J. biosoc. Sci. (1998) **30**, 533–539

REPLY TO LYNN

N. J. MACKINTOSH

Department of Experimental Psychology, University of Cambridge

Do males and females differ in general intelligence? Lynn (1994, 1998) and I (Mackintosh, 1996) agree that the answer to this question must depend on the answer to two prior questions: what is the definition of general intelligence, and what tests best measure it? Our disagreement arises partly from different answers to these prior questions, but also from differences in our reading of some not wholly consistent evidence. What answers might one give to these prior questions? They are not independent, for a possible (severely operational) definition of general intelligence is that it is a person's score on one particular IQ test battery. Let us start with that simple possibility.

WAIS full-scale IQ

If we accept the Wechsler tests as the best available measure of general intelligence, then Lynn's claim that men are more intelligent than women is correct. As he has shown, and contrary to much received opinion, there is now good evidence that the average score of men on the WAIS and WAIS-R is significantly higher than that of women (even though on the original Wechsler–Bellevue test, women obtained rather higher scores than men). There is also a somewhat smaller difference in favour of males on the WISC.

I doubt, however, that anyone would accept the evidence of male superiority provided by the Wechsler tests if every other IQ test ever invented revealed either no sex difference or one favouring females. In practice, therefore, we must surely turn to a second definition.

Conclusion. There *is* a sex difference in favour of men on the WAIS-R, but this alone could never be sufficient to establish a sex difference in general intelligence.

Overall IQ score on any diverse test battery

Do other test batteries give the same answer as the Wechsler tests? Lynn (1998) rapidly and, from his point of view wisely, passes such evidence by. The fact is that other large-scale surveys, employing a variety of different test batteries, have yielded every possible outcome. Some have found evidence of male superiority, e.g. of about 5·5 IQ points on the ASVAB; others of even greater female superiority, e.g. of 7·9 IQ points on the GATB (both studies reported by Jensen, 1998). But the most common outcome has been that the sex difference, while still oscillating between marginal

superiority for one sex or the other, is too small to be taken seriously. Herrnstein & Murray (1994) reported a difference of 0·9 IQ points in favour of males; but Lubinski & Humphreys (1990) and Eliot (1983) found differences of 0·3 and 0·5 points respectively in favour of females; while Feingold (1988), analysing the 18-year-old scores in the 1980 standardization sample of the DAT, found a marginally larger difference of 1·65 points in favour of females.

Conclusion. There is no *consistent* evidence of male superiority on other large-scale test batteries. Different test batteries yield all possible outcomes.

Overall test score on a 'good' test of intelligence, or the general factor, g, extracted from large test batteries

Is it possible to cut through this confusion by establishing that some IQ test batteries are better measures of general intelligence than others, and that the good tests all agree in pointing to one conclusion? I do not believe that there is any principled argument sufficient to establish that some test batteries (e.g. those that yield male superiority) are inherently better measures of general intelligence than others that yield no such evidence. Lynn's discussion is unconvincing. He is, of course, happy to follow a certain popular consensus which sees the Wechsler tests as the best available general test battery. But other authorities are more sceptical of the virtues of the WAIS (e.g. Carroll, 1993, p. 702). Lynn states that the WAIS is a good test because it measures a wide range of abilities, including verbal and spatial abilities and non-verbal reasoning. But Wechsler expressly decided *not* to include tests of abstract reasoning in his test batteries, since, for reasons best known to himself, he believed that they were not necessarily good measures of intelligence (Wechsler, 1958, p. 62). And the WAIS does not include any such sub-test, as is evident from Snow, Kyllonen & Marshalek's (1984) analysis of the interrelationship between various IQ tests. They used multidimensional scaling to suggest that most of the WAIS sub-tests are measures of verbal ability or Gc, and that none is a measure of non-verbal reasoning or Gf.

Lynn (1998) also seeks to dismiss other test batteries that yield evidence of female superiority, such as the DAT, since it includes measures of 'minor cognitive skills like spelling and clerical accuracy' (which happened to yield substantial female superiority). It is true that the DAT is partially geared towards measures of school attainment, such as spelling, that do not appear in most IQ test batteries. Another such sub-test is mechanical reasoning (questions about pulleys and levers), on which males do better than females. If we exclude both spelling and mechanical reasoning, the overall difference among 18 year olds in the 1980 standardization sample on the remaining six sub-tests reduces to 1·2 IQ points in favour of females. If we just exclude spelling, the overall difference becomes 1·0 points in favour of males, but at this point the decision what to include has surely become merely arbitrary.

Is not general intelligence defined as *g*, the general factor common to all IQ tests? Could we not ascertain whether there is a sex difference in *g*? Jensen (1998) agrees with Lynn (1994) in seeing this as an answer to our question, but unlike Lynn believes that there is no sex difference in *g*. This illustrates the central problem with this suggestion: *g*, defined as the first principal component of a given battery of tests, is no more stable than the average of scores on different test batteries. Indeed it may be less stable. As

Lynn (1994) showed, male superiority on the first principal component of the WAIS is actually rather greater than on overall WAIS IQ, amounting to about 4 IQ points, because those sub-tests on which males outscore females have higher loadings on the general factor than do those sub-tests on which females outscore males. Jensen demonstrated that there were no significant sex differences on the general factor extracted from other test batteries. Factor analysis of the DAT would most probably indicate female superiority in *g* since males outscore females on only two of the eight sub-tests. The study of sex differences makes it abundantly clear, as I argued before (Mackintosh, 1996), that the general factor extracted from one test battery cannot be the same as that extracted from another.

Conclusion. Neither factor analysis, nor appeal to the inherent superiority of one test battery over another, will serve to resolve the impasse. Males outscore females on some test batteries; females outscore males on others. The differences are usually trivially small.

The average of Gf, Gc and Gv

Lynn (1994) suggested that a hierarchical model of general intelligence, which he attributed to Gustafsson (1984), defines general intelligence as the average of scores on tests of three second-order abilities: non-verbal reasoning (fluid intelligence or Gf), verbal ability (crystallized intelligence or Gc) and spatial ability (Gv). Lynn's suggestion betrays a serious misunderstanding of the nature of hierarchical models. No such model would simply take the unweighted means of scores on tests of secondary abilities as its measure of general intelligence. Such scores need weighting in accordance with their loadings on the higher-order general factor. Gustafsson (1984) argued that Gc and Gv loaded significantly less strongly on the general factor than did Gf: indeed, he suggested that the loading of Gf on the general factor was essentialy unity, this being the basis for the view, which Lynn dismisses as 'too narrow to command assent', that general intelligence be defined as abstract reasoning ability.

In my earlier comment on Lynn's views, I wondered why one should not include other factors besides Gf, Gc and Gv, such as Gs or perceptual speed, in one's hierarchical definition of general intelligence. Lynn dismisses this as conceptually incorrect, since it would amount to the addition of a first-order factor to three second-order factors. Maybe it would in Gustafsson's model, but not in others. Carroll (1993) has suggested that there are at least half a dozen second-order factors (including Gs) that should be accepted.

Lynn's 'hierarchical' definition thus fails to address seriously the question of what secondary abilities should be included, and with what weighting. He also takes a cavalier attitude to the question of what tests should be accepted as good measures of the three second-order abilities he does recognize. Lynn (1994) used scores on the Wechsler verbal scales as his measure of Gc, the average performance on the DAT verbal and abstract reasoning tests since 1947 as his measure of Gf, and the average of the data reported by Linn & Peterson (1985), in their meta-analysis of sex differences on spatial IQ, as his measure of Gv. He offered no serious justification of his choice of these particular measures. The WAIS is virtually unique in finding evidence of substantial male superiority on measures of Gc. Why did he not take the average sex

difference on verbal abilities, reported in Hyde & Linn's (1988) meta-analysis? (Excluding SAT-V scores, this yielded a female advantage of 1·65 points. The SAT-V test, although correlating moderately with other measures of Gc, is clearly measuring something else, since we know that scores on tests of Gc were increasing from one year to the next at a time when SAT-V scores were declining quite sharply.) Why did he not use the score on the language test in the latest standardization of the DAT, which yields a female advantage of 6 IQ points? Since Lynn's original article appeared, Hedges & Nowell (1995) have published a meta-analysis of large-scale surveys of various abilities. On 'reading comprehension', the female advantage was 1·35 points; on vocabulary, the male advantage was 0·3 points. There can be no serious justification for Lynn's use of a male advantage of 2–3 IQ points on measures of Gc.

No one disputes that males obtain higher average scores than females on most tests of spatial ability or Gv. For Lynn's purpose, the average of the difference summarized by Linn and Peterson may be as fair an estimate as any, although as Linn and Peterson noted, and as Voyer, Voyer & Bryden (1995) have confirmed, this average conceals differences ranging from no more than 1 or 2 points on some kinds of test to 10 or more points on others. For anyone concerned to achieve a deeper understanding of the nature of spatial ability, the interesting question is why different types of test yield such different outcomes.

I turn to the proper measure of Gf in the next section.

Conclusion. Since there is a significant sex difference on most tests of spatial ability, it is obvious that the definition of general intelligence as the arithmetic mean of scores on tests of spatial ability plus one or two other kinds of test (of, say, verbal ability or abstract reasoning) will yield an overall difference in favour of males. But no serious model of intelligence would accept such an arithmetic mean as a definition of general inteligence, and it is obvious that the effects of the spatial tests would be diluted by the addition of yet more other kinds of test. And it would surely be more sensible to acknowledge that males outscore females on tests of spatial ability, and perhaps even to attempt to understand why such a difference occurs, rather than to insist that men are generally more intelligent than women because spatial tests form a major part of our measure of general intelligence.

Abstract reasoning

I suggested, following Gustafsson (1984) and Snow *et al*. (1984), that it might be worth equating general intelligence with Cattell's Gf. This seems a particularly sensible suggestion in the present context because it might help to cut through the imprecision that inevitably accompanies definitions of general intelligence as the average score on a diverse set of tests, some favouring males, other females.

Once again, the question necessarily arises: what tests shall we accept as good measures of Gf? Lynn (1994) took the average of the verbal and abstract reasoning test of the DAT in all standardization samples since 1947, yielding a male superiority of 3·4 points. Why? There is clear evidence that virtually all sex differences on the DAT have been declining since 1947 (Feingold, 1988). In the 1980 standardization, the sex difference among 18 year olds on these two tests amounted to 0·15 IQ points in favour of males. If the numerical ability test is added, even this small difference vanishes.

As I mentioned in my original article, if we took Hedges & Nowell's (1995) data as an up-to-date measure of Gf, one of the two studies in their survey yielded a male advantage of 0·6 IQ points, the other a female advantage of 3·3 points.

Finally, as I said, 'the paradigm test of non-verbal, abstract reasoning ability is, of course, Raven's Matrices'. What do the data from Raven's tests say? I noted that Court's thorough review had reported all possible results: male advantage, female advantage and no difference (Court, 1983). Lynn (1998) dismisses this because 'It does not analyse the sex difference by age and therefore fails to address the crucial component of my theory that it is only at the age of about 16 that the male advantage in abstract reasoning begins to appear'. In fact, it is a reasonably simple matter to distinguish between studies of children and of adults in Court's review. Allowing for some arbitrary judgements, I counted some 20 studies of adults, of which 10 yielded no sex difference, 7 one in favour of males and 3 one in favour of females. That might perhaps average out to a small male advantage. But there are some problems with one of the larger studies yielding an overall male advantage (Heron & Chown, 1967), which is one of the three now relied upon by Lynn (1998) to establish an overall male advantage of 5·2 IQ points. Let us examine these three studies in more detail. The most notable feature of Heron and Chown's results is the interaction of the sex difference in test scores with age and social class. There was some decline in test scores with age in both males and females, but this decline was more pronounced for females than males, especially in the lowest two classes. Thus among 20–40 year olds in social classes I and II, there was no difference in test scores; among 50–70 year olds in classes IV and V, the male advantage was over 10 IQ points.

The interaction with age is also evident in the other two studies cited by Lynn (1998). Neither provided tabular data, so the numbers in Lynn's Table 1 are simply his estimates of the overall scores from graphically presented data. I shall not attempt any similar precision. But it is evident from Fig. 4 of the paper by Wilson *et al.* (1975) that there was essentially no difference in the test scores of the two sexes between the ages of 18 and 25, a modest difference in favour of men from age 25 to 35, and a larger difference amounting to perhaps 4–5 IQ points from the age of 40. The sex $\times$ age interaction was significant. Deltour's (1993) data show a similar interaction, the size of the sex difference after the age of 60 being about twice that of the difference observed between the ages of 20 and 40 (although all the differences were larger than those reported by Wilson *et al.*, 1975).

Why does the male advantage in Raven's scores increase with age? Since all three studies were cross-sectional, there is no way of knowing whether it is because women's performance on Raven's tests deteriorates more rapidly than that of men as they grow older, or because earlier generations of women obtained lower scores on Raven's tests than men, but this difference has begun to disappear. As I noted above, the sex difference among 18 year olds on the verbal and abstract reasoning tests of the DAT declined from 5 IQ points to 0·15 points between the 1947 and 1980 standardizations (Feingold, 1988). This points strongly to a secular change. And the only longitudinal study of which I am aware reveals no increase in the sex difference on tests of abstract reasoning as people grow older (Schaie, 1996). It seems probable, therefore, that we are seeing a secular change from one generation to the next.

What is the size of the sex difference among young adults today on tests of abstract

reasoning? Flynn's analysis of Israeli Raven's scores suggests that males outscore females by between 1·4 and 1·7 IQ points (Flynn, 1998). Over the past 2 years, I have collected the scores of 180 Cambridge undergraduates, half male, half female, on Raven's Advanced Matrices: they show a male advantage of less than 0·5 IQ points. When added to the data of Feingold (1988) and Wilson *et al*. (1975), the implication is that the difference is somewhere between zero and 1·7 points, a conclusion consistent with Lynn's analysis of the British standardization of the DAT (Lynn, 1992). Data from Belgium (Deltour, 1993) and Ireland (Lynn, 1996) imply a somewhat larger difference. It is not clear why. But set against these results, one of the two studies in Hedges & Nowell's (1995) survey found a female advantage of just over 3 points in a sample of over 17,000 American high-school seniors.

Conclusion. If general intelligence is defined as Cattell's Gf, best measured by tests such as Raven's Matrices or the verbal and abstract reasoning tests of the DAT, then the sex difference in general intelligence among young adults today in the USA, Britain or Israel is trivially small, surely no more than 1–2 points either way. If I was thus over-confident in my assertion that there was *no* sex difference, Lynn's (1998) suggestion that there might be a male advantage of 5·5 points seems a serious over-estimate. The difference may be larger in some countries than others, and there is good reason to believe that the male advantage was larger in earlier generations than today. Neither of these qualifications, however, provides much comfort for Lynn's general thesis that a sex difference in general intelligence of some 4 IQ points in favour of males is predicted by the sex difference in brain size.

References

Carroll, J. B. (1993) *Human Cognitive Abilities*. Cambridge University Press, Cambridge.

Court, J. H. (1983) Sex differences in performance on Raven's Progressive Matrices: a review. *Alberta J. educ. Res*. **29**, 54–74.

Deltour, J. J. (1993) *Echelle de Vocabulaire Mill Hill et du SPM de JC Raven*. Editions L'Application des Techniques Modern SPRL; Braine de Chateau, Belgium.

Feingold, A. (1988) Cognitive gender differences are disappearing. *Am. Psychol*. **43**, 95–103.

Flynn, J. R. (1998) Israeli military IQ tests: gender differences small; IQ gains large. *J. biosoc. Sci* **30**, 000–000.

Gustafsson, J. E. (1984) A unifying model of the structure of intellectual abilities. *Intelligence* **8**, 179–203.

Hedges, L. V. & Nowell, A. (1995) Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science* **269**, 41–45.

Heron, A. & Chown, S. (1967) *Age and Function*. Churchill, London.

Herrnstein, R. J. & Murray, C. (1994) *The Bell Curve*. Free Press, New York.

Hyde, J. S. & Linn, M. C. (1988) Gender differences in verbal ability: a meta-analysis. *Psychol. Bull*. **104**, 53–69.

Jensen, A. R. (1998) *The g Factor*. Praeger, Westport, CT.

Linn, M. C. & Peterson, A. C. (1985) Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Dev*. **56**, 1479–1498.

Lubinski, D. & Humphreys, L. G. (1990) A broadly based analysis of mathematical giftedness. *Intelligence* **14**, 327–355.

Lynn, R. (1992) Sex differences on the Differential Aptitude Test in British and American adolescents. *Educ. Psychol*. **12**, 101–106.

Lynn, R. (1994) Sex differences in intelligence and brain size: a paradox resolved. *Personal. individ. Diffs* **17**, 257–271.

Lynn, R. (1996) Differences between males and females in mean IQ and university examination performance in Ireland. *Personal. individ. Diffs* **20**, 649–650.

Lynn, R. (1998) Sex differences in intelligence: a rejoinder to Mackintosh. *J. biosoc. Sci.* **30**, 000–000.

Mackintosh, N. J. (1996) Sex differences and IQ. *J. biosoc. Sci*. **28**, 559–571.

Schaie, K. W. (1996) *Intellectual Development in Adulthood: The Seattle Longitudinal Study*. Cambridge University Press, Cambridge.

Snow, R. E., Kyllonen, P. C. & Marshalek, B. (1984) The topography of ability and learning correlations. In: *Advances in the Psychology of Human Intelligence*, Vol. 2. Edited by R. J. Sternberg. Hillsdale, NJ, Erlbaum.

Voyer, D., Voyer, S. & Bryden, M. P. (1995) Magnitude of sex differences in spatial ability: A meta-analysis and consideration of critical variables. *Psychol. Bull*. **117**, 250–270.

Wechsler, D. (1958) *The Measurement and Appraisal of Adult Intelligence*. (4th edn.). Williams & Wilkins, Baltimore.

Wilson, J. R., De Fries, J. C., McClearn, G. E., Vandenberg, S. G., Johnson, R. C. & Rashad, M. N. (1975) Cognitive abilities: Use of family data as a control to assess sex and age differences in two ethnic groups. *Int. J. Aging & hum. Dev.* **6**, 261–276.