Sex Differences in Variability in General Intelligence

A New Look at the Old Question

Wendy Johnson,[1,2] Andrew Carothers,[3] and Ian J. Deary[1]

[1]MRC Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, United Kingdom, [2]Department of Psychology, University of Minnesota—Twin Cities, [3]Public Health Sciences, University of Edinburgh Medical School, United Kingdom

ABSTRACT—*The idea that general intelligence may be more variable in males than in females has a long history. In recent years it has been presented as a reason that there is little, if any, mean sex difference in general intelligence, yet males tend to be overrepresented at both the top and bottom ends of its overall, presumably normal, distribution. Clear analysis of the actual distribution of general intelligence based on large and appropriately population-representative samples is rare, however. Using two population-wide surveys of general intelligence in 11-year-olds in Scotland, we showed that there were substantial departures from normality in the distribution, with less variability in the higher range than in the lower. Despite mean IQ-scale scores of 100, modal scores were about 105. Even above modal level, males showed more variability than females. This is consistent with a model of the population distribution of general intelligence as a mixture of two essentially normal distributions, one reflecting normal variation in general intelligence and one reflecting normal variation in effects of genetic and environmental conditions involving mental retardation. Though present at the high end of the distribution, sex differences in variability did not appear to account for sex differences in high-level achievement.*

The variability hypothesis, which posits that general intelligence may be more biologically variable in males than in females, has a long history in both scientific and political writings. In recent years, it has received renewed attention as an explanation for the presence of greater numbers of males than females in technology, engineering, and the highest levels of scientific research. As is often the case in areas of research involving demographic group differences, much of the literature has been emotionally charged, and the empirical data have often been ambiguous. Even when observed results have seemed clear, researchers have raised issues involving adequacy of overall sample size and differences in relative selectivity of male and female samples, relevant experiential background, participant responses to the testing situation, and rates of physical and cognitive maturation that call into question both the relevance of the observations to the question and the appropriateness of any attribution to underlying biology. Moreover, issues related to variation have often been confounded with issues related to central tendency in distributions of scores. In this article, we review the development of the variability hypothesis, present new distributional data from two unique datasets consisting of almost entire populations, and review the implications of these data in the context of the environment in which we live.

Because the subject is emotionally charged, we lay out our definitions of the terms we use and note that our usages are not the only senses in which these terms can be and are often used. Our usages reflect only a desire for clarity and brevity, and they are intended to avoid involvement in cultural struggles over "correct" terminology. We use the term *general intelligence* to mean the ability to use combinations of preexisting knowledge and abstract reasoning to solve any of a variety of problems designed to assess the extent to which individuals can benefit from instruction or the amount of instruction necessary to attain a given level of competence. The problems can be posed either verbally or figurally. We assume that tests composed of such problems are, to varying degrees, valid measures of the construct to which we refer, particularly within cultural groups sharing a common place and time. We use the term *sex* to refer to males and females generally, regardless of the sources of any differences between them. Finally, we use the term *biological* to refer

Address correspondence to Wendy Johnson, Department of Psychology, University of Minnesota—Twin Cities, 75 East River Road, Minneapolis, MN 55455; e-mail: wendy.johnson@ed.ac.uk.

to innate characteristics or processes (usually involving genes) that have potential for development from birth and are inherent in the essential nature of the individual. We chose this term because it was most commonly used in the early writings about the variability hypothesis. The potential for development of biological characteristics is contingent on the existence of environmental circumstances that make the development possible.

This is not an article about values. Values create the emotionally charged climates pervading discussions of sex differences, making it difficult to evaluate scientific data objectively. Values are extremely important and appropriately form the basis of many actions and social contracts. But the laws of nature are not responsible to us or to our values and may not conform to them. It is important to understand the laws of nature as completely as possible within our circumstances in order to actualize our values as we intend. This article is thus an attempt to make an objective exploration of scientific data about the way the laws of nature related to the variability hypothesis are manifested.

HISTORY OF THE VARIABILITY HYPOTHESIS

The question of biologically based sex differences in variability in general intelligence dates at least from the writings of Charles Darwin on evolution. Shields (1982) has described much of this history in detail. From the beginning, Darwin stressed the importance of variation in the process of natural selection, but the relation between sex and variability only attracted interest when he articulated the process of sexual selection (Darwin, 1897). Darwin used the term *sexual selection* to refer to individual differences in reproduction rates resulting from the relative presence of secondary sex traits that confer advantages in mating but not in survival, and he used the term *variability* to describe individual tendencies to deviate from the median. To Darwin, the existence of greater male variability in sexually selected traits was a fact, but its cause was unknown. He believed, however, that this variation was maintained through the tendency of adult characteristics to be transmitted only to offspring of the same sex, which in turn helped to establish and maintain sexually selected traits. It was not a major focus of his attention.

To some scientists, however, it was the critical point needed to establish that males were the superior sex in the human species and thus responsible for the species' evolutionary advance. Though Darwin described evolution as essentially random, many scientists adopting and extending the concept thought and wrote of it as progressive or directed toward advancement because it involves the transmission of variation that contributes to adaptive survival. To them, variability was the mechanism for the attainment of greatness or extreme positive deviation from the norm, and because males were more variable, they were superior. For example, Geddes and Thomson (1890) proposed that sex evolved through the development of sex differences in metabolism that pervaded all the biochemical processes in every cell of the body. These metabolic differences paralleled the differences between reproductive cells and extended to intellectual and temperamental traits: ova are large and relatively inert, sperm are small and active; thus females were proposed to be biologically efficient, passive, submissive, conservative, and receptive, whereas males were thought to be biologically spendthrift, enterprising, progressive, and imaginative. In so doing, they appear to have confused the tendency for the individual to deviate from the median consistently (i.e., interindividual variability) with lack of individual consistency (i.e., intraindividual variability). This is an issue to which we will return.

Ellis (1894), the influential sexologist, was the first to write clearly about the question of interindividual variability and to make use of empirical data to substantiate his claims. He proposed that males showed greater interindividual variability than did females in both physical and psychological characteristics, including what we would today call general intelligence. Unlike many of his era, he focused on sex differences in the incidences of negative as well as positive deviations from the median to provide evidence for this, noting that the scientific literature of the day indicated that there were more males than females in homes for the mentally deficient as well as more male than female "geniuses" (p. 366). Though he did not articulate the biological mechanisms responsible for the sex differences in variability, he believed that the negative and positive deviations were from the same biological tendency and that both male and female distributions were symmetrical. Thus, he believed that the greater negative deviations from the median in males were compensated by greater positive deviations.

This sparked a sharp rebuttal from Karl Pearson (1897). His specific points were primarily statistical, but he also argued conceptually that a biologically based sex difference in interindividual variability could only exist if genetic transmission of traits was sex-specific, which he considered impossible. Reasoning that evolutionary theory implied that the intensity of the struggle for survival increased selection pressures and thus reduced variability, he concluded that men should be less variable than women because, at least in recorded history, men have had "a harder battle for life." (p. 259). He dismissed the empirical data amassed by Ellis, claiming that variability could only be compared for traits that could not be considered secondary sex characteristics. In his mind, this limited consideration of the question to physical variation in normal individuals and the data he considered and analytical techniques he used substantiated his conclusion.

Despite Pearson's eminence and the unassailable quality of his statistical analyses, it was Ellis' hypothesis of greater male variability that prevailed during the early years of the 20th century, and researchers increasingly focused their attention on variability in general intelligence. In part, this could be attributed to the appeal of Francis Galton's (1869/1952) study of hereditary genius. Galton compiled a list of eminent people

throughout western history and examined the families of these individuals to detect patterns in the transmission of genius. He concluded that genius was transmitted genetically, which meant to him that it was an innate quality that would inevitably manifest if possessed, whatever the environmental conditions. There were very few women on Galton's list and on the lists of other researchers who followed with similar studies, and this was taken as evidence in support of the variability hypothesis with respect to general intelligence.

The data provided by the development of mental tests further substantiated the position of the variability hypothesis in the minds of educators because the tests imbued this position with quantitative and scientific authority. Edward Thorndike was one of the foremost advocates of both the validity of mental testing and biologically based greater variability in general intelligence of males. He (Thorndike, 1906) believed that there were, at most, small sex differences in mean levels of mental abilities but that the clear sex difference in variability meant that men would be much more heavily represented than women at the highest levels of ability. To him, this had implications for education: women should be channeled into programs preparing them for more practically oriented occupations, and the highest level professional programs should be reserved for men. Though this view was well received by many of his contemporaries, women had started to enter graduate school programs in the United States. By 1910, they accounted for 11% of the PhDs granted (Woody, 1929). This may have been part of the reason for ongoing confusion between sex differences in intellectual variability and sex differences in mean levels of mental abilities and about the interpretation that the observed differences were biologically based. That is, Thorndike was correct that if men actually are more variable in ability than are women, it is reasonable to expect differences in their rates of participation in select graduate school programs, even if there are no mean level ability differences in the population as a whole. At the same time, the existence of sex differences either in means or variances in ability says nothing about the source or inevitability of such differences or their potential basis in immutable biology.

These points tended to be obscured in the writings of the time, as they often are today. For example, Thompson (1903) examined sex differences in mental abilities, motor skills, sensory abilities, and emotional processes in a sample of 25 women and 25 men. Though the sample size would be considered unacceptably small by today's standards, the work was notable for the large number of traits included and for the care taken to match male and female participants for age and educational and social background. She found small mean differences in sensory thresholds, memory, and speed of association that favored women and similar differences in sensory discrimination and puzzle solving that favored men. These are patterns typical of more modern studies as well (Halpern, 2000; Jensen, 1998). She interpreted her results as demonstrating the inconsistency and tenuous nature of the assumptions on which theories of the biological basis of sex differences in abilities, such as Geddes and Thomson's (1890), were based and noted that cultural differences in treatment of boys and girls could easily explain the small differences she observed. This led her into criticism of the variability hypothesis, which her data could not address. She argued correctly, however, that even if men are more variable in ability than women, it is not reasonable to infer that positive intellectual traits such as originality and inventiveness are in general characteristic of men but not of women.

One early twentieth-century scholar who addressed the variability hypothesis from a thorough empirical and relatively dispassionate perspective was Leta Hollingworth. Seeking to examine characteristics that were unlikely to be subject to environmental influences, she and her associate Helen Montague collected physical measurements such as length, weight, and head size of over 2,000 neonates (Montague & Hollingworth, 1914). They found no evidence of greater variability in males than in females, though they acknowledged that this conclusion about physical measures in neonates said little about relative amounts of variability in mental abilities. Hollingworth also pointed out that the arguments for the variability hypothesis to date (and most of those since then) relied on the assumption that mental abilities are normally distributed in both sexes (Hollingworth, 1914). This assumption is important because it implies that any difference in variability dictates an associated difference in range, and the argument that greater male variability implies greater numbers of males at the highest levels of ability rests on the existence of this difference in range. But few data about the degree to which the distributions of abilities were actually normal were available then (or have become available since then).

Hollingworth (1914) demonstrated hypothetically that, even without altering the normal distribution's property of symmetry, deviations from the normal distribution's *kurtosis*, or the thickness of the tails of the distribution, could produce very different results with respect to the relative numbers of males and females at the highest levels of abilities while still maintaining greater overall male variability. In particular, the conclusions depended highly on whether or not the range of scores was actually greater in males than in females. She also pointed out that it was impossible to infer that biology was responsible for the greater numbers of both eminent and institutionalized feeble-minded men, as straightforward social explanations could also explain these data. Women were both socially constrained from attaining eminence by their responsibility for child-rearing and protected from institutionalization for feeble-mindedness by their socially expected economic dependence on men.

With the inception of World War I, the focus of psychologists interested in mental ability testing shifted from sex differences to the development and administration of fast, accurate, and economical tests to army inductees. Tests were also developed to identify individuals with more specific kinds of cognitive abilities. The primary goal in developing these tests was to make

possible the tailoring of education to the individual, and overall statements about the relative suitability of males and females for various levels of education largely disappeared from the literature (Shields, 1982). Studies did continue to quantify variability and compare the results by sex, and many made use of increasingly large samples. The ways in which they did so were inconsistent, however, and conclusions were not always accurate. For example, Fraiser (1919) compiled grade-level achievement test results for over 60,000 13-year-olds from 20 U.S. cities. Noting that the coefficient of variation (the ratio of the standard deviation, *SD*, to the mean) was 1.6 for males and 1.4 for females, he concluded that the difference was so slight that no difference in variability could be said to exist, despite the fact that the difference indicated a variance ratio that was both highly statistically significant (p effectively 0), and meaningful in terms of the expected numbers of males and females at the extremes of the overall distribution: 1.40:1 males to females in the lowest grade, 1.35:1 in the next lowest, and 1.28:1 in the highest grade—the ratio was 0.87:1 in the next highest grade, but this was in the presence of an overall difference of a quarter of a grade in favor of females.

Another landmark study was conducted by McNemar and Terman (1936). They searched the literature for studies examining sex differences in variability and, in an era before meta-analysis, summarized the results of the tests for significant differences. Though they observed that studies using samples of college students tended to show greater male variability, they recognized that such samples were not representative of the population. Overall, they concluded that there was no consistent evidence for greater male variability. Jensen (1973) reviewed many of the same studies, but, again, prior to the development of formal meta-analysis, examined and cumulated variance ratios to summarize them across studies rather than simply tabulating those that showed significant differences. Using this approach, he concluded that males tended to show greater variability in general intelligence than females.

In 1932 and 1947, the Scottish Council for Research in Education administered the same group general intelligence test to nearly all schoolchildren born in 1921 and 1936 (Deary, Whalley, Lemmon, Crawford, & Starr, 2000; Scottish Council for Research in Education, 1933; Scottish Council for Research in Education, 1949). The stated purposes of the first survey were to quantify the number of people in Scotland who were "mentally deficient," and to "obtain data about the whole distribution of the intelligence of Scottish pupils from one end of the scale to the other" (Scottish Council for Research in Education, 1933, p. 23). Children who were deaf, blind, or absent from school on the testing day were not tested, nor were students from a few private schools as well as some from schools that received inadequate numbers of tests. Children in special education programs were specifically included unless their handicaps prevented them from producing a valid score in the group testing situation. In total, the Scottish Mental Survey of 1932 (SMS32) tested 87,498 children, about 95% of the surviving population born in 1921. The Research Council intended, at the time of the first administration, to repeat the survey on some future occasion. They had thus made requests for funding to do so and had preserved the materials and notes regarding the survey administration for this future use.

The specific impetus for the second survey was a decline in the birthrate since 1921 and concern that this decline might have caused a dysgenic trend in intelligence (Scottish Council for Research in Education, 1949). Considerable care was thus taken to ensure that the conditions of test administration and sample ascertainment were similar to those used in 1932, despite the fact that the schools still suffered disruption from World War II. In particular, the fact that the same test would be used was not generally known before the tests were mailed to the schools, and teachers received very specific administration instructions consistent with those given in 1932. The schools provided demographic information for all students on the rolls on the day of administration, with one indication of missing score for those who were absent that day and a separate indication for those who did not complete the test due to mental or physical handicap. In total, the Scottish Mental Survey of 1947 (SMS47) obtained demographic information for 75,211 children, of whom 4,406 (5.86%) were absent on the day of testing and an additional 532 (.7%) were deemed unable to complete the test. The best information available indicated that this latter group had somewhat lower general intelligence than those who scored 0 on the test in 1932. They were included in the tabulations of 1947 test results with scores of 0.

To the present date, these are the two most clearly population-representative samples of general intelligence test scores that have been compiled. This, along with the consistency with which the tests were administered, continues to make them uniquely valuable for addressing questions regarding the distribution of general intelligence. Although the 1932 sample consisted of more than 87,000 children, there was no significant sex difference in mean scores (in the data used here,[1] the mean for males was .28 raw points higher than that for females or .018 standard deviation). The 1947 sample, which consisted of almost 71,000 children, showed a significant difference of 1.72 raw points in favor of females or .109 standard deviation. In both samples, the variances of males' scores exceeded those of females', with variance ratios of 1.12 and 1.17 respectively. There were, however, similar and substantial departures from normality in both samples. Anastasi (1958) used these departures to attribute the sex differences in variances to excesses of males with scores that were below average but not in the retarded range. It is these two Scottish samples that form the basis of the new empirical analyses we describe in this article, so we will discuss the dis-

[1]For these analyses, we deleted the results of participants from Fife, Wigtown, and Angus counties from SMS47 because the data ledgers for these counties were missing from SMS32. Some entries from these counties from private schools do remain in both surveys.

tributions of scores they produced in detail in the sections to follow.

Though studies of sex differences in both mean levels and variability of general intelligence have continued over the last 50 years, it has become increasingly clear that any difference in mean levels is small. Moreover, scientific evidence for general intelligence notwithstanding (Carroll, 1993; Horn & Knapp, 1973, 1974; Hunt, 2001; Jensen, 1998; Lubinski & Benbow, 1995; Messick, 1976), there is strong popular interest in the idea that intelligence reflects multiple specific abilities (Gardner, 1993; Guilford, 1985). This has led to both relative popular acceptance of the observation that males tend to perform better on some kinds of tests whereas females perform better on others and to a shift in focus to investigating these more specific differences in both means and variability. In this climate, Maccoby and Jacklin (1974) carried out one of the first major meta-analyses of studies of sex differences published in American journals in the preceding 10 years. They concluded that sex differences in tests of composite cognitive abilities were negligible, but that, beginning in adolescence, females tend to perform slightly better on many kinds of verbal tests, whereas males tend to perform better on quantitative and visuospatial tests. There is also a large literature from a variety of perspectives suggesting that these three groups of abilities are the lines along which specific cognitive abilities most readily segregate (e.g., Johnson & Bouchard, 2005; Snow, Corno, & Jackson, 1996; Snow & Lohman, 1989; Vernon, 1964).

Though it is not often referenced, Maccoby and Jacklin (1974) also noted that scores of males were more variable than those of females in quantitative and visuospatial tests, though there was no apparent sex difference in variability on verbal tests. This general pattern of results has generally been corroborated by more recent studies (N.S. Cole, 1997; Deary, Irwing, Der, & Bates, 2007; Feingold, 1992; Hedges & Nowell, 1995; Lohman & Lakin, in press; Strand, Deary, & Smith, 2006). With the exception of Strand, Deary, and Smith (2006), however, most of the prominent studies have been carried out in the United States. Feingold (1994) examined the available smaller studies from different countries, concluding that results outside the United States were far less consistent. Clearly, the issue of the very existence of sex differences in the distribution of general intelligence is not resolved, though it might be reasonable to summarize the preponderance of the evidence as favoring greater male variability in at least quantitative and visuospatial tests.

What has been lacking is analysis of the male and female distributions of general intelligence in samples that are large and representative enough of the full population to reveal their distributions clearly. We used the Scottish Mental Surveys of 1932 (Scottish Council for Research in Education, 1933) and 1946 (Scottish Council for Research in Education, 1949) to carry out novel and considerably more extensive analyses to address the questions of the extent to which the population distribution of general intelligence is actually normal and to which there is greater male than female variability, particularly at the high end of general intelligence.

THE DISTRIBUTIONS OF TEST SCORES IN THE SCOTTISH MENTAL SURVEY DATA

The Scottish Council for Research in Education referred to the test used in the SMSs as the "verbal test" because it required literacy and numeracy to understand and complete the items. It was closely related to the Moray House Test (MHT) No. 12, which was one of the series used in secondary school placement in England for many years. It was administered with a time limit of 45 min and has a maximum score of 76 with 75 items. The test includes items of various types, including verbal analogies; identification of synonyms or antonyms of a presented word; and reasoning, arithmetic, and spatial problems. Thus, the three major areas of ability were all represented. In both SMS surveys, possible scores ranged from 0 to 76, but 75 was the maximum score earned.

When placed on an IQ scale with a mean of 100 and an *SD* of 15, it is readily apparent that there were departures from the normal distribution in the SMS32 and SMS47 scores (see Fig. 1; also, see Footnote 1).[2] The most obvious departure was that neither distribution was symmetric: in each survey, more scores were below the peak of the distribution than above it, indicating negative skewness. Skewness was −.211 and −.347 in SMS32 and SMS47, respectively. These skewness values do not sound excessive by common standards, but they were highly statistically significant in these large samples. If we want to understand the distribution of general intelligence at the level of the population, these skewness results matter.

The other dimension of departure from the normal distribution in the SMS data was kurtosis. Formally, kurtosis is the standardized fourth moment of the distribution, or the mean of the fourth powers of its random variables. Like skewness, it is generally measured on a scale for which the normal distribution has a value of 0. Positive values of kurtosis indicate distributions with higher probabilities of values near the mean than the norm, as well as higher probabilities of extreme values. In contrast, negative values of kurtosis indicate distributions with both lower probabilities of values near the mean and lower probabilities of extreme values than the norm. These distributions have higher probabilities of values some moderate distance from the mean than the norm; that is, broader "shoulders." The scale on which kurtosis is measured is somewhat unusual: The smallest value is −2 and the largest is infinite. Distributions with positive kurtosis are termed *leptokurtic*; those with negative kurtosis are termed *platykurtic*. Overall, both distributions of SMS scores

[2]For these analyses, we deleted the results of participants from Fife, Wigtown, and Angus counties from SMS47 because the data ledgers for these counties were missing from SMS32. Some entries from these counties from private schools do remain in both surveys.

were platykurtic, with kurtoses of −.581 and −.495 for SMS32 and SMS47, respectively. Again, these values do not sound excessive by usual standards, but they were highly statistically significant in these large samples, and they matter at the level of the population. In combination with the negative skewness, they suggested the existence of an exaggerated number of scores in the lower score ranges, with the greatest exaggeration occurring somewhere in the middle of the range of scores below 100.

We now have knowledge of the existence of many both specific and nonspecific conditions that disrupt general intelligence, with both genetic and environmental etiologies. These may result in diagnosable mental retardation, defined as an IQ less than 70 accompanied by limitations in adaptive functioning (Raymond, 2006), but they also may disrupt intelligence, though not sufficiently to result in diagnosable mental retardation. Given this, we can probably make more progress in investigating the variability of general intelligence if we consider the possibility that its population distribution is not normal, but rather something close to *mixed normal*, or the combination of two normal distributions. In this conceptualization, one of the distributions would represent the portion of the population not suffering from any condition that disrupts general intelligence, and the other would represent the portion of the population that does suffer from such conditions. This idea is not new. It was perhaps first proposed by Roberts (1945), though Burt (1957, 1963) might be credited with the first empirical description. He noted both a prolonged lower tail and overall "downward asymmetry" in the distributions of general intelligence resulting from standardization samples of the Stanford–Binet in Britain and suggested that these deviations from normality could be attributed to a mixture of "multifactorial and unifactorial" genetic influences on general intelligence. Today, we would call these different forms *polygenic* and *mendelian*, respectively. The issue has also been addressed by Feingold (1995), Humphreys (1988), and Jensen (1998), as well as by researchers in the areas of mental retardation and giftedness (e.g., Robinson, Zigler, & Gallagher, 2000).

To anyone who has spent even a little time with people suffering from conditions that disrupt general intelligence, it is clear that its distribution is still not uniform—it follows some kind of peaked form. Even if the disruptions of general intelligence involved in these conditions are uniform, this distribution could result from differences in the genetic and environmental backgrounds in which the conditions occur. In the normal distribution, the mean, modal, and median values are the same. For the IQ-scaling process we used to produce the SMS scores in both surveys, all three of these values would be 100 if the data were completely normally distributed. But the IQ-scaling process that set the score means at 100, in combination with the negative skews in the overall distributions, resulted in median and modal values that were greater than 100. The extent to which this was true was similar in the two surveys: the median and modal values, respectively, were 101.18 and 103.71 in SMS32 and 101.60 and 105.85 in SMS47. In SMS32, the modal value was at the 56.5 percentile; the SMS47 modal value was at the 61.3 percentile. The medians were lower than the modes because the left shoulders of both distributions were broader than the right shoulders.

We used this information to estimate normal distributions representing the scores of the populations not suffering from any condition that disrupts general intelligence. The results are shown as the normal percent curves in Figure 1. They were the same for each survey, and we used an underlying mean, median, and modal value of 105 to create them. This was based on the modal values from both distributions because, given the skewness and kurtosis properties of the empirical distributions, the modal values most clearly represented the central tendencies of the distributions we wanted to model. We then modeled the proportion of the 50% of the populations that would fall at 105 and at each score value above it if the data were in fact normally

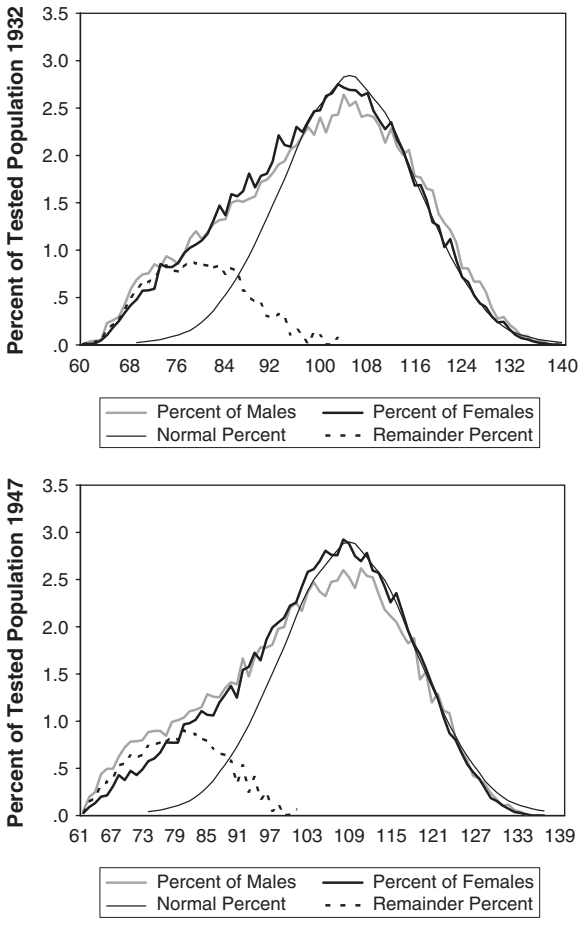

Fig. 1. Distributions of Moray House Test (MHT) scores for the populations of school children born in 1921 and 1936 and tested in the Scottish Mental Surveys of 1932 and 1947 at age 11. The *y* axis is the percentage of the tested population, and the *x* axis is the test score. The normal percent is a normal distribution based on the properties of the distribution in excess of the mode, and the remainder percent was created by separating out a random portion of those causing the negative skew in the overall distribution.

distributed along the range of scores we had from 105 to 140. We did the same in reverse to create the portion of the curve falling below 105. As the figure shows, even those in each survey scoring above the hypothesized central tendency of this model normal distribution did not follow it exactly, but the departures from it were dramatic below the central tendencies.

These departures represented the portions of the population suffering from conditions that disrupt general intelligence. To estimate their distributions, we estimated the numbers of them that fell at each IQ score by taking the differences between the numbers that actually fell at each IQ score in each survey and the numbers that we would have expected based on the estimated underlying normal distribution representing those not suffering from conditions that disrupt general intelligence. The results are shown as the remainder percents in Figure 1. This was 19.68% of the total sample for SMS32, and 18.18% for SMS47. As expected, the distributions of these scores in both samples also proved to be at least roughly normally distributed. The mean was 79.60 for SMS32 and 77.37 for SMS47.

The percentages of the total survey populations we assigned to represent those with conditions that disrupt general intelligence deserve some comment. At almost 20%, they are much higher than any estimates of the incidence rates of diagnosed mental retardation in the population. As most IQ tests are structured, the IQ criterion of 70 for mental retardation translates to 2 *SD*s below the mean of 100. Prevalence rates for mental retardation are commonly given as 2%–3% of the population, which is consistent with the normal distribution of IQ presumed in structuring the tests. Observations of actual prevalence rates for diagnosed mild retardation, however, vary widely from 0.5% to 8.0% (Roeleveld, Zeilhuis, & Gabreels, 1997; Simonoff et al., 2006) and vary even more for more severe retardation. Reasons for this variation are not clear, but rates tend to be lower among those of higher socioeconomic status and higher in population-based samples than in samples collected from educational programs or for test norming.

As the mean IQ in our distribution of disrupted general intelligence was almost 80, many of the individuals we modeled as belonging to that distribution would not have been formally diagnosed as retarded. In developing our model of the population distribution of general intelligence as being comprised of the mixture of two roughly normal distributions, we thus explicitly allowed for the possibility that the conditions of both genetic and environmental origins that disrupt general intelligence can have their effects even on individuals of otherwise high general intelligence background; the conditions markedly disrupt the intelligence of these individuals, but they may not do so severely enough to render the individuals formally diagnosable as mentally retarded.

The observation that the population distribution of general intelligence is not normal has implications for the norming procedures used for intelligence tests, which tend to exclude institutionalized mentally retarded individuals and those with known brain damage or severe behavioral or emotional problems from the standardization samples. The definitions of such cases are subjective, but presumably they would comprise some but not all of those suffering from conditions that disrupt general intelligence. Thus, we should expect to see some negative skew in most standardization samples as currently gathered. If this is the case, scoring as if the scores were normally distributed will produce modal scores above 100. The indicated negative kurtosis value indicates medians in excess of 100 as well. Moreover, the extent to which we observe these characteristics should vary with the definitions for exclusion from standardization samples. There are many difficulties involved in appropriate interpretation of IQ scores for individuals, but this only adds to them. It also introduces uncertainty regarding the conclusions that can be drawn from research studies making use of IQ scores from standardized tests, particularly those examining sex differences in means.

SEX DIFFERENCES IN VARIABILITY IN THE SCOTTISH MENTAL SURVEY DATA

Figure 1 shows the separate empirical distributions for males and females. In both surveys, the percentages of the populations in the very centers of the distributions were smaller for males than for females: The males' scores were dispersed more widely about the means. As Hollingworth (1914) had described in her hypothetical examples, this alone caused greater *SD*s in the male than in the female scores (15.45 for males vs. 14.52 for females in SMS32, 15.59 for males vs. 14.32 for females in SMS47; these differences were highly statistically significant using any of several nonparametric tests for differences in variance). Unlike the hypothetical examples featured by Hollingworth (1914), however, the ranges of scores were greater for males than for females. The test had a floor, a point we discuss in greater detail below. More males obtained the minimum raw score of 0, which translated to an IQ of about 61 (457 males vs. 276 females in SMS32, 413 males vs. 197 females in MSM47; these don't show as single "clumps" in Figure 1 because the data there have been adjusted to remove the effects of age, slightly spreading out the 0 scores earned by SMS participants who were born throughout a single year but tested on a single day).

Males earned the highest scores as well. In SMS32, there were 5 boys scoring IQ-scaled scores of 140 or 141, and 1 girl scoring 140. In SMS47, there were 3 boys scoring 137–139, and no girls in this range. Negative skewnesses were very similar for males and females in both samples (−.22 for males vs. −.20 for females in SMS32, −.31 vs. −.37 in SMS47), but males' scores were more platykurtic than females' (−.63 vs. −.54 in SMS32, −.61 vs. −.39 in SMS47). The difference accounted for the large excess of males over females around IQ of 70 in both surveys, though the "bulge" of males over females was particularly noticeable in SMS47. The existence of these bulges is completely

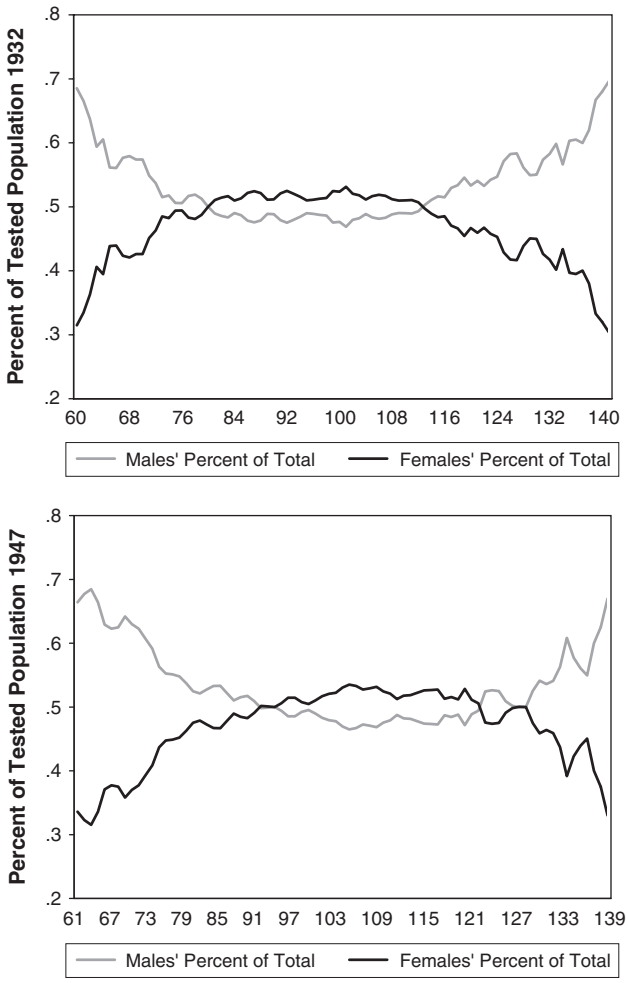

Fig. 2. Proportions of males and females receiving IQ-equivalent Moray House Test (MHT) scores from the populations born in 1921 and 1936 and tested in the Scottish Mental Surveys of 1932 and 1947 at the age of 11. The proportions are 3-point moving averages.

consistent with the existence of conditions that disrupt general intelligence beyond the normally existing variation.

The SMS data indicated greater variance in males than in females. There is some evidence that this greater variance existed at the high end of general intelligence, but there is also evidence that the lower half of the distribution of general intelligence made a greater contribution to this greater variance than did the upper half. This can be seen more clearly in Figure 2. The figure shows the proportions of males and females who participated in the SMS surveys at each level of MHT score, standardized to account for the small differences in numbers of males and females. The curves are of course not completely regular, but the overall pattern is clear and the same in both surveys. Females' scores were more closely clustered in the middles of the overall distributions, causing them to make up more than their shares of the population there. The range over which this was true was not the same in the two surveys: It ran from about 80 to about 115 in SMS32 and from about 90 to about 120 in SMS47. Males' scores were more dispersed, causing them to make up more than their shares of the populations at both extremes of the distributions. The extent to which this was true grew with absolute distance from the middles of the distributions, and the imbalances toward males at the extremes of the distributions were greater than the imbalances toward females in the middles of the distributions. This of course occurred because so many more participants of either sex scored in the middles of the distributions than at their extremes.

The imbalances towards males at the extremes of the distributions were reasonably similar at both the high and low extremes, due to the negative skews in the overall distributions for both surveys and the similarities of the amounts of negative skew in males and females in both surveys. Another way to understand the implications of this is shown in Figure 3. This figure shows the differences in male and female scores for males and females scoring at the same quantiles of their distributions. Thus, for example, in SMS32, males scoring at the fifth percentile of their distribution had scores .1 *SD*s lower than females scoring at the fifth percentile of their distribution. In SMS47, the difference

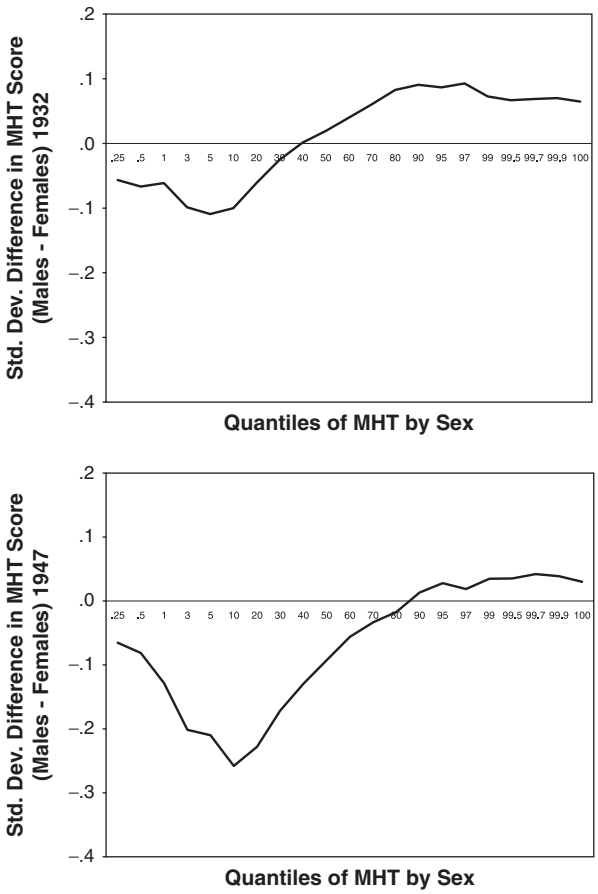

Fig. 3. Standard deviation differences in Moray House Test (MHT) scores between males and females at indicated quantiles of the sex-specific distributions. For example, in SMS47, males at the 10% quantile of their distribution of scores scored almost a third of the full population standard deviation lower than did females at the 10% quantile of their distribution of scores.

was greater at that quantile: males' scores were .2 *SD*s lower than females' scores. This way of displaying the differences in the male and female distributions better highlights the differences between the two surveys. The shapes of the two curves are only very generally similar, with small differences favoring females at the very bottom of the distributions of general intelligence that grow and then fall off to 0 as one moves up the distributions. The curves are relatively smooth, so once the differences reach 0, they continue in favor of males, though the rate of increase in the differences falls off.

Beyond this general similarity, the curves are somewhat different. The differences favoring females in the lower range of the 1932 distribution never reached the magnitudes of the analogous differences in the 1947 distribution. The differences favoring males in the 1947 distribution never reached the magnitudes of the analogous differences in the 1932 distribution. In SMS32, male and female scores were identical when both were at the 40th percentiles of their respective distributions. In SMS47, this point was reached at the 85th percentile. This may be due to random sampling variation, but it is also possible that male performance generally suffered between 1932 and 1947 relative to female performance, or that female performance generally improved relative to male performance. Whatever the explanation, it provides evidence that even when overall patterns are generally similar, male–female ratios at specific points in the distributions of general intelligence may not be stable over time.

It is the male–female variance ratios at various levels of the distribution of general intelligence that have received the most attention in analyses and critiques of the variability hypothesis. Thus, Figure 4 shows these ratios for both surveys. The data clearly refute Anastasi's (1958) claim that the sex differences in variances in the Scottish Mental Surveys were due solely to excesses of males with scores in the low, but not retarded, range. As discussed in more detail below, it is true that there were such excesses, but they did not account for all of the differences in variances. The figure shows clearly that there were more males than females at the high levels of both distributions. The differences were not as extreme as some that have been noted, but they were nonetheless substantial. In addition, though the difference curves in Figure 3 do indicate changes in the variance ratios over time, the changes do not appear to have been very large at the high end of general intelligence where much attention is often focused. That is, in SMS32, the ratio of males to females reached about 2.3 at an IQ equivalent of 140, but it reached about 2.0 at that level in SMS47. Despite an irregular jag in the SMS47 data, however, it was about 1.4 in both surveys at an IQ-equivalent of 132. Moreover, the variance ratios at the high ends of the distributions were pretty much the same as those at the low ends in both surveys.

In Figure 1, we show that both distributions of general intelligence could be more accurately modeled as mixtures of two approximately normal distributions that might be conceived as

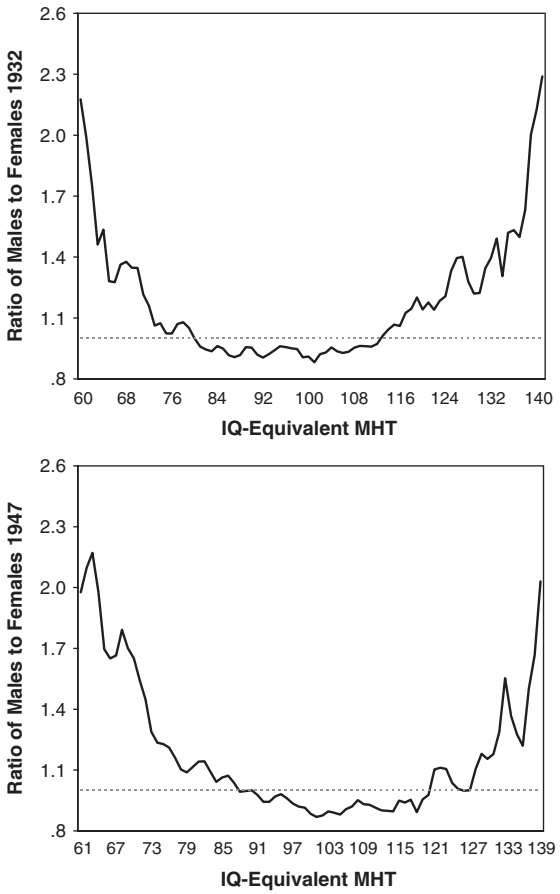

Fig. 4. Ratios of males to females at indicated IQ-equivalent Moray House Test (MHT) scores from the populations of Scottish schoolchildren born in 1921 and 1936 and tested in the Scottish Mental Surveys of 1932 and 1947. Ratios are 3-point moving averages.

describing those with and without conditions that disrupt general intelligence. Table 1 presents descriptive statistics for the two distributions from each survey, separately by sex. The table shows that the process we used to create the two distributions in each survey was very effective in removing the skews from the distributions representing those without conditions that disrupt general intelligence for each sex and that even the distributions representing those with conditions that disrupt general intelligence were relatively symmetrical, especially for SMS47.

Table 1 also shows that the means for both surveys for the distributions representing those without conditions disrupting general intelligence were very similar for males and females: There were effect size differences of .064 favoring males in SMS32 and .013 favoring females in SMS47. Even in these distributions, from which we removed representation of individuals with conditions that disrupt general intelligence, male variability was greater than female variability: the variance ratios were 1.080 for SMS32 and 1.086 for SMS47. Again, this was partly because the distributions remained more platykurtic for males than females (though only very slightly so in SMS32), but in SMS47 it was also because the range was greater in males

TABLE 1

Descriptive Statistics by Sex for the SMS32 and SMS47 Distributions Representing Those With and Without Conditions That Disrupt Intelligence

	Without disruption		With disruption	
Variable	Male	Female	Male	Female
SMS32				
Mean	105.37	104.62	78.91	80.33
Median	105.38	104.64	78.54	80.03
SD	11.90	11.45	8.16	8.24
Skewness	−.013	.004	.334	.250
Kurtosis	−.476	−.399	−.290	−.400
Proportion of total	.800	.817	.200	.183
SMS47				
Mean	104.94	105.08	76.73	78.22
Median	104.98	105.09	76.62	78.39
SD	11.36	10.90	7.60	7.36
Skewness	−.001	−.004	.156	−.020
Kurtosis	−.550	−.500	−.688	−.600
Proportion of total	.798	.850	.202	.150

than in females (70.79–139.50 in males vs. 74.19–136.48 in females).

Greater proportions of males than females were assigned to the distributions representing those with conditions that disrupt general intelligence, and the disruptions suffered by the males appeared to be greater. In SMS32, 20.02% of males fell here, but only 19.33% of females did so, for a ratio of 1.036. In SMS47, 20.25% of males fell here, but only 15.96% of females did so, for a ratio of 1.269. Even in SMS47, however, this ratio was much lower than the ratios often given for specific conditions known to disrupt general intelligence. Again, this is likely to be at least partly because the ratios commonly given focus on cases that are diagnosable as mentally retarded, and the mean IQs for these distributions here were close to 80 and thus well above the threshold for diagnosis of mental retardation. Moreover, the means for females were greater than those for males, with effect sizes of .173 in SMS32 and .199 in SMS47. This suggests that females that fell in these distributions were less affected than the males. There are many biologically based conditions that disrupt general intelligence in the ways we intend to be representing here that would be expected to affect males more often than females, and we know of some environmentally based conditions that are more likely to affect both equally often (such as substance and chemical exposure, birth trauma, etc.). We also know, however, that our lists to date of either biological or environmental mechanisms are not exhaustive.

These distributions were both slightly positively skewed. This resulted from the platykurtosis of the overall distributions (the "broader shoulders"), but these smaller distributions remained platykurtic as well. Males in these distributions were less variable than females in SMS32 and more variable in SMS47, with variance ratios of 0.981 and 1.066, respectively. This was because the male distribution in SMS47 was positively skewed, whereas the female distribution was not. In SMS32, both male and female distributions were positively skewed, though the male distribution was somewhat more so than the female.

The data from our model of the overall distribution of general intelligence as a mixture of two reasonably normal distributions thus acts only to corroborate the observation that greater male variability in general intelligence is not limited to the low end of general intelligence. This conclusion rests, however, on some specific assumptions about the validity of the MHT that was administered in the SMS to measure general intelligence. It is important to lay out these assumptions to understand the implications they may have had on the results that we are presenting. All tests rely on these assumptions; they are not unique to the MHT. We turn now to these issues.

MEASUREMENT ISSUES IN OUR USE OF THE MHT

To use the MHT to model the distribution of general intelligence as we have, we, of course, have to believe that it is a good test and that it actually does measure general intelligence relatively well, in terms of both test administration and test construction. One issue with respect to test administration is that the SMSs were not administered blind to sex, which may have introduced subtle measurement biases in unknown ways. But this is a problem with the administration of all mental ability tests, and we have no reason to suspect that there is any unique administration difficulty with the SMSs in this regard. The usual way to show that a mental ability test measures general intelligence well is to establish that it is appropriately correlated with other tests that are purported to measure general intelligence. This was accomplished for the MHT in the SMSs by administering the Stanford Version of the Binet–Simon scale to a randomized subset of 1,000 of the SMS32 participants and the Terman–Merrill Revision of the Stanford–Binet to a randomized subset of 1,215 of the SMS47 participants. Correlations between the MHT and the other tests were all in the range of .78 to .81 for both males and females. These correlations are not much lower than the test–retest validities for any of these kinds of tests, so it is very appropriate to consider that the MHT as used in the SMSs does measure general intelligence quite well, in the sense in which we use the term.

This is not enough, however, to validate fully the analyses we carried out. First, our detailed analyses of the distribution of MHT scores relied on the assumption that the metric on which the test measures is linear; that is, that an additional point scored represents an equivalent additional amount of general intelligence, no matter where along the range from 0 to 76 it occurs. There is no way to test this assumption because we have no gold-standard, absolute scale on which to measure general intelligence and all general intelligence tests to which we might compare the MHT also rely on this untested assumption. If we were certain that general intelligence followed a normal

distribution, we could use violations of normality to test for linearity of measurement, but of course that is exactly the property we wanted to test. If we had the item scores, we could use difficulty parameters and test information functions from item-response theory to estimate the extent to which the difficulty parameters are evenly spaced and to which the items are similarly discriminating, but this information was not retained and only the full scores are available. To the extent that the test metric is nonlinear, the actual deviations from the normal distribution could be very different from our estimates here.

We can, however, develop some information regarding this assumption. A well-designed test with linearity of measurement across the full range of general intelligence will have neither floor nor ceiling. That is, both perfect scores and scores of 0 should be very rare at best, so that it is reasonable to conclude that, throughout the range of general intelligence, any two individuals with the same score have the same general intelligence, within the measurement accuracy of the test. There were no perfect scores in either SMS survey, so the test apparently effectively had no ceiling. It did, however, have a floor: There were many individuals in each survey who received scores of 0 and it is very likely that there was considerable variation in these individuals' actual general intelligence beyond what might have distinguished them due to the small differences in their ages. There is no question that this floor contributed to the distributional deviations from the normal that we observed in both surveys. Without this floor, however, the distributions would probably have been even more negatively skewed. Though overall kurtosis might have changed considerably, the broad-shouldered features that drove some of our observations about the relative variability of males and females would still have been present, and there were more many more male than female scores of 0 in both surveys. Moreover, we were particularly interested in sex differences in variability at the high end of general intelligence, and these data would not have been affected at all if the measurement floor had not been present.

Second, our distributional analyses relied on the assumption that the MHT was, at time of testing, measurement invariant with respect to sex. That is, we assumed that the test measured the same construct of general intelligence in the same way in males and females. Again, this assumption can be rigorously tested if the item scores are available, but testing is impossible without them. It is rare for general intelligence tests to show full measurement invariance with respect to sex (e.g., Dolan et al., 2006; Ilai & Willerman, 1989; Johnson & Bouchard, 2007b; Lim, 1994; Mackintosh & Bennett, 2005; Reynolds, Keith, Ridley, & Patel, 2008; van der Sluis, Derom, et al., 2008; van der Sluis, Posthuma, et al., 2006), but the extent to which violations of measurement invariance impact measurement of general intelligence as tapped by general intelligence test scores is very controversial, and any potential impact on our analysis of the distribution of scores in the population is unclear. Again, though, some limited relevant information is available. Anastasi (1958) pointed out that of the male and female SMS 47 participants who earned equivalent MHT scores and also completed the Terman–Merrill Stanford–Binet (TMSB), the males had higher TMSB scores than did the females. This does not necessarily indicate a failure of measurement invariance in the MHT, but it does indicate that the MHT was measuring general intelligence in males and females in a systematically different way than was the TMSB. The question is relative rather than absolute as we do not have a gold standard measure of general intelligence.

It is thus relevant to consider observations made about the standardization process of the TMSB in evaluating the validity of the MHT. In the years after this revision of the Stanford–Binet was developed, several critical reports were published. In a special report in *Psychological Bulletin*, Garrett (1943) noted that the standardization sample did not adequately represent children with parents from lower occupational groups. He also reported that the verbal scale correlated more highly with the full-scale IQ than did the performance scale (.80 vs. .65) and that the memory scale was highly correlated in the younger ages but that it then rapidly dropped off with increasing age. R. Cole (1948) presented evidence that many of the items in this revision of the test did not appear to be valid in Britain. His validity criterion was teachers' ratings of general intelligence, which only serves to emphasize the point here that validity is itself a far from absolute standard. It does appear reasonable, however, to conclude that the systematically different scoring for males and females of the MHT from the TMSB is not a strong reason to suspect the measurement invariance across sex of the MHT.

Though we cannot be sure that the MHT meets the assumptions that underlie our examination of the distributional properties of the scores, it is reasonable to conclude that the MHT meets these assumptions about as well as any test would. Given that the SMSs effectively surveyed their respective populations, the combination of the breadths of the surveys and the absence of any clear weaknesses in the test renders our analysis as strong and conclusive as we can probably realistically expect. Like any analysis of this kind, its findings are rooted in the times, places, and circumstances of the populations tested. Nonetheless, it provides strong evidence for important deviations from normality in the distribution of general intelligence and for greater male than female variability in general intelligence. This greater variability is displayed both at the high and low ends of the distribution.

A CAVEAT

All the population-level data we have presented here, nearly unique though they are, are limited in that they come from children in early adolescence, at ages 11–12, before cognitive maturity. Lynn (1999) has suggested that females develop neurologically more rapidly than males, and prior to age 16, have achieved a higher proportion of the cognitive ability they will

manifest in adulthood. Because of this, he proposes, there is no sex difference in general intelligence in childhood, but a sex difference favoring males emerges after age 16. The evidence for this hypothesis is mixed (e.g., Dolan et al., 2006; Ilai & Willerman, 1989; Johnson & Bouchard, 2007a; Lim, 1994; Mackintosh & Bennett, 2005; Mau & Lynn, 2001; Reynolds et al., 2008; van der Sluis, Derom, et al., 2008; van der Sluis, Posthuma, et al., 2006), but the hypothesis is plausible. If it were true, the data we present about sex differences in mean and variability could not be considered generally applicable across the lifespan, though greater male than female variability has also been observed in less comprehensive samples of children (Arden & Plomin, 2006) and adults (e.g., Hedges & Nowell, 1995; Strand, Deary, & Smith, 2006). At the same time, even if the hypothesis that there is no mean sex difference in childhood but that there is a sex difference beginning at age 16 were true, the attribution to hardwired sex differences in neurological development might not be. It would remain possible that the mean sex difference develops because of different socialization patterns as males and females move from childhood into their adult sex roles. Were this the case, our analysis using data at age 11 would be more indicative of biological potential than an analogous analysis using young adult data.

CONCLUSIONS AND FUTURE DIRECTIONS

In this article, we reviewed the history of the hypothesis that general intelligence is more biologically variable in males than in females and presented data from two samples consisting of almost entire populations that test this hypothesis. These data, which in many ways are the most complete that have ever been compiled, substantially support the hypothesis. In addition, they suggest that the population distribution of general intelligence shows substantial deviations from the normal and that it is better conceptualized as a mixture of two normal distributions: one reflecting those with genetic and environmental syndromal conditions that disrupt general intelligence, and one reflecting those without such conditions. Males were more heavily represented than females in the distributions representing those with conditions that disrupt general intelligence. Thus, there was greater variability among males than among females at the low ends of the overall distributions of general intelligence. The differences in variability were smaller, but there was also greater variability among males at the high ends of the overall distributions of general intelligence. The departures from normality that we observed in the population distributions of general intelligence suggest that the assumptions underlying the norming procedures for many general intelligence tests should be reviewed.

Though it is not possible to measure this exactly due to a lack of consistent measurement scales, it appears that the sex difference in variability in general intelligence that we observed would not account for the sex differences in participation at the highest levels of mathematics and science occupational performance. For example, even at the highest levels of general intelligence in the SMS data, the ratios of males to females were only about 2:1. Halpern et al. (2007), however, reported male–female ratios ranging from 6.9:1 to 14.4:1 for tenure-track faculty in elite universities in physical sciences, mathematics, and engineering. Thus, as Lubinski and colleagues have reported for the Study of Mathematically Precocious Youth (e.g., Lubinski & Benbow, 1992, 2006, 2007), sex differences in career motivation and occupational interest likely contribute, as may vulnerability to identity threat from situational cues (Murphy, Steele, & Gross, 2007) and sex differences in self-confidence (e.g., Deaux, 1976; Heatherington et al., 1993). All of these, of course, possibly have roots in biological differences, differences in socialization experiences, and differences in the personal and professional trade-offs required to maintain high-level careers in math and science fields (Halpern et al., 2007). Covert sex discrimination, as reflected in studies such as those that compare ratings of work products or resumes when labeled with male or female names (e.g., Bowen, Swim, & Jacobs, 2000; Davison & Burke, 2000; Swim, Borgida, Muruyama, & Myers, 1989), also likely continues to play some role. This suggests that, to the best of our current understanding of the laws of nature, there remains plenty of room for discussion of the actualization of values.

Acknowledgments—Wendy Johnson holds a Research Council of the United Kingdom Fellowship. She and Ian Deary are members of the Medical Research Council Centre for Cognitive Ageing and Cognitive Epidemiology. The University of Edinburgh is a charitable body, registered in Scotland, with registration number SC005336.

REFERENCES

Anastasi, A. (1958). *Differential psychology* (3rd. ed.). New York: Macmillan Publishing.

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences*, *41*, 39–48.

Bowen, C., Swim, J.K., & Jacobs, R.R. (2000). Evaluating gender biases on actual job performance of real people: A meta-analysis. *Journal of Applied Social Psychology*, *30*, 2194–2215.

Burt, C. (1957). The distribution of intelligence. *General Psychology*, *48*, 161–175.

Burt, C. (1963). Is intelligence normally distributed? *British Journal of Statistical Psychology*, *16*, 175–190.

Carroll, J.B. (1993). *A survey of human cognitive abilities*. New York: Cambridge University Press.

Cole, N.S. (1997). *ETS gender study: How females and males perform in educational settings*. Princeton, NJ: Educational Testing Service.

Cole, R. (1948). An item analysis of the Terman-Merill revision of the Stanford-Binet Tests. *British Journal of Psychology*, *1*, 137–151.

Darwin, C. (1897). *The descent of man and selection in relation to sex* (2nd ed.). New York: Appleton.

Davison, H.K., & Burke, M.J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*, 225–248.

Deary, I.J., Irwing, P., Der, G., & Bates, T.C. (2007). Brother-sister differences in the *g* factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence*, *35*, 451–456.

Deary, I.J., Whalley, L.J., Lemmon, H., Crawford, J.R., & Starr, J.M. (2000). The stability of individual differences in mental ability from childhood to old age: Following up the Scottish Mental Survey of 1932. *Intelligence*, *28*, 49–55.

Deaux, K. (1976). *The behavior of women and men*. Monterey, CA: Brooks/Cole Publishing.

Dolan, C.V., Colom, R., Abad, F.J., Wicherts, J.M., Hessen, D.J., & ver der Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, *34*, 193–210.

Ellis, H. (1894). *Man and woman: A study of human sexual characters*. London: Walter Scott.

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

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross cultural perspective. *Sex Roles: A Journal of Research*, *30*, 81–92.

Feingold, A. (1995). The additive effects of differences in central tendency and variability are important in comparisons between groups. *American Psychologist*, *50*, 5–13.

Fraiser, G.W. (1919). A comparative study of the variability of boys and girls. *Journal of Applied Psychology*, *3*, 151–155.

Galton, F. (1952). *Hereditary genius: An inquiry into its laws and consequences* (2nd ed.). New York: Horizon Press. (Original work published 1869)

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences* (2nd ed.). London: Fontana.

Garrett, H.E. (1943). The standardization of the Terman-Merrill revision of the Stanford-Binet Scale: A special review. *Psychological Bulletin*, *40*, 194–201.

Geddes, P., & Thomson, J.A. (1890). *The evolution of sex*. New York: Scribner.

Guilford, J.P. (1985). A 60-year perspective on psychological measurement. *Applied Psychological Measurement*, *9*, 341–349.

Halpern, D.F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Erlbaum.

Halpern, D.F., Benbow, C.P., Geary, D.C., Gur, R.C., Hyde, J.S., & Gernsbacher, M.A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51.

Heatherington, L., Daubman, K.A., Bates, C., Ahn, A., Brown, H., & Preston, C. (1993). Two investigations of female modesty in achievement situations. *Sex Roles*, *29*, 739–754.

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

Hollingworth, L.S. (1914). Variability as related to sex differences in achievement. *American Journal of Sociology*, *19*, 510–530.

Horn, J.L., & Knapp, J.R. (1973). On the subjective character of the empirical base of Guilford's structure of intellect model. *Psychological Bulletin*, *80*, 195–196.

Horn, J.L., & Knapp, J.R. (1974). Thirty wrongs do not make a right. *Psychological Bulletin*, *81*, 502–504.

Humphreys, L.G. (1988). Sex-differences in variability may be more important than sex differences in means. *Behavioral and Brain Sciences*, *11*, 195–196.

Hunt, E.B. (2001). Multiple views of multiple intelligence. *Contemporary Psychology*, *46*, 5–7.

Ilai, D., & Willerman, L. (1989). Sex differences in WAIS-R item performance. *Intelligence*, *13*, 225–234.

Jensen, A.R. (1973). *Educational differences*. London: Methuen.

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

Johnson, W., & Bouchard, T.J. Jr. (2005). The structure of human intelligence: It's verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *35*, 393–416.

Johnson, W., & Bouchard, T.J. Jr. (2007a). Sex differences in mental abilities: *g* masks the differences on which they lie. *Intelligence*, *35*, 23–39.

Lim, T.K. (1994). Gender-related differences in intelligence: Application of confirmatory factor analysis. *Intelligence*, *19*, 179–192.

Lohman, D., & Lakin, J. (in press). Consistencies in sex differences on the Cognitive Abilities Test across countries, grades, and cohorts. *British Journal of Educational Psychology*.

Lubinski, D., & Benbow, C.P. (1992). Gender differences in abilities and preferences among the gifted: Implications for the math/science pipeline. *Current Directions in Psychological Science*, *1*, 61–66.

Lubinski, D., & Benbow, C.P. (1995). An opportunity for empiricism: Review of Gardner's *Multiple Intelligences*: The Theory in Practice. *Contemporary Psychology*, *46*, 935–938.

Lubinski, D., & Benbow, C.P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, *1*, 316–345.

Lubinski, D., & Benbow, C.P. (2007). Sex differences in personal attributes for the development of math-science expertise. In S.J. Ceci & W.M. Williams. *Why aren't more women in science* (pp. 79–100). Washington, DC: American Psychological Association.

Lynn, R. (1999). Sex differences in intelligence and brain size: A developmental theory. *Intelligence*, *27*, 1–12.

Maccoby, E.E., & Jacklin, C.N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Mackintosh, N.J., & Bennett, E.S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, *33*, 663–674.

Mau, W., & Lynn, R. (2001). Gender differences on the Scholastic Aptitude Test, the American College Test, and college grades. *Educational Psychology*, *21*, 133–136.

McNemar, Q., & Terman, L.M. (1936). Sex differences in variational tendency. *Genetic Psychology Monographs*, *18*, 1–66.

Messick, S. (1976). *Individuality in learning*. San Francisco: Jossey-Bass.

Murphy, M.C., Steele, C.M., & Gross, J.J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*, 879–885.

Pearson, K. (1897). *The chances of death and other studies in evolution*. London: Arnold.

Raymond, F. (2006). X-linked mental retardation: A clinical guide. *Journal of Medical Genetics*, *43*, 193–200.

Reynolds, M.R., Keith, T.Z., Ridley, K.P., & Patel, P.G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, *36*, 236–260.

Roberts, J.A.F. (1945). On the difference between the sexes in dispersion of intelligence. *British Medical Journal*, *1*, 127–138.

Robinson, N.M., Zigler, E., & Gallagher, J.J. (2000). Two tails of the normal curve: Similarities and differences in the study of mental retardation and giftedness. *American Psychologist*, *55*, 1413–1424.

Roeleveld, N., Zeilhuis, G.A., & Gabreels, F. (1997). The prevalence of mental retardation: A critical review of recent literature. *Developmental Medicine and Child Neurology*, *39*, 125–132.

Scottish Council for Research in Education (1933). *The intelligence of Scottish schoolchildren: A national survey of an age group*. London: London University Press.

Scottish Council for Research in Education (1949). *The trend of Scottish intelligence*. London: University of London Press.

Shields, S.A. (1982). The variability hypothesis: The history of a biological model of sex-differences in intelligence. *Signs*, *7*, 769–797.

Simonoff, E., Pickles, A., Chadwick, O., Gringas, P., Wood, N., Higgins, S., et al. (2006). The Coryden Assessment of Learning study: Prevalence and educational identification of mild mental retardation. *Journal of Child Psychology and Psychiatry*, *47*, 828–839.

Snow, R.E., Corno, L., & Jackson, D. III. (1996). Individual differences in affective and conative functions. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243–310). New York: MacMillan.

Snow, R.E., & Lohman, D.F. (1989). Implication of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–332). New York: Macmillan.

Strand, S., Deary, I.J., & Smith, P. (2006). Sex differences in Cognitive Abilities Test scores: A UK national picture. *British Journal of Educational Psychology*, *76*, 463–480.

Swim, J., Borgida, E., Muruyama, G., & Myers, D.G. (1989). Joan McKay vs. John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin*, *105*, 409–429.

Thompson, H.B. (1903). *The mental traits of sex*. Chicago: University of Chicago Press.

Thorndike, E.L. (1906). Sex in education. *Bookman*, *23*, 211–214.

Van der Sluis, S., Derom, C., Theiry, E., Bartels, M., Polderman, T.J.C., Verhulst, F.C., et al. (2008). Sex differences on the WISC-R in Belgium and the Netherlands. *Intelligence*, *36*, 48–67.

van der Sluis, S., Posthuma, D., Dolan, C.V., de Geus, E.J.C., Colom, R., & Boomsma, D.I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence*, *34*, 273–289.

Vernon, P.E. (1964). *The structure of human abilities*. London: Methuen.

Woody, T. (1929). *A history of women's education in the United States* (Vol. 2). New York: Science Press.