related differences in intelligence (Jensen 1994b; Jensen & Johnson 1994).

It may very well be that Jensen's interest in the study of sex differences in intelligence was tempered by the many inconsistencies in the results. This may explain why it took him so long to tackle the topic and pass a devastating judgment. Even then, it was not at all a spin-off of own interests. To the contrary, he was explicitly asked to add a chapter on sex differences to a manuscript, long time previously submitted to a major publisher for evaluation. Ironically, that publisher eventually declined to publish what later became probably the best book ever written on intelligence — *The g factor: The science of mental ability*. Praeger Press published the book in 1998. Luckily, the relatively short chapter on sex differences survived the transfer. On just 13 pages Jensen demonstrates his characteristic perfectionism, his preference for ruthless empiricism, and his blessed lack of taste for easy compromises.

2.2. The Rude Awakening

Jensen's characterization of previous research in the area of sex differences was devastating. He concluded:

> Past studies of a sex difference in general ability have often been confounded by improper definitions and measurement of "general ability" based on simple summation of subtest scores from a variety of batteries that differ in their group factors, by the use of unrepresentative groups selected from limited segments of the normal distribution of abilities, and by the interaction of sex differences with age-group differences in subtest performance. These conditions often yield a mean sex difference in the total score, but such results, in principle, are actually arbitrary, of limited generality, and are therefore of little scientific interest. The observed differences are typically small, inconsistent in direction, across different batteries, and, in above-average samples, usually favor males" (1998: 531).

This is simply another way of saying that most previous studies of sex differences in intelligence fail to obey strict scientific criteria, and that progress in the field depends on better definitions and methods. The rest of this chapter is devoted to a discussion of what this means.

2.3. Proper Criteria

Jensen's incisive characterization of previous studies, developed on top of Spearman's original insights, suggests that future studies of sex differences in general ability must conform to the following objective principles:

- Make sure samples are truly representative;
- Present a proper operational definition of intelligence;
- Incorporate a multitude of tests that differ widely in content;
- Go analytically well beyond simple summed scores; and
- Control for potential confounders (e.g. contamination by group factors or sex-age interaction).

3. The Proper Study of Sex Differences in Abilities

This section first expands on the above criteria, and then scrutinizes how closely a number of methodologically quite different studies come to conforming to the criteria. The overall purpose of this maneuver is first, to illustrate that studies can in fact be ranked by quality, so that we can grade the trustworthiness of their conclusions with respect to the existence of a sex difference in general ability, and second, to see whether the claim of a sex difference survives closer scrutiny.

3.1. Criteria

3.1.1. Subject sampling It has been argued that the typical greater male variability in general ability poses a special problem that makes proper subject sampling critically important. A sample restriction towards the high end of the bell-shaped (Gaussian) curve would for example favor male superiority, whereas sample restriction towards the left side would favor female supremacy. In both cases this would misrepresent a true sex difference in the general population. More generally, a failure to recognize the typically greater male variance in test scores ". . . may cause both the direction and magnitude of the mean sex differences in test scores to vary across different segments of the total distribution for the general populations. The observed sex difference will therefore often vary across groups selected from different segments of the population distribution" (Jensen 1998: 536). With respect to educational bias this implies, for example, that samples should preferably be drawn from populations of elementary school children, because increasingly harsh sample restrictions apply the further we go up the ladder from elementary levels to junior high and beyond. A further consideration here is whether the male/female variance ratio on various subtests relates in balanced samples to the subtests' *g* loadings.

Allik *et al.* (1999: 1140–1941) dispute the importance of sample restriction based on higher male variability. Their main argument is based essentially on Feingold's (1992) extensive review of variability on the national norms of several standardized test batteries. They stress the fact that the theory of greater male variability is by no means conclusively established. Then again, it is worth noting that Feingold's own conclusion was that males were consistently more variable than females in quantitative reasoning, spatial visualization, spelling and general knowledge, mostly abilities that are heavily *g*-loaded. Anyway, we may concur with Allik *et al.*'s recommendation that further evidence is needed before we draw firm conclusions, as well as with Feingold's notion that ". . . sex differences in variability and sex differences in central tendency have to be considered together to form correct conclusions about the magnitude of cognitive gender differences" (p. 61).

3.1.2. Test item variation The proper study will include a minimum of nine tests that all differ widely in content area. The more varied the tests, the more likely it is that bias in one direction cancels out bias in another direction, according to classical test theory. As mentioned previously, verbal and spatial tests typically benefit females and males differently; and their simultaneous presence in a test battery would tend to balance out the sex biasing effects.

3.1.3. Lexical definitions With respect to the importance of definition for measurement, Jensen (1998) suggests, for reasons detailed in *The g Factor*, that we better give up all talk about "intelligence" or "intelligence in general". These terms are used by too many in too many different contexts to be of any scientific use. It is important to realize that there are basic differences between, on the one side:

> ". . . the simple sum or mean of various subtest scores [which] is a datum without scientific interest or generality." or "ability in general" . . . "an arbitrary variable that fails to qualify conceptually as a scientific construct", and,
>
> on the other side:
>
> "general ability, defined as g, [which] rests on the correlations among test scores" (Jensen's 1998: 537 emphasis).

In other words, summed or averaged subtest scores are no scientifically acceptable alternatives to measures of general ability based on inter-test correlations. Of course, in practice we might use IQ scores as a reasonable proxy to general intelligence *g*, because most standardized IQ tests load fairly high on *g*. However, in the process of establishing whether there is a sex difference, no IQ test results will suffice as a sole basis for deciding whether an observed difference in general ability *g* is real, or rather a mirror of biased test item composition.

None of the newly defined "intelligences" will solve this problem. Sternberg's (1988) Triarchic, Practical or Successful intelligences have not yet demonstrated proper construct and predictive validity (see Chapter 19 in this volume), and the various subcomponents of the theory have also been criticized (Kline 1991; Messick 1992). Four of Gardner's Multiple "intelligences" (Gardner 1983, 1993) correlate closely enough to suggest considerable redundancy, and they seem to mainly tap general ability *g* (linguistic, logical-mathematical, spatial and musical), whereas the remaining "intelligences" (intra-personal, inter-personal and bodily-kinesthetic) neither inter-correlate

well nor do they correlate noticeably with the first four "intelligences". There is also the problem that the latter "intelligences" do not reflect *g* well, and that their predictive validity has not yet been documented. Goleman's (1995) Emotional "intelligence" seems based more on psychobiographical anecdotes than on solid data obtained through nationally representative samples. No doubt, psycho-biographic evidence can be an interesting starting point, but it is a long shot from a serious empirical mission to establish an ability test. Left to itself, it surely does not suffice as a basis for sweeping generalizations, and it in no way substitutes for the predictive validity of this "intelligence". Finally, none of the new intelligences seem to satisfy obligatory criteria for legal or moral use (see Chapter 19 in this volume).

It is therefore a relief to realize from Jensen's many contributions over the years, culminating in *The g Factor* book (1998) that a precise lexical definition of intelligence is really not needed. It is no more a must than is the exact lexical definitions of time, space and gravity. What all these constructs really need is an operational definition, which will give us an idea of what kind of reality lies behind the constructs.

3.1.4. Operational definitions The problem of how to properly estimate *g* is actually by and large solved by now. A chapter on sex differences is not the right place to discuss in details what a "good" *g* is, and the reader is referred to Jensen (1998), Jensen & Weng (1994), or to Carroll (1993, or Chapter 1 in this volume) for detailed expositions. It suffices to say here that all the different factor analytic solutions that allow for the existence of a *g* factor will identify *g*. Thurstone's simple structure rotation method is the exception to the rule, because it expressly forbids the appearance of a *g*. Even if all the *g* variance remains in the factor structure all the time, the mathematical solution does not allow it to appear as a separate higher order factor. The considerable predictive validity of psychometric *g*, and the fact that the heritability estimate for the various *g*'s *increases* over the life span (i.e. it becomes larger with time rather than smaller, as previously expected), when derived as a second or third order factor, also speaks well for its usability, as does its multiple correlations with a variety of biological traits.

However, and this is vitally important in the present context: I have found that most factor analytic solutions are perhaps less than optimal in the search of a sex difference in *g*. The average sex difference may be rather small, and this means that it has to be protected by all means from the masking effect of confounding by other ability dimensions, or it will not be possible to identify at all. A brief comparison of the various factor analytic approaches demonstrates this analytic point, and the forthcoming grading of studies illustrates the important empirical implications.

3.1.5. Principal component (PC) and principal factor (PF) analyses A drawback of PC and PF analyses is that they are somewhat sensitive to test bias. If the test battery contains a predominance of, say, visuo-spatial tests, second or third order PC1 and PF1 *g* run the risk of being contaminated by this over-representation, perhaps even as much as to side with general intelligence *g* in the case of a grossly biased test battery. In that case we would get a definitely false impression of a male superiority in general ability *g*. Contrariwise, a surplus of verbal fluency tests in the battery would easily induce the illusion of a female superiority in *g*. The mathematical reason for this outcome is straightforward: The PC1 and PF1 *g* factors derive directly from inter-test correlations. As such, they reflect to some extent the kind of abilities the test items cover.

3.1.6. Hierarchical factor analysis (HFA) Hierarchical factor analysis is much less sensitive to this error of over-sampling same- or similar-ability type tests. One thus begins the analytic process by first identifying the primary or group factors using PF (or in some special cases PC) analysis, and forcing an oblique rotation of factor axes to determine their correlations. Another step is to derive the second-order (or third order in the case of a large and varied test battery) *g* from correlations among the group factors at the primary level. The bottom of the hierarchy of factors is, in other words, the least general: the factors there arise on the basis of correlations between only a few of the tests (say, a number of interrelated verbal tests or some interrelated visuo-spatial tests). The factors at the next higher level are a function of the correlations among a few group factors (say, verbal abilities or visuo-spatial abilities.) The highest second or third level is inhabited by the general ability factor *g*, which is a function of what is common to all the lower order group factors and test. In other words, because the sources of variances due to test specificity and possible group factor biases are sorted out already at lower levels, the higher order *g* factor emanates as a largely uncontaminated function of general ability, reflecting mostly the variance that is common to all factors below.

3.1.7. Orthogonal Schmid-Leiman rotation The HFA analytic solution can be optimized by including a Schmid-Leiman (SL) transformation (Schmid & Leiman 1957). This procedure orthogonalizes all factors between and within all levels in the hierarchy, making them totally uncorrelated. One advantage is that the structure is easier to interpret. Another advantage is, that it prohibits the appearance of a general ability factor where none is present in a correlation matrix. This could happen in the case of the PC and PF approaches.

3.1.8. Test for differences There are basically three ways to test for the significance of an observed sex difference in *g*, here presented in increasing order of scientific interest.

The least informative method is to factor *g* by any of the relevant factor analytic methods, and then simply use a *t*-test to see if the male-female difference is significant.

The next step, which is proposed by Arthur Jensen (and explained in details in Appendix B in Jensen 1998: 589–591), is to determine whether the vector of disattenuated *d* values correlates significantly with the vector of disattenuated *g* loadings. The sex difference on each subtest may, for example, be expressed in terms of effect size *d* (calculated according to the formula: $d = (X_M - X_F)/\text{sigma}$, where X_M is the male mean, X_F is the female mean, and sigma is the pooled SD). The *d* effects are now arranged according to size in a vector matrix, preferable after correction for attenuation or reliability. The next step is to determine whether the vector matrix of attenuated *d* values correlate significantly with the vector matrix of likewise attenuated test *g* loadings. Both Pearson product-moment correlation and Spearman's rank-order correlation coefficients are calculated, as a divergence between the two coefficients may

reveal a hidden non-linearity. In case the magnitude of the sex differences are related to the tests' *g* loadings, we can with Rushton (see Chapter 9 in this volume) talk about the demonstration of a "Jensen Effect", a shorthand phrase that saves the many words needed to describe the monotonous calculation of relating *g* to test differences.

Unfortunately, there is an important caveat to the use of the correlated vector method in the search for a sex difference in *g*. Thus, the number of tests — and not the number of subjects — determines the degrees of freedom in the correlated matrix analysis. This means that the likelihood of committing a type 2 error is rather high with this method. After some testing I have therefore come to the conclusion that the use of the correlated vector approach, while perfectly suited for many other purposes, is counterproductive in the pursuit of a small sex difference in *g*.

The last mean for testing for a sex difference is also the best. Here sex differences are first expressed in terms of point-biserial correlations (r_{pbs}) between sex and scores on each of the various subtests. Jensen states (1998: 538–542) that: "The point-biserial correlation ... is simply a Pearson product-moment correlation that expresses the relationship between a metric variable (e.g. test scores) and a dichotomous variable (in this case sex, quantitized as male = 1, female = 0 . . .".) The formula for I_{pbs}, allows for corrections for inequalities in sample sizes and SDs. The I_{pbs} is then fitted into the correlation matrix along with the various subtests inter-correlations and subjected to factor analysis. A PC, PF, or a HFA SL orthogonal analysis will then reveal how heavily each factor dimension loads on sex, including the *g* dimension.

3.1.9. Confounding factors The last requirement for a proper study of a sex difference in *g* is the analytic capability to identify and control for further confounders, such as sex-age interaction. It is common knowledge that the developmental tempo of boys and girls differ considerably. To conclude that there is an absolute sex difference in *g* among, say, 12-year boys and girls without taking the developmental advantage of girls into account would be risky at best. Age is usually taken to mean actual or chronological age in studies of school children, but it would actually be more correct to compare the sexes on basis of their biological rather than chronological age in developmental studies of children.

3.2. Summary of Analytic Considerations and Outline for the Optimal Study

We are now able to condense the analytic considerations and outline the optimal approach in a search for a sex difference in *g*. A study can be trusted only if it is based on representative samples of males and females, if it incorporates a large number of tests (i.e. ≥9), if there is no over-sampling of a particular type of ability in the test battery, and if data are subjected to a HFA analysis. The inclusion of the Schmid-Leiman (1957) orthogonal transformation is obligatory, because a small sex difference in *g* would otherwise too easily drown in contamination, either from first order group factors, some of which clearly favor females and others favor males, or from test specificity. The correlated vector analysis admittedly indicates the *g*-load of a sex difference, a so-called Jensen Effect, but the degrees of freedom are restricted to the number of tests, which

makes it a too conservative estimate of a sex difference in *g*. The best solution is, as stated by Jensen (1998) to include the point-biserial correlations in the factor matrix along with the inter-test correlations, to inspect the loading of sex on *g*, and to test the factored correlations coefficient for significance.

4. Selective Review of Sex Differences Research

4.1. The Development of a Quality Questionnaire

The question of whether a sex difference in general ability *g* exists after proper methodological control is, as previously mentioned, technically quite demanding. In an attempt to keep matters as simple as possible, this section reviews only a few studies. Rather than aiming for an exhaustive overview, the examination is meant to illustrate a number of specific methodological points, and therefore includes studies that vary sufficiently in quality to illustrate just that. The simple point scale described in table 10.1 is then used to grade the studies in accordance with how well they conform to the criteria for a scientifically sound study of sex differences in general ability *g*.

Use of the quality questionnaire is straightforward. One point is given to a study if the sample is fairly representative, rather than restricted to, say, university students. As previously mentioned, correct sampling is particularly important in sex differences research, because differences in male and female distributions can greatly influence the size and direction of the sex difference, depending on the set point of the ability scale. This point deserves repetition because too many studies use biased samples and no controls.

The derivation of a good *g* depends on a sufficient number of tests. The minimum number of tests for a sound hierarchical factor solution is 9 (Jensen (1998: 85). Studies including nine or more tests are awarded one point.

A study is granted one point for diversity if it includes a wide variety of tests mapping obviously different abilities. A way to check diversity is to see if the factor analytic solution allows for the derivation of at least three different first order factors.

The application of a HFA analysis is awarded one point, because it allows for close to virtual independence among factor dimensions. Studies deriving *g* by non-hierarchical PC or PF analysis earn no point, because this *g* is too easily contaminated by influences from the other factor dimensions when a sex difference is suspected.

Studies adding a Schmid-Leiman (1958) orthogonal rotation to the hierarchical solution are awarded a further point, as this transformation secures correlational independence among all factors, in addition to allowing for easier interpretation of the factor structure.

Studies just summing subtest scores are not given points, as they reflect unspecific "intelligence in general" rather than general intelligence. Structural equation modeling studies are not awarded points either, because the *g* thus derived may be contaminated by non-*g* factors (Bollen 1989). In this connection is it interesting to note that Gustavsson (1992) was unable to support the widely accepted Spearman hypothesis that

the white-black difference in *g* increases the more *g*-loaded a test one uses (e.g. Nyborg & Jensen 2001). Gustavsson analyzed the same WISC-R data used in the Jensen & Reynolds (1983) study, but he used a LISREL program to factor analyze it. The explanation for the divergence in results may be that LISREL and other structural equation solutions may produce a contaminated *g* that gives unexpected results.

One point is given if point-biserial correlations are calculated, included in the inter-test correlation matrix before factor analysis and, finally, tested for significance after factoring.

The use of a *t*-test or other straightforward statistical procedure to test for significance of sex differences in *g* earns no point. Neither does a correlated vector analysis. It will be remembered that the purpose of a correlated vector calculation is to inspect whether observed sex differences in various tests, often expressed in terms of d effects, correlate with the *g*-loading of the various tests, routinely after control for attenuation (a Jensen Effect). The problem is that a correlated vector calculation is a severe test that provides a highly conservative estimate of *d*–*g* relationships, because the degrees of freedom are restricted to the number of tests. This means that the relationship does not attain significance unless the sex difference is large, which seems unlikely (see later). In other words, the risk of committing a type 2 error of not correctly identifying a real sex difference, is unacceptably large, using this test. Obviously, proper use of the correlated vector method further presumes the availability of an uncontaminated *g* measure.

The scale thus allows for a total of 5 points in the case of a hierarchical analysis, and six points when the Schmid-Leiman transformation is added. Conversely, any study given less than five points provides an unacceptable shaky basis for conclusions about the existence of a sex difference in general intelligence.

4.2. Ranking of Studies

4.2.1. The Colom and García-López (2002) study. This study — "Sex differences in fluid intelligence among high-school graduates" — was specifically designed to take a stand in the controversy over whether there is a sex difference in general intelligence. Where Lynn (1994, 1999) argued that there is a difference, Colom and *García-López* explicitly intended to demonstrate that this is not true. They did so by subjecting 301 females and 303 males to two tests: Cattell's Culture-Fair (CF) Intelligence test (scale 3) and Raven's Advanced Progressive Matrices (APM). They further subjected 1,471 females and 1,997 males to the PMA Inductive Reasoning (IR) Test from the Primary Mental Abilities Battery. The idea behind using these tests was that they basically tap reasoning *g* ability (G_f), precisely the general intelligence proxy Lynn claims males are better at than females.

Colom and García-López made three observations: No sex differences on the CF; a significant female advantage on the IR ($p = 0.000$); and a male advantage on the APM ($p = 0.000$). Their verdict: "Given that there is no systematic difference favoring any sex in the measures of G_f, and that there is no sex difference in the best available measure of G_f (the Culture-Fair Test), it is concluded that the sex difference in fluid intelligence is non-existent". They also concluded, that "The data reported in this study disconfirm

the case set out by Lynn (1994, 1999) . . .", and further that they are ". . . contrary to the [results] reported by Allik *et al.* (1999). The main conclusion is that Lynn's notion of a sex difference in general intelligence is falsified.

How much confidence may we ascribe to the strong conclusions from this critical study? Not much according to the quality criteria in Table 10.1. First, subject sampling is biased (for details, see the comments to the Colom *et al.* (2000) and the Allik *et al.* (1999) studies). Second, the number of tests is too small to satisfy the minimum criteria. Third, the test battery does not satisfy the minimum diversity criteria. Fourth, intra-test scores are simply summed. Fifth, the more or less undetermined sex differences are averaged across tests. Colom and García-López thus mention that their measures of "Fluid intelligence (G_f) is usually seen as the core of intelligence behavior . . ." and they refer in this matter to Carroll (1993). However, Carroll (see Chapter 1 in this volume) has arrived at the conclusion that G_f largely dissolves when *g* is controlled for. Given

Table 10.1: Rough and ready grading scale for *g*-sex studies. Studies granted 5 points or less run an unacceptable high risk of committing either a type 1 or a type 2 error, that is, they permit no firm conclusion about the existence of a sex difference in general intelligence. Maximum is 6 points.

Qualifiers	**No = 0 Yes = 1**
Sample:	
Representative populations.	
Tests:	
Large number of tests (≥ 9).	
Diversity of tests	
Analytic Method:	
Hierarchical factor analysis (HFA).	
Orthogonal Schmid-Leiman transformation.	
(No points for:	
(1) Simple summing over standardized scores (reflecting "intelligence in general".)	
(2) Structural equation modeling (as it may give a *g* contaminated by non-*g* factors (Bollen, 1989)).	
Test:	
Inserting point-biserial correlations into the inter-test correlation matrix, co-factoring it, and then testing whether sex loads significantly on *g*.	
(No points for:	
Correlated vector analysis because it is too easy to make a type 2 error (i.e. not seeing a true difference)).	

this is correct, there is little reason to discuss G_f separately as a measure of general intelligence. Finally, by averaging summed scores over tests that probably differ in their *g*-loading, Colom and García-López end up with a less than optimal *g*-measure. In fact, the study earns no point.

4.2.2. Raven matrices. Court (1983) reviewed the entire literature (117 studies from five continents) on sex differences on Raven's Standard Progressive Matrices test for adults and the Colored Progressive Matrices test for children — "Sex differences in performance on Raven's Progressive Matrices: A review". The Raven tests may be interesting in the present context to the extent they are good proxies for pure *g*, and their *g*-loading indeed amounts to about 0.80 (Jensen 1998: 38). Court concluded that there is no consistent evidence for a sex difference in the Raven tests, and that these studies represent all degrees of representativeness.

According to the quality scale, the analysis is tentatively awarded one point for sample representativity, even though it is difficult to know for sure if some bias sneaked into this huge compilation of rather different studies, rather than just cancelled out. The Raven study neither qualified for points by representing many and very different tests, nor did it earn points for the total score based on summing.

The single point earned leads to the conclusion that the many studies using the Raven tests do not permit any sober conclusions about the existence of a sex difference in general intelligence.

4.2.3. The Colum *et al.* (2000) (1) study. This study — "Negligible sex differences in general intelligence" — subjected two samples, totaling no less than 10,475 adult subjects (6,219 males and 4,256 females, average age 23.12 years), to two cognitive test batteries, one with five and the other with six different tests. Raw data were factor analyzed and several tests for differences were applied in order to see whether the sexes differed in general ability, after having secured that the male and female factor structures were identical, as reflected in sufficiently high factor congruence coefficients. The main conclusion of the study was that there are only negligible sex differences in *g*.

Sampling bias is a problem in this study, as it is in most other studies. It is true that the majority of adult applicants for a private university may not actually pass the score level required for admission to a state university, but there are good reasons to believe that even these applicants do not represent an unselected population sample. Any bias toward the right side of the ability distribution will in general deflate the *g* loadings obtained, relative to samples better representing the general population, in addition to enhancing the probability of finding a male advantage to the extent their distribution is wider than that of the females. The possibility of ". . . some statistical sampling error . . ." is appropriately discussed by Colom *et al.* (p. 60), and related to the fact that there was close to 30% more males than females in the sample and to the possibility of some female self-selection. Thus, no point is given for proper sampling.

The two studies offered data from five or six different tests for analysis, respectively. This obviously reduces the likelihood of tapping into a large variety of abilities and in this way counter a possible ability bias. The study gets no points for using the first un-

rotated PF approach. A hierarchical analysis would have been a better choice, but this would require access to data from a minimum of nine tests from which at least three primary factors can be derived. As the study stands, the PF1 *g* measure most likely was contaminated to some undetermined degree by influences from the other factor dimensions (see below).

With respect to testing, it is interesting to note that when the sex differences in *g* are measured by Pearson's rs, *Rhos*, and *Taus*, they all turned out to be highly significant ($p < 0.000$) in both the first and second study. Colom *et al.* (p. 65) dismissed the significance of this finding as they ". . . could be explained by the non-g variance included in the *g* factor scores. It should be remembered that the *g* factor scores are not a pure measure of *g*". While the latter certainly is true, the size and consistency of the observed sex difference in *g* were considerable. The study earns no point for applying the method of correlated vectors, which by the way, did not come out statistically significant (Spearman $r = 0.000$; $p = 0.999$, after proper disattenuation). It is worth bearing in mind that it is highly unlikely that a correlated vector calculation based on only five or six tests (or even on the pooled 11 tests) will come out significant in sex difference research. The study earns one point for factoring in the point-biserial correlations. Curiously enough, the resulting loading of sex on *g* of 0.216 (found in Colom *et al.* 2002: 34) was never tested for significance. Given $N = 10{,}475$, I find the correlation to be highly significant ($p = 0.000$, Fisher $z = 0.219457$.) Had they done this testing, the authors obviously would have been forced to conclude that they had found a very real sex difference in *g* in male favor. On the other side, it should be realized that the derivation of *g* was based on the PF approach, which means that they probably operated with a contaminated *g*. That by itself renders any conclusion suspect.

The Colum *et al.* (2000) study of sex differences in *g* earns, in other words, a total of one point on the quality scale. This disqualifies it as a sound basis for deciding in the matter of a sex difference in general intelligence *g*. The methodological shortcomings, the divergence of data, and the observation of a significant sex load on *g*, should perhaps have tempered their main conclusion, of ". . . no sex difference in general intelligence" (p. 66).

4.2.4. The Jensen and Reynolds (1983) study. Jensen and Reynolds found a small but significant sex difference (M–F) of $d = 0.161$ ($p < 0.01$) in *g* factor scores in a study — "Sex differences on the WISC-R".

The study earns one point for drawing upon the national standardization sample of 6- to 16-year-old boys and girls, making it representative, at least for that age range. The study earns another point as data emanated from a fairly large number of subtests. However, as mentioned earlier, Wechsler purposely removed all test items during the construction of the test reflecting a large sex difference, carefully balanced out the remaining items so that what females would gain on verbal score side, males would gain on performance score, in order to present a neutral IQ test with no offensive overall sex difference. A test deliberately twisted that way cannot be trusted as an objective measure of sex differences. It may not represent a realistic distribution of abilities out there. No point could be given for the use of PF analysis. The study thus earned a total of 2 points on the quality scale.

Jensen (1998: 538) later carefully de-emphasized the value of the study, and the reasons he gave are illustrative. Basically, he argued, the precise size of the sex difference could not be estimated, because PF *g* factor scores might have been "... somewhat contaminated by small bits of the other factors and test specificity measured by the various subtests. This might either have increased or decreased the mean difference". I agree that the methodological flaws make it unlikely that this study can form a solid basis for the conclusion that there is no genuine sex difference in general ability *g*.

4.2.5. The Lynn *et al.* (2002) study. This study — "Sex differences in general knowledge" — applied a newly developed general information test (Irwing *et al.* 2001), covering 19 domains of general knowledge. The sample consisted of 469 female and 167 males. A second-order general factor was extracted that arguably reflects *g*, in addition to six first-order factors. Significant male superiority was found on the general factor and on four first-order factors, whereas females came out superior on one first-order factor. There were no sex differences in the remaining first-order factors.

Judging the methodological merits of this study, sampling cannot be claimed to be unbiased; the study drew on an unequal number of male and female undergraduates from three different academic areas. It gets one point for applying a large number of tests, but no points for diversity as general knowledge or information is but one, though heavily *g*-loaded, component of a broad-spectrum test battery. The study obtains one point for applying a hierarchical factor solution, but none for involving rather sophisticated MIMIC (multiple indicators and multiple causes models (Jöreskog & Sörbom 1993). Here, sex appears as the single predictor, the second-order general knowledge factor as a latent variable, and the first-order factors and domain levels of general knowledge as multiple indicators. Various models were tested using LISREL maximum likelihood estimation. A model with six effects on general knowledge domains was finally accepted. The coefficient for an effect of sex on overall general knowledge factor amounted to –0.42 (or 0.51*d*), but females performed better on factors reflecting Family and Fashion. The problem with structural equation modeling is, as mentioned previously that it may produce a contaminated *g* (Bollen 1989). The study concluded that males have more general knowledge than females. However, the earned total of two points on the quality scale makes it likely that we cannot learn much about sex differences in *g* from this study.

4.2.6. The Allik *et al.* (1999) study. This study — "Sex differences in general intelligence in high school students: Some results from Estonia" — found a substantial sex difference in *g* when testing 1,201 applicants for entry to the University of Tartu, Estonia. Raw data from four tests — verbal, reasoning, spatial abilities and scholastic knowledge — were factored, and a *g* was derived by the PC1 method. A large male effect size lead on *g* of $d = 0.65$, equal to 9.75 IQ points was found.

The sample consisted of applicants striving to enter a Social Science Faculty at the university. Obviously, this means that the sample was no more representative than the ones used in the Colum *et al.* (2000) and the Colom & García-López (2002) studies.

Potential university students are in general located well to the right of the mean in the normal ability distribution. Moreover, there were more than double as many female as male applicants (838 vs. 363, respectively). The battery consisted of four tests. The study thus misses the optimum requirements for numbers and diversity, even if each of the available tests loads heavily on *g*. The study gets no points for applying the PC approach, as the PC1 *g* measure is unacceptably vulnerable to biasing influence from tests with a clear male advantage, such as reasoning and spatial ability. The study checked for a sex difference in *g* by using a simple test for significance.

In other words, this study is compromised by sample bias, by the incorporation of a few and not widely varied tests, by the likely contamination of the PC1 *g* measure from other factors, and by the use of a simple statistical test for sex differences. It accordingly did not earn any points on the quality scale, and the finding of a rather large sex difference in general ability *g* cannot be trusted.

4.2.7. The Jensen (1998) analysis. Jensen analyzed five test batteries for which data were available for large and representative samples that encompass the full range of abilities in the general population, and presented the results in Jensen (1998: 538–541). Sex differences on each subtest were represented by point-biserial correlations that were inserted into the matrix of subtest inter-correlations. The loadings on sex on each of the factors, including *g*, were then determined after factor analysis. I leave out many of the interesting details of the analysis, and go directly to Jensen's two main conclusions:

- The sex difference in psychometric *g* is either totally nonexistent or is of uncertain direction and of inconsequential magnitude;
- The generally observed sex difference in variability of test scores is attributable to factors other than *g*.

However, there are methodological problems with this study. One of these relates to a critique voiced by Mackintosh (1996: 567). Mackintosh argued, "... research on sex differences suggests that different batteries yield significantly different general factors". He accordingly concluded that for the analysis of sex differences in *g* "... little will be gained by further pursuit of the precise nature of general intelligence defined in this way". Lynn (1999) follows up with further detailed critique of the analysis, some of which resonates with Mackintosh's. Lynn thus also finds that "... The nature of *g* depends on the kind of tests in the battery from which it is extracted", and further that the nature of the test batteries favored females and males to unequal extents, respectively. Lynn concluded that it is incorrect to average such greatly disparate estimates of *g* and then concludes that there is no sex difference in general ability. For reasons stated before and below I agree with Mackintosh's and Lynn's conclusion that the *g* measure was probably flawed. However, I disagree with Lynn's argument that a "global IQ obtained by summing the subtests . . ." will suffice. This measure is flawed, too. By the way, it should be remembered that this critique of the use of more or less contaminated *g* measures applies in particular to sex difference research. In most other cases, a *g* derived by different factor analytic solutions will usually also behave as a good *g* (Jensen & Weng 1994).

Lynn states incorrectly (1999: 8–9) that Jensen applied the PC method in the analysis. The analysis was actually based on the PF approach. Not that this matters much in the present context, because PC and PF analyses suffer equally with respect to the nagging problem of factor contamination. As in a PC analysis, one encounters the obligatory problem that ". . . contamination is especially significant when one extracts *g* as the first factor (PF1) in a principal factor analysis" (Jensen 1998: 86). Closer examination of data from the General Aptitude Test Battery, one of the five tests subjected to analysis, illustrates the nature of this problem rather well. Females performed significantly better than males in this test. However, the *g* factor was clearly compromised by a psychometric sampling excess of psychomotor tests that typically favor females. Jensen is, of course, too experienced to not note the danger. When he removed the female biased tests and performed a follow-up factor analysis, the remaining cognitive variables then showed only negligible sex loadings on PF1.

The analysis earns quite a number of points on the quality scale. One point is given for unbiased sampling, one for the large number of tests involved, and one for great test variety. It gets no points for the PF approach, but one for inserting point-biserial correlations in the factor analysis. However, the study earns a total of four points. This disqualifies it, according to the quality scale, as a trustworthy basis for conclusions about sex differences in *g*.

4.2.8. The Aluja-Fabregat *et al.* (2000) study. The title of this study is: "Sex differences in general intelligence defined as *g* among young adolescents". The investigation involved two independent samples of 678 primary school children in the first, and 887 children in the second. The average age was about 13 in both groups of volunteers, and there were an almost equal number of girls and boys.

With respect to the analytic approach, the authors state: "It seldom makes a difference whether *g* is represented by the highest order factor in a HFA analysis or by the first unrotated principal factor in a principal factor analysis. These typically have a congruence coefficient of +0.99 or more (Jensen & Weng 1994). We have used the first unrotated principal factor solution to extract *g*" (p. 815). This would be acceptable in cases where *g* research is tracing large group differences, such as among races or between social levels, but not in the search for a sex difference in *g*. Here test and sample bias may become THE basic problems to be dealt with effectively in order not to draw conclusions on the basis of the contaminated *g*. The Aluja-Fabregat *et al.* study thus earns no point for the PF approach. The point-biserial correlations were actually entered into the inter-test correlation matrix. The sex loading on *g* was –0.194 and –0.150 in the first and second study, respectively.

The authors offer an interesting interpretation of this: ". . . the percentage of *g* variance due to sex differences is 0.817 (first sample) and 0.420 (second sample) . . ." and this suggests, "A negligible sex difference in general intelligence defined as *g* in young adolescents" (pp. 818–819). This surely is not the optimal basis for deciding in the matter. The correct way is to check whether the two coefficients are statistically significant, and then a rather surprising result surfaces. Given a total *N* of 678 in the first sample, I find that the loading of sex on *g* of –0.194 is highly significant in the first sample ($p = 0.000$, Fisher $z = 0.196490$). In the second sample with $N = 887$, the sex loading on *g* of –0.150 also is highly significant ($p = 0.000$, Fisher $z = 0.151140$). The inescapable conclusion seems to be, that there are now two independent confirmations of a very convincing sex difference in *g* in **female** favor!

It will be remembered that the study was explicitly designed to disprove Lynn's (1994, 1999) developmental hypothesis. It says that boys have in fact higher *g* than girls, but this will be camouflaged by the girl's earlier maturation, so young girls ought to score similarly or even higher than same age boys. The statistical outcome of the Aluja-Fabregat study actually supports the hypothesis it was designed to falsify. We, nevertheless, cannot trust this conclusion. The two PF1 *g* measures were probably contaminated to an undetermined degree by a test bias, pointing in a female direction in both independent samples. One piece of evidence for this is that the girls in both samples outperformed the boys in all but an attention test. Another rather puzzling observation is that the girls in both samples outperformed the boys not only on the Math but also on the Natural Science test. Boys in practically all other studies surpass girls in these areas. This raises the suspicion that either the tests were not of sufficient complexity, or the female samples were biased in an upward direction.

The Aluja-Fabregat *et al.* (2000) study is tentatively awarded a total of three points: One for sampling, and two for using many and varied tests. This falls short of the five points needed for a solid *g*-sex study with reduced risk of committing type 1 or type 2 errors.

4.2.9. The Colom *et al.* (2002) study. This study — "Null sex differences in general intelligence: Evidence from the WAIS-III" — examined 703 females and 666 males, aged 15–94, from the Spanish standardization of the WAIS-III test. A male advantage of 3.6 IQ points was found in "ability in general", which is not far away from the 3.8 male average IQ lead observed by Lynn (1994).

Colom and colleagues, nevertheless, came to a very different conclusion, resting upon two other main findings. First, the non-significant outcome of the ". . . method of correlated vectors contradicts the conclusion that could be derived from the simple summation of the standardized mean group difference (*d*). Because of the greater scientific adequacy of the method of correlated vectors to test the null hypothesis concerning sex differences in general intelligence defined as *g*, we can conclude that there is no sex difference in general intelligence" (p. 34). Second, the factor loading of sex on *g* ". . . suggests a null sex difference". From this they deduced that the Ankney-Rushton paradox (a larger male brain predicts a male IQ lead) is irrelevant ". . . because there is no sex difference in general intelligence" (p. 34).

One of the eminent features of the study is, that Colum *et al.* first calculated point-biserial correlations among subtests' scores and the sex variable and included them within the matrix of subtest inter-correlations. They then performed a hierarchical Schmid-Leiman type factor analysis of the full matrix, which presently is the most adequate way to check for a sex difference in *g*. What they found was that sex loads 0.159 on *g*. Unfortunately, they lost this vital information again by combining the load value with all other available sex loadings on *g* coming from different studies. A meager average sex load of 0.02 came out of this averaging of sex loadings on *g* — clearly not an impressive figure!

The best approach is to directly test the observed sex load of 0.159 for significance. I did this, and found that the male lead in g is highly significant ($N = 1{,}369$; $p < 0.000$; Fisher $z = 0.160361$). This is all the more remarkable as the Colom *et al.* study operates with an excellent and probably entirely uncontaminated Schmid-Leiman transformed g identified in a representative sample.

Thus, rather than showing null sex differences, the overall conclusion of this methodologically sober study is that males significantly excel females in general intelligence g. The study gets one point for a representative sample, two points for using many and varied tests, one for a hierarchical factor analysis, and one for the Schmid-Leiman transformation. The earned total of five points indicates that the outcome of this study requires serious attention. I will take the last point for significance testing!

4.2.10. The Nyborg (2003) study. A final study — "Sex related differences in general intelligence, g, and group factors: A representative hierarchical orthogonal Schmid-Leiman type factor analysis" — found no sex difference in g before age 14, but identified a significant sex difference in the adult group of 52 males and females (as reported in Nyborg, 2001).

This is arguably the most carefully sampled study of those reviewed so far (Nyborg, unpublished data). The selection procedure began with a computer search in the late 1970s in the Danish Folkeregister for every twentieth child that was either 8, 10, 12, 14 or 16 years old, either a boy or a girl, and attending a school either in the countryside, in a suburb or in a larger city. Information about the socio-economic status of the parents, defined by father's occupational status, was also collected and categorized at five levels. If the twentieth child, or the parents, refused participation in the scheduled 20 years cohort-sequential study, the twenty-first (or in two cases the twenty-third) child on the computer list was invited. No particular pattern of reasons to refuse participation could be spotted in retrospect. Five preliminary age categories were established on basis of the results from this preliminary search protocol. The groups consisted of 8, 10, 12, 14 and 16+ year olds, respectively. When about 50% of the children were tested and filed, the distribution of all socio-economic and personal characteristics of the children were inspected for each group. The categories were then filled up with additional children, so that each age category finally mustered a total of 15 boys and 15 girls. During the fill-up process, great care was taken to ensure that each and all categories ended up being representative with respect to the general Danish socio-economic population distribution while also conforming to the nationwide proportional representation of rural, suburban and city residency. Data on children participating in the cross-sectional parts of the study were included in the present analysis, as were data on children participating in the longitudinal part of the study, but who had been examined only once. The particular selection procedure resulted in a total of 376 children and adults, with an identical number of girls and boys in each category.

All subjects were exposed to a large and varied battery of 20 or 21 ability tests (20 for the pooled 8 to 14 year-old-group, and 21 for the 16+ year-old group, with one subtest, Coding, making up the difference). The substantial number of highly varied tests permitted application of a hierarchical oblique factor analysis, which was supplemented with the Schmid-Leiman transformation. The factor structure coefficients for boys and girls were close to unity. A second order factor g and seven first-order factors were derived. Point-biserial correlations were computed, fitted into the inter-test correlation matrices, and factored in order to inspect the loading of sex on g, and tested for significance. The study also included a correlated vector analysis for Jensen effect, in addition to a traditional d effect analysis.

Point of departure for the analysis was a test of Lynn's prediction of a moderate but significant sex difference in g. The prediction could not be supported for the pooled 8 to 14-year-old children's sample, but the results of the adult sample actually confirmed the hypothesis. Thus, the point-biserial loading of sex on g was thus only 0.009 in children (ns), but reached 0.272 in the adult hierarchical orthogonal g factor analysis, which is significant (one-tailed $p = 0.026$) despite the very low $N = 52$. A correlated vector calculation reached significance neither for the children nor for the adult group, also as expected. Children's average sex difference d effect size was 0.18 or 2.62 IQ points, and the corresponding adult values were 0.26 or 3.94 IQ points, with positive signs indicating a male advantage in intelligence in general. The adult raw sex difference in g was 0.37 SD or 5.55 IQ points.

The study earns one point for being representative, and two for operating with a large battery of highly varied tests that allows for an adequate operational definition of g. It earns three points for factoring in the point-biserial correlations, for taking the hierarchical factor approach, for optimal orthogonalization through the Schmid-Leiman transformation, and for testing the loadings for significance. In other words, all likely precautions were taken against the likelihood of g-contamination in this study, due to the carefully chosen sample, the particular analytic approach, and the presence of a rich, varied and highly g-loaded test battery. The maximum of six points earned means that we can ascribe at least the same degree of confidence to the conclusions of this study as we did to the Colom *et al.* (2002) study.

This concludes the selective review of studies. Studies earning less than five points on the quality scale may either find a female advantage, a male advantage, or no sex difference in g, but none of these can be trusted due to the risk of contamination. Only two recent studies obtain five or six points, and both studies identify a significant adult advantage in g.

5. Discussion

This chapter specifically addressed the problem why sex difference research on g has been plagued for so long by analytic inconsistency and incompatible findings, and thus has provided little guidance for a scientifically based opinion whether there is in fact a sex difference in general intelligence which could explain, at least in part, the obvious sex-differentiated achievement in education, jobs, and societal power structures, as well as the repeated observation of an average male advantage in brain sizes.

The strategy chosen was to take point of departure in the analytic and empirical disagreement among two of the most prominent combatants in the field. On the one side, there is Lynn (1994, 1999) who uses the sum standardized scores, and finds an average significant male superiority in general intelligence, and uses the on average larger male

brain to explain this difference. Jensen (1998) is, on the other side, moved neither by Lynn's finding of a sex difference in summed scores nor by the interpretation of its basis. Using the more advanced factor analytic inter-test correlation approaches, Jensen documents considerable inconsistency in the data: sometimes females obtain a higher score, sometimes males are in front, and sometimes there is no sex difference at all in *g*. In short, there is no reliable sex difference in *g*.

The way the present chapter addressed this analytic and empirical dilemma was, basically, to systematize Jensen's analytic critique in terms of the development of a brief catalogue of criteria for a proper study of sex differences in *g*. Then a number of typical studies were judged against these criteria, with the hope that this strategy promised a double advantage: to examine in detail whether Lynn's claimed male *g* advantage is a fact, and whether it would survive even at the highest levels of a Jensenist hierarchy of increasingly more demanding methodological environments.

5.1. Three Major Conclusions

The grading of *g*-studies allowed for three major conclusions.

First, studies granted less than five points on the quality scale routinely produce unreliable or inconsistent results, which means that their conclusions about the existence of a sex difference in *g* cannot be trusted. Most of these studies do not sample properly. Many studies remain satisfied with the summing of standardized scores, which makes them particularly vulnerable to the possibility of arriving at a contaminated IQ or "intelligence in general". To make things worse for sex differences research, the PC or PF analytic approaches cannot be trusted either, because the "general intelligence" or *g* thus derived may take on color to an undetermined degree from the non-*g* factors in the matrix. Jensen (1998: 539–540) elegantly demonstrated this danger in the previously discussed analysis of the GATB test battery. This test contains an unusual number of psychomotor tests for vocational aptitudes that favor females. It will be remembered that the observed female advantage in *g* disappeared after proper control for this test bias. The obvious implication is, that studies based on PC or PF analysis can neither be taken to confirm nor reject the possibility of a sex difference in *g*. We have to apply methodologically more stringent studies that take stronger precautions against test bias and the associated contamination of *g* by group or test specific factors, before they deserve our confidence in conclusions about a sex difference in *g*.

The second conclusion is that, provided the requirements set forth in the quality questionnaire in Table 10.1 are met, a significant adult sex difference in *g* appears. The Colom *et al.* (2002) and the Nyborg (2001, 2003) studies are the only ones that offer a hierarchical Schmid-Leiman transformation solution, in addition to conforming to other stated quality criteria. In both those cases, a moderate male lead in *g* is identified, as the factored point-biserial *g*-loading of sex on *g* came out unexpectedly significant at the two-tailed $p < 0.000$ level in the Colom case and, despite a critically low *N* of 52 in the Nyborg case, at a predicted one-tailed $p = 0.026$ level. The loading of sex on *g* did not reach significance in the 8 to 14-year-old child sample in the latter case.

The third conclusion is that the moderate male *g* advantage of 0.37 SD probably goes a long way in explaining why it was so difficult to pin it down in a multitude of inconsistent studies. A difference of that size easily drowns in studies not following the most stringent methodological rules, and where all sorts of influences from group factors and test specificity may contaminate the *g*. The cure against this danger seems to be to sample carefully, to use many and highly varied tests, and to exploit the mathematical approach behind the Schmid-Leiman orthogonalized transformation of the hierarchically organized factor dimensions. Such strict requirements for measuring *g* probably do not reach the same importance in studies of the well-documented much larger race or social level differences in *g*.

Obviously, the final establishment of an adult sex difference in *g* needs more than two replications!

5.2. Theoretical Implications

There are at least two theoretical lines of interests in knowing whether the sexes differ in *g*.

First, Jensen (1998: 541) states that no difference in *g* means that there is no sex difference in the "... general conditions of the brain's information-processing capacity that cause positive correlations among all of the modular functions on which there is normal variation and which account for the existence of *g*". It furthermore means that, "... the true sex differences reside in the modular aspects of brain functioning." In other words, the finding of a real sex difference in *g* would force us to acknowledge that whatever causes the positive manifold among abilities, observed by Spearman (1904), would be subjected to an effect of a general and not just specific nature.

Given that there is a small but real sex difference in *g*, and further given that the difference does not seem to appear before puberty (as suggested by the Nyborg 2001, 2003), one might speculate that some sex-related brain differentiation is taking place among boys and girls around that time. This could involve the general conditions of the brain's information-processing capacity, or it could provoke modular differentiation, or both. In either case, it is likely that individual and group differences in gonadal or adrenal hormones at puberty might be involved, because such hormones are known to significantly affect brain development (Nyborg 1994a). A study by Nyborg & Jensen (2000) suggested that only extremely high or low plasma testosterone concentrations significantly affect adult *g*, and then in a downward direction. However, it is worth keeping in mind that this sample consisted of middle-aged men with fully mature brains. The study might not adequately reflect the complex *g*-hormone concentration connections in much more sensitive younger people. In any case, the question whether sex hormones affect general or only modular aspects of the brain's information-processing capacity in a sex-related way can be fully answered only in carefully designed longitudinal studies. No such study is presently available, but one has been on its way since 1976 (Nyborg, unpublished data). The Nyborg (2001, 2003) analysis, referred to previously, took its data from this much more comprehensive study.

The second theoretical implication of a male advantage in *g* has to do with the Ankney-Lynn-Rushton brain size-IQ paradox mentioned previously. It will be remembered that Lynn (1994) was able to predict rather accurately a male lead in IQ from knowledge of a male lead in brain size. His calculations are presented in Table 10.2.

Given a male lead in brain size of SD = 0.78, and provided that the mean correlation between brain image size and IQ is 0.35 (Rushton & Ankney 1996), all we have to do is to multiply the male lead in brain size with the brain size-IQ correlation. We then get an SD of 0.27. When multiplied by 15 this 0.27 SD value translates into a male IQ lead of about 4 points. This theoretical prediction of intelligence from brain size matches the observed male average IQ lead of 3.8 quite well, but is for obvious reasons restricted to "intelligence in general". The Nyborg (2001, 2003) study allows us to re-test the prediction, but this time using an uncontaminated measure of *g* or "general intelligence". Like in the Lynn case, the arithmetics is straightforward. The observed sex difference in head circumference, a proxy for brain size, was SD = 0.87. This gives, when multiplied by the observed head circumference — *g* correlation of 0.34, an SD = 0.30, which, when multiplied by 15 turns into a predicted male IQ advantage of 4.437. The observed male lead in *g* was 0.37 which, when multiplied by 15, corresponds to a male IQ advantage of 5.55 points. In other words, Lynn underestimated the sex difference in *g* by 0.44 IQ point when using the questionable "intelligence in general" IQ measure.

5.3. The Very High End Male g Hypothesis

Obviously, the importance of the observed sex difference in *g* is not to be found in the group mean. No sensible prediction can be made for any individual male or female by referring to a mean average difference of just 0.37 SD. However, a brief consultation of the characteristics of Gaussian distribution theory teaches us that even a moderate mean advantage in *g* will have a considerable effect on the male/female ratio of individuals with high or very high *g*. In fact, the higher scoring group will be exponentially over-represented above a given high cut-off point on the scale. The growing disparity is graphically illustrated in Figure 10.1.

Equally obviously, a larger male variability would enhance this pattern, whereas the larger number of surviving females in each age category throughout life would to some small extent counter it. This means that the idealized curves must be inspected with caution, and this applies in particular at the extremes, as it is not given that they follow the normal distribution here. Finally, the sex difference in *g* variability in the Nyborg (2001, 2003) studies is larger than what is typically seen in most studies of IQ. It remains to be seen whether a sex difference of that size in *g* variability materializes in future *g* studies, or is just an artifact of the present study.

With these provisos in mind, we can begin to speculate about the practical predictive validity in the real world, outside the test room, of the male mean *g* advantage at the right side of the distribution. But first we have to realize that real life situations constitute a much, much broader test basis than the sex difference reported in this

Table 10.2: Predictions of **IQ** (intelligence in general) and *g* (general intelligence) from a sex difference in brain volume or head circumference, a proxy for brain size (from Nyborg (2002).

Study	**A: Observed sex differences in brain size**	**B: Correlations between volume[a] or circumference[b] and IQ[1] or g^2**	**C: A × B**	**D: Predicted male lead in**	**E: Observed male lead in**
	***d* effects**	***r***	**SD units**	**IQ[3] or g^4 C × 15**	**IQ[5] or g^6**
Lynn (1999): IQ data averaged over several studies	0.78	0.35[a1]	0.27	4.05[3]	3.85[5]
Nyborg: Observed data in a specific study of *g*	0.87	0.34[b2]	0.30	0.30[4] *(IQ 4.44)*	0.37[6] *(IQ 5.55)*

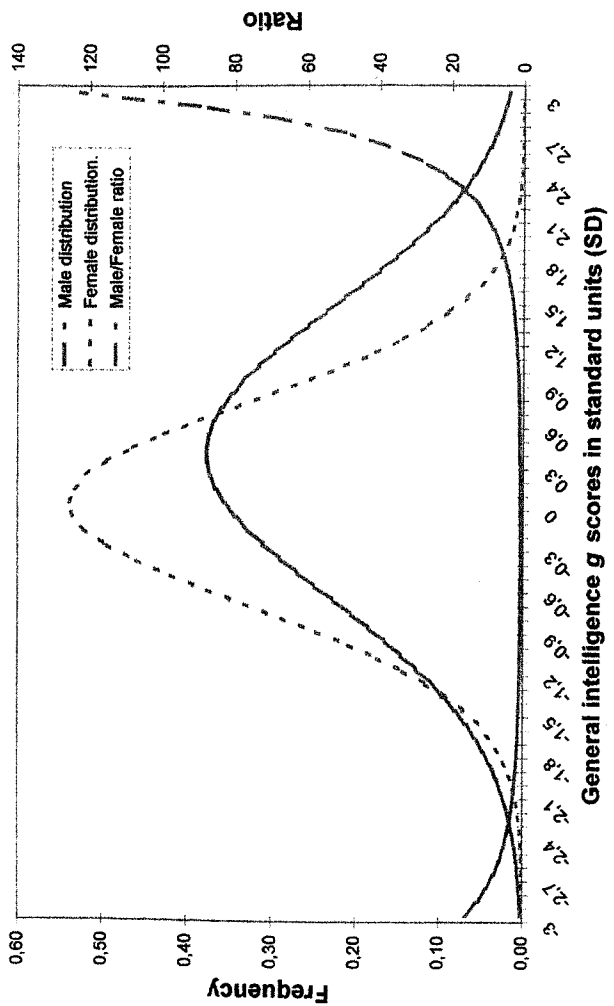

Figure 10.1: Male and female distributions and ratio as a function of male $g = 0.36$ (SD 1.06) and female $g = -0.01$ (SD 0.74). The theoretical ratio of males to females with $g = 3$ (IQ = 145) is about 120 : 1. 2.15% of the population obtains a g score ≥ 2 SDs, and only 0.13% a g score ≥ 3 SDs (from Nyborg 2002).

chapter, even if they emanated on the basis of relatively wide-ranging batteries of standardized tests in the two confirmatory studies. It has thus been argued that, in the largest possible scale, individuals coping with recurrent complex life problems can be viewed as participating in a gigantic longitudinal intelligence test (Gordon 1997; Gottfredson 1997, and Chapter 15 in this volume). In this all-encompassing perspective, all people can be seen as examinees in a gigantic and extremely varied set of tests, and many of these everyday tests undoubtedly are highly g-loaded. Given the male average lead in g, we can expect that more males will come out with a higher everyday success score, and more so the heavier the g-load of the every-day tests. For simplicity I will call this the "Very high end male g hypothesis". There actually is a growing body of evidence to support this hypothesis. Males typically outperform females in most top level educational, vocational, and political power areas — with a disproportionately higher proportion of males found as complexity and demands increase. This tendency is seen most clearly in complex problem-solving mathematics (Benbow 1992), engineering and physics (Lubinski *et al.* 2000), and in other areas calling for high spatial ability (Shea *et al.* (in press)). The "very high end male g hypothesis" might also provide part of the explanation for the massive male preponderance in high-level chess competition, musical composition, theoretical physics, economy and in the numerous other areas of demonstrated high-level male dominance.

This said, it is vitally important to realize that there is no question that a multitude of factors besides the sex difference in g will also have to be fitted into a realistic equation for predicting unequal participation in challenging areas of life (e.g. Nyborg, in press; Nyborg & Jensen 2001).

5.4. *The Future of Sex Differences Research on g*

Mainly thanks to Arthur Jensen we can now safely assume that both the definition and measurement of general intelligence are on safe ground and need little further attention. What needs closer attention, though, is the definition and measurement of sex. The division of mankind into male and female categories is awfully crude. More sophisticated approaches are possible but, unfortunately, vicious politically correct attempts to gloss over decisive biological differences among male and female are sure to surface whenever the attention is turned toward biologically based modeling of sex. Proofs of this are the widespread preference for terms like gender (supposed to have an environmental basis) over sex (supposed to have a biological basis), or the more wide-angled claim (with little hard evidence) that sex is just a social construct, or the aggressive claim that those who cannot see this must house a hidden hostile political agenda against females. This is empirically irresponsible and promotes more heat than light.

A realistic biological approach to sex differences research involves two things: deep knowledge of the operational definitions of sex, and the development of fine-grained, multi-spectered and continuous ways to classify males and females.

With respect to operational definitions, sex can be analyzed at many levels: the level of chromosomal sex (XX or XY, or multiple combinations thereof), sex hormonal sex

(simplified here as estrogen or androgen in various ratios), internal sex organ sex, external sex organ sex, sexual inclination, gender identity, or sexually differentiated phenotypes. Usually, the sexual development at all these levels tends to co-vary, but any or all of the levels may co-vary differently in a given individual (Nyborg 1994a, 1994b). All causal modeling of sex will have to incorporate the interaction patterns among these levels, and only lazy social researchers may think they can get away with anything less.

It is an interesting hypothesis that, the causal factors that guide the coordinated sex-related development of the body and brain functions may also explain both the moderate sex-related difference in *g* and the much larger sex-related differences in group factors like verbal and spatial abilities. Sex hormones seem a good first choice here. Obviously, without sex hormones there will be no phenotypic sex-related differences at all. By implication, there will be no sex differences in body, brain, behavior and in *g*. A fetus, whether it is chromosomally male or female, will inevitably develop into a female in the absence of prenatal testosterone, *t*. *t* is a potent androgenic so-called male hormone — so-called because *t* can be aromatized into the so-called female hormone estradiol which, when present in sufficiently high concentrations, may exert masculinizing effects. Contrariwise, irrespective of male or female chromosomal complement, if the fetus is exposed to sufficiently high concentrations of *t*, it inevitably develops into a male with all the related bodily, brain and behavioral characteristics. One exception to this rule is, that if it is unable to induce *t* receptor molecules, the fetus will — despite it's male *t* level and even it's male XY karyotype — develop a female phenotype.

All this means that hormonally guided body and brain development is better considered a continuous than as a categorical phenomenon. It further means that, depending on the time-table for hormone exposure and transient individual differences in local body or brain receptor sensitivity, a given hormone exposure may during development provide an otherwise predominantly female individual with a couple of typical male traits, or an otherwise predominantly male individual with a couple of outstanding feminine traits, even if chromosomal and hormonal sex usually coincide in a species-specific evolutionary economic way (Nyborg 1994a).

With respect to the development of more fine-grained, multi-faceted and continuous ways to classify males and females, we can now begin to profit from an improved understanding of how sex hormones are capable of making multiple continuous mixings of male and female traits possible, on top of an individual's chromosomal sex. The causal basis is that hormones may alter the transcription of familial genes, whenever they are present in sufficient concentrations and specific receptor molecules can be induced in the target tissues. This mechanism allows us to classify males and females into more fine-grained categories, such as hormotypes (Nyborg 1984, 1994a, 1994b, 1997a, 1997b): I have tentatively established five different androtypes for XY males (low *t* A1s to high *t* A5s), and five estrotypes (low estradiol E1s to high estradiol E5s) for XX females (plus two A0 or A6, and two E0 or E6 categories for abnormally low or high hormone values — 1 or 99 percentiles, respectively).

The general Trait Co-variance — Androgen/Estrogen (GTC–A/E) model then generates a large number of empirically testable predictions about covariant body, brain, and behavioral development (see Nyborg 1994a, 1994b, 1997a, 1997b, for details).

With respect to estrotypes, young E1 individuals with relatively low early plasma estrogen and relatively high testosterone levels are thus expected to slowly develop high *g*, in addition to forming a slightly masculinized personality, and to expose few social interests. Such E1 females are expected to do well in male occupations like engineering, but they also encounter little reproductive success. They are thus expected to give birth to few and late children. A test in the U.S. of the GTC–A/E model confirmed that they have, in fact an unexpected high rate of unprovoked abortion (Hausman 1999). Clinical evidence suggests that E1 females, who for various medical or environmental reasons have been exposed prenatally to non-physiological amount of androgens, tend to score higher on academic achievement measures than normal females (Hoyenga & Hoyenga 1979, 1993). Another line of clinical evidence suggests that sex hormone treatment may affect the development of intelligence in young girls with Turner's syndrome (Turner 1938; Nyborg 1990; Nyborg & Nielsen 1981). These girls lack some X chromosome material and most remain psychosexually infantile, unless given proper hormone treatment. A pseudo-experimental study suggested that *t* treatment enhances *g*, whereas estradiol treatment elevates spatial abilities and depresses verbal abilities, and growth hormones do not affect intelligence (Nyborg *et al.* 2001; under revision).

With respect to males, we know that *g* is linearly and positively related to job status and income (Nyborg & Jensen 2001; see also Chapter 15 in this volume). Further analyses of a large sample of 4,000+ males suggest that plasma *t* is inversely related to these outcome variables, and that physical dominance and violence in interpersonal situations is related to high plasma *t*, whereas formal dominance, educational level and IQ is related to low *t* in A1 males (Nyborg, in press).

Obviously, this chapter is not the place for a more detailed elaboration of the host of modifying variables that come into play in the tangled webs of basically unexplored hormonally, genetically and environmentally based molecular, anatomical and functional interactions. The interested readers may consult Nyborg (1994a, 1994b, 1997a, 1997b) for suggestions on how to test the models for these complex relationships. Moreover, quite sophisticated analytical tools have recently been developed, so as to make it possible to not only directly take effects of hormones on intelligence and personality into account, but also to incorporate and test the power of the many indirect effects in such highly complex nexus (Netter *et al.* 2000; Reuter *et al.*, submitted.)

In conclusion we can — essentially thanks to Arthur R. Jensen — now remain satisfied with already available solid operational definitions and practical measures of general intelligence *g*. The next move in the study of sex-related differences in *g* is, accordingly, to enlighten its biological side. We already have some promising lines of evidence and models, telling us how to proceed in the future. The ultimate task is to unravel the material basis of physical *g* "... whereby physiology will achieve the greatest of all its triumphs" (Spearman 1927). Undoubtedly, an important part of this accomplishment consists of carefully modeling interaction effects of genes, hormones and environmental factors, with brain and body differentiation, and to inspect how all this affects specific and general life achievement measures in a coherent, consistent and causally satisfying way, including the possibility of experimental control.

With all this out in the open, the contemporary crude and dichotomous sex categorization analyses would surely abate to multi-faceted gene-hormone-body-brain-

behavior-environment covariant interaction analyses (Nyborg 1997a, 1997b). This would hopefully assist in illuminating the causes for why low *t* A1 males and high *t* E1 females tend to converge developmentally towards high *g*, an androgynous body type with minimal sexual differentiation, and similarities in personality, and why high *t* A5 males and high estradiol E5 females tend toward developing low *g*, early and pronounced sexual development by maximal differentiation between their male or female body characteristics.

I feel quite confident that the two grand old masters of general intelligence — Charles Spearman and Arthur Robert Jensen — would not be too unhappy with the current situation. We may now begin to combine the almost unlimited possibilities in modern biological and molecular sciences with behavioral genetics' increased precision in identifying environmental effects. By probing deep down into the physical foundation of human nature, it might be possible to finally understand the basically physical and chemical nature of *g*, and the mechanisms for how hormones may affect *g* in a sex-related way by modulating the expression of familiarly transmitted genes. The "very high end male *g*" hypothesis may help us to understand why a modest average sex difference in *g* may have such large implications at the highest steps of the educational, occupational and political power hierarchies. There may, in fact, be no scientifically acceptable alternative approaches.

References

Allik, J., Must, O., & Lynn, R. (1999). Sex differences in general intelligence in high school students: Some results from Estonia. *Personality and Individual Differences, 26*, 1137–1141.

Aluja-Fabregat, A., Colom, R., Abad, F., & Juan-Espinosa, M. (2000). Sex differences in general intelligence defined as *g* among young adolescents. *Personality and Individual Differences, 28* (4), 813–820.

Ankney, C. (1992). Sex differences in relative brain size: The mismeasure of woman, too? *Intelligence, 16*, 329–336.

Ankney, C. (1995). Sex differences in brain size and mental abilities: Comments on R. Lynn and D. Kimura. *Personality and Individual Differences, 18*, 423–424.

Benbow, C. P. (1992). Academic achievement in mathematics and science of students between ages 13 and 23: Are there differences among students in the top one% of mathematical ability? *Journal of Educational Psychology, 84*, 51–61.

Bollen, K. A. (1989). *Structural equation with latent variables*. New York: Wiley.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Brody, N. (in press). Construct validation of the Sternberg Triarchic Abilities Test (STAT): Comment and reanalysis. *Intelligence, 30*.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, U.K.: Cambridge University Press.

Colom, R., García, L. F., Juan-Espinosa, M., & Abad, F. (2002). Null sex differences in general intelligence: Evidence from the WAIS-III. *Spanish Journal of Psychology, 5* (1), 29–35.

Colom, R., & García-López, O. (2002). Sex difference in fluid intelligence among high school graduates. *Personality and Individual Differences, 32*, 445–451.

Colum, R., Juan-Espinosa, M., Abad, F., & García, L. (2000). Negligible sex differences in general intelligence. *Intelligence, 28* (1), 57–68.

Court, J. H. (1983). Sex differences in performance on Raven's Progressive Matrices: A review. *Alberta Journal of Educational Research, 29*, 54–74.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62*, 61–84.

Gardner, H. (1983). *Frames of mind*. New York: Basic Books.

Gardner, H. (1993). *Creating minds*. New York: Basic Books.

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam.

Gordon, R. A. (1997). Everyday life as an intelligence test: Effects of intelligence and intelligence context. *Intelligence, 24*, 203–320.

Gottfredson, L. (1997). Why *g* matters: The complexity of everyday life. *Intelligence*, 24, 79–132.

Gottfredson, L. Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 30* (in press).

Gould, S. J. (1996). *The mismeasure of man*. New York: Norton & Company.

Gustavsson, J-E. (1992). The "Spearman hypothesis" is false. *Multivariate Behavioral Research, 27*, 265–267.

Halpern, D. F., & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review, 12* (2), 229–246.

Hausman, P. (1999). On the rarity of mathematically and mechanically gifted females: A life history analysis. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 60* (6-B), 3006.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: The Free Press.

Hoyenga, K. B., & Hoyenga, K. T. (1979). *The question of sex differences. Psychological, cultural, and biological issues*. Boston, MA: Little, Brown and Company.

Hoyenga, K. B., & Hoyenga, K. T. (1993). *Gender-related differences*. Boston, MA: Allyn and Bacon.

Irwing, P., Cammock, T., & Lynn, R. (2001). Some evidence for the existence of a general factor of semantic memory and its components. *Personality and Individual Differences, 30*, 857–871.

Jensen, A. R. (1975). A theoretical note on sex-linkage and race differences in spatial ability. *Behavior Genetics, 5*, 151–164.

Jensen, A. R. (1978). Sex linkage and race differences in spatial ability: A reply. *Behavioral Genetics, 8*, 213–217.

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

Jensen, A. R. (1987). Psychometric *g* as a focus of concerted research effort. *Intelligence, 11*, 193–198.

Jensen, A. R. (1988). Sex differences in arithmetic computation and reasoning in prepubertal boys and girls. *Behavioral and Brain Sciences, 11*, 198–199.

Jensen, A. R. (1993). Psychometric *g* and achievement. In: B. R. Gifford (Ed.), *Policy perspectives on educational testing* (pp. 117–227). Boston: Kluwer Academic Publishers.

Jensen, A. R. (1994a). Phlogiston, animal magnetism, and intelligence. In: D. K. Detterman (Ed.), *Current topics in human intelligence* (Vol. 4), *Theories of intelligence* (pp. 257–284). Norwood, NJ: Ablex.

Jensen, A. R. (1994b). Psychometric g related to differences in head size. *Personality and Individual Differences, 17*, 597–606.

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

Jensen, A. R., & Johnson, F. W. (1994). Race and sex differences in head size and IQ. *Intelligence, 18*, 309–333.

Jensen, A. R., & Reinolds, C. R. (1983). Sex differences on the WISC-R. *Personality and Individual Differences, 4*, 223–226.
Jensen, A. R., & Weng, L-J. (1994). What is a good *g*? *Intelligence, 18*, 231–258.
Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.
Kline, P. (1991). Sternberg's components: Non-contingent concepts. *Personality and Individual Differences, 12*, 873–876.
Lubinski, D., Benbow, C. P., & Morelock. M. J. (2000). Gender differences in engineering and the physical sciences among the gifted: An inorganic-organic distinction. In: K. A. Heller, F. J. Monks, R. J. Sternberg, & R. F. Subotnik (Eds), *International handbook for research on giftedness and talent* (2nd ed., pp. 627–641). Oxford, U.K.: Pergamon Press.
Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences, 17*, 257–271.
Lynn, R. (1997). Sex differences in intelligence: Data from a Scottish standardization of the WAIS-R. *Personality and Individual Differences, 24* (2), 289–290.
Lynn, R. (1999). Sex difference in intelligence and brain size: A developmental theory. *Intelligence, 27* (1), 1–12.
Lynn, R. (2001). *The science of human diversity: A history of the Pioneer Fund*. New York: University Press of America.
Lynn, R., Irwing, P., & Cammock, T. (2002). Sex differences in general knowledge. *Intelligence, 30* (1), 27–39.
Maccoby, E. E. (1966). Sex differences in intellectual functioning. In: E. E. Maccoby (Ed.), *The development of sex differences*. Stanford, CA: Stanford University Press.
Mackintosh. N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science, 28*, 559–572.
Messick, S. (1992). Multiple intelligences or multilevel intelligence? Selective emphasis on distinctive properties of hierarchy: On Gardner's Frames of Mind and Sternberg's Beyond IQ in the context of theory and research on the structure of human abilities. *Psychological Inquiry, 3*, 365–384.
Neisser, U., Boodoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S., Halpern, D., Loehlin, J., Perloff, R., Sternberg, R., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.
Netter, P., Toll, C., Rohrmann, S., Hennig, J., & Nyborg, H. (2000). Configural frequency analysis of factors associated with testosterone levels in Vietnam veterans. *Psychologische Beiträge*, (Band 42), 504–514.
Nyborg, H. (1984). Performance and intelligence in hormonally-different groups. In: G. Vries, J. Bruin, H. Uylings, & M. Corner (Eds), *Sex differences in the brain. Progress in Brain Research* (pp. 491–508). Amsterdam: Elsevier Biomedical Press.
Nyborg, H. (1990). Sex hormones, brain development, and spatio-perceptual strategies in Turner's syndrome. In: D. Berch, & B. Bender (Eds), *Sex chromosome abnormalities and human behavior: Psychological studies*. Boulder, CO: Westview Press.
Nyborg, H. (1994a). *Hormones, sex, and society: The science of physicology*. Westport, CT: Praeger.
Nyborg, H. (1994b). The neuropsychology of sex-related differences in brain and specific abilities: Hormones, developmental dynamics, and new paradigm. In: P. A. Vernon (Ed.), *The neuropsychology of individual differences* (pp. 59–113). San Diego: Academic Press.
Nyborg, H. (1997a). Molecular man in a molecular world: Applied physicology. *Psyche & Logos, 18* (2), 457–474.
Nyborg, H. (1997b). Personality, psychology, and the molecular wave: Covariation of genes with hormones, experience, and traits. In: J. Bermudez, B. De Raad, A. M. Perez, A. Sanchez-Elvira, & G. L. van Heck (Eds), *Volume of Personality Psychology in Europe* (Chapter 16, pp. 159–173). Tilburg, The Netherlands: Tilburg University Press.

Nyborg, H. (2001). Early sex differences in general and specific intelligence: Pitting biological against chronological age. Paper presented at the Second Annual Conference for Intelligence Research (ISIR). Cleveland, OH, December 6–8th.
Nyborg, H. (2002). IQ and *g*: The art of uncovering the sex difference in general intelligence. Paper presented at the Third Annual Conference for Intelligence Research (ISIR). Vanderbilt University, Nashville, TN, December 5–7th (p. 37)
Nyborg, H. (2003). Sex-related differences in general intelligence, g, and group factors: A representative hierarchical orthogonal Schmid-Leiman rotation type factor analytic study (submitted).
Nyborg, H. (in press). Multivariate modeling of testosterone-dominance associations. *Behavior and Brain Sciences*.
Nyborg, H. *The Jutland School Project: A 25 year cohort-sequential study* (unpublished data)
Nyborg, H., & Jensen, A. R. (2000). Black-white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences, 28*, 593–599.
Nyborg, H., & Jensen, A. R. (2001). Occupation and income related to psychometric *g*. *Intelligence, 29* (1), 45–55.
Nyborg, H., & Nielsen, J. (1981). Sex hormone treatment and spatial ability in women with Turner's syndrome. In: W. Schmid, & J. Nielsen (Eds), *Human behavior and genetics* (pp. 167–182). Amsterdam: Elsevier/North-Holland Biomedical Press.
Nyborg, H., Nielsen, J., Naeraa, R., & Kastrup, K. (2001). *Estrogen and androgen treatment affect the development of general and specific intelligence differently in young girls with Turner's syndrome*. Paper presented at the ISIR meeting, Cleveland, Ohio, December 4–7th.
Nyborg, H., Nielsen, J., Naeraa, R., & Kastrup, K. (under revision). Estrogen and androgen, but not growth hormone, treatment affect the development of general and specific intelligence differently in young girls with Turner's syndrome.
Pakkenberg, B., & Gundersen, H. J. (1997). Neocortical neurone number in humans: Effects of age and sex. *Journal of Comparative Neurology, 384*, 312–320.
Reuter, M., Netter, P., Hennig, J., Mohiyeddini, C., & Nyborg, H. (2003). Test of Nyborg's General Trait Covariance (GTC) model for hormonally guided development by means of structural equation modeling. *European Journal of Personality* (in press).
Rushton, J. P. (1992). Cranial capacity related to sex, rank and race in a stratified sample of 6,325 military personnel. *Intelligence, 16*, 401–413.
Rushton, J. P., & Ankney C. D. (1996). Brain size and cognitive ability: Correlations with age, sex, social class, and race. *Psychonomic Bulletin and Review, 3*, 21–36.
Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61.
Segerstråle, U. (2000). *Defenders of the truth: The battle for science in the sociobiology debate and beyond*. Oxford: Oxford University Press.
Shea, D. L., Lubinski, D., & Benbow, C. P. (in press). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*.
Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201–293.
Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition*. London: Macmillan.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.
Spearman, C., & Jones, L. W. (1950). *Human ability: A continuation of The abilities of man.* London: Macmillan.
Sternberg, R. J. (1988). *The triarchic mind: A new theory of intelligence.* New York. Viking Press.
Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence? Contemporary viewpoints on its nature and definition.* Norwood, NH: Ablex.
The Editors (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12,* 123–147, 195–216, 271–275.
Turner, H. (1938). A syndrome of infantilism, congenital webbed neck, and cubitus valgus. *Endocrinology, 23,* 566–574.
Wolman, B. B. (Ed.) (1985). *Handbook of intelligence: Theories, measurements, and applications.* New York: Wiley.

Part IV

The *g* Nexus