Constructing difference: assessment in mathematics education[1]

Dylan Wiliam, King's College London

In L. Burton (Ed.) *Which way social justice in mathematics education?*
(pp. 189-207). Westport, CT: Praeger Press, 2003.

Introduction

Since the pioneering work of Maccoby and Jacklin (1974), there has been a great deal of research on sex differences in mathematics. Most of this research has treated these differences as substantial, as more or less real and in some way inherent, innate, or even inherited and has either focused on measuring the size of the difference, or on exploring possible causes such as differences in brain physiology, child-rearing practices, or biases in processes of schooling.

This chapter will challenge the prevailing view. In the first section, I will show that differences between males and females are in fact extremely small—of an order that most statisticians would regard as negligible— and in the second section argue that these differences are not in any sense 'natural' but rather are constructed through the social processes that have shaped what we choose to call mathematics. Since the results of assessments in mathematics have profound consequences for the lives of students, adopting a particular definition of mathematics is therefore a moral, as much as an epistemological enterprise, which is discussed in the third section, and this is examined through the key concept of validity in the fourth section. In the final section, I develop a principle for the use of assessments in selection, based on the idea that if we cannot make selections fairly, we should make them at random.

The (mis-) use of statistical significance testing

If we test a sample of students, and find that the average scores of the males is higher than that of the females, the first question that arises is can we take this difference seriously? After all, it could be that the achievement of males and females in the whole population is equal but that the males who were selected for inclusion in the sample just happened, by chance, to be of higher attainment than the girls. By using straightforward principles of statistics, we can calculate the probability that a specified difference between males and females emerges just through the vagaries of random sampling (this, of course, requires that our sample was, indeed, random, which, as we shall see below, is often not the case).

Suppose we have a mathematics test on which we know that the scores for a population of students range from around 20 to 80, with an average of 50[2]. If we choose 15 students from this population at random then the average score of these 15 students is unlikely to be exactly 50. Sometimes we will get an average above 50 and sometimes we will get less than 50. If we repeated this process of drawing samples of 15 students at random from the population, then approximately 10% of the time, the average score of the sample of 15 students will be over 55, and, conversely, 10% of the time, the 15 students will average less than 45 on this test. This is why that if we tested a class of 15 males and 15 females from this population, and found that the males averaged 5 points higher than the females, it would not be safe to conclude that the males in the population were actually performing higher than the females. It could be simply that the 15 males who were selected to 'represent' the males in the population were not, in fact, representative.

For this reason, we use statistical significance testing to tell us how likely a given result is to arise by chance. If the males outscored the females by 5 marks, we can calculate that such a result (ie a difference of 5 marks either way between males and females) could have arisen by chance with a truly random sample with a probability of around 20%. We would almost certainly attribute this to random variation, and conclude that we had no grounds for deciding that males in the population have higher attainment than the females. If, however, the 15 males in our sample out-scored the females by 13 marks, such a result would arise by chance just one time in a thousand in a population where the average scores of males and females were exactly equal. In this case, we would probably decide that it was not reasonable to hang on to the idea that the males' performance across the whole population was the same as that of the females. In the language of statistical significance testing, we would reject the null hypothesis (ie that the males and the females in the population had equal scores, on average) and instead decide that, on the basis of our sampling, that the males in the population from which the sample was drawn did have higher scores than the females. However it is important to realise that we have not established that the males in the *population* outscore the females by 13 marks. All we have shown is that it is highly unlikely (in fact one in a thousand unlikely) that the scores of the males and the females in the population are equal. In fact statistical significance tells us nothing about the size of the difference between males and females—only that it is likely there is a difference. This is because statistical significance testing mixes up two quite different elements: the size of the differences between groups, and the faith we can have in the conclusion that the difference is real, as opposed to an artefact of our sampling.

Where we draw the line that marks the boundary between acceptance and rejection—effectively equivalent to determining the strength of evidence we require—however, is not clear. We would probably not regard a result that could occur with a probability of 20% as evidence that our null hypothesis had to be rejected, but would do so if our observed result only occurred one time on a thousand. The value generally used is 5% but this is quite arbitrary. In some cases, it might be justifiable to use a higher figure, such as 10%, and in many other cases, it would be appropriate to require a much heavier burden of proof, for example by setting our cut-off at 1% or even lower.

As we saw, a difference of 5 marks is not significant across a group of 30 students, but, using the conventional 5% cut-off, the same 5-mark difference *would* be significant across a group of 100 students. In fact, if we had a group of 10,000 students, with equal numbers of males and females, then a difference between males and females of just one half of a mark would be statistically significant. In other words, with such a large sample we can be sure that the difference between males and females is unlikely (ie less that 5% likely) to have arisen by chance. The difference between males and females would therefore be *statistically significant*, because we can be reasonably sure that it is really there, and not just some fluke in our sample, but it is not a *large* difference.

Much early research on sex-differences in mathematics (and, in my view, too much recent research) has used statistical significance as the primary tool for investigating the size of differences in performance between males and females. Instead, we need to focus not on the level of statistical significance, but on the *size* of the differences.

Of course we cannot use differences in marks or scores to indicate the size of differences, because as Willingham and Cole (1997) note, a difference of 1 mark between males and females on (say) the five-point scale of the Advanced Placement (AP) examination is a far greater difference than a one mark difference on the 200 to 800 scale used in the Educational Testing Service's SAT. The solution is to divide the difference in scores between males and females by a measure of how spread out are the marks for the whole group (the rationale for this is that the measure of spread provides some sort of a standardised 'measuring stick' for the data). The most generally used measure of spread used is the standard deviation of the scores. The resulting quantity is variously called the standardised effect size or the standard difference, and is generally denoted d or D. Conventionally, positive values of D denote differences where females outscore males.

Willingham and Cole (1997) calculated the standard sex difference for 75 national surveys of students undertaken in the United States at the end of schooling (grade 12), and a summary of the data is given in table 1.

How these standard differences should be interpreted is a matter of some controversy. Cohen (1988) proposed that absolute values of D (ie ignoring plus and minus signs) from 0.20 to 0.49 be classed as small, 0.50 to 0.79 as medium and those of 0.80 or more as large. Glass, McGaw and Smith (1981) however suggest that there can be no hard-and-fast rules about such interpretation, and Abelson (1995) points out that apparently small effects can be very noteworthy if produced by small interventions, or in areas where previous research has failed to find any effect at all.

As might be expected, the research studies reviewed by Willingham and Cole indicate clear differences in favour of females in language and language-based subjects, and differences in favour of males in natural science, spatial skills and mechanical comprehension. In mathematics, males outperform females in mathematical concepts, but surprisingly perhaps, females outperform males in computation. However, it is important to note two cautions in interpreting these data.

Type	Number	Mean	Range	H-spread[3]
Verbal: writing	2	+0.57	+0.54 to +0.59	-
Verbal: language use	6	+0.43	+0.33 to +0.53	0.11
Verbal: reading	10	+0.20	+0.00 to +.46	0.15
Verbal: vocabulary/reasoning	8	+0.06	-0.07 to +0.21	0.08
Mathematics: computation	2	+0.18	+0.17 to +0.19	-
Mathematics: concepts	14	-0.11	-0.28 to +0.15	0.16
Natural science	7	-0.17	-0.36 to +0.17	0.16
Social science	3	+0.02	-0.15 to +0.20	0.16

Geopolitical	5	+0.20	-0.43 to -0.12	0.12
Study skills	4	+0.20	+0.06 to +0.30	0.14
Perceptual processing	4	+0.31	+0.23 to +0.42	0.12
Spatial skills	2	-0.14	-0.25 to -0.03	-
Short-term memory	2	+0.23	+0.18 to +0.27	-
Abstract reasoning	2	+0.10	-0.03 to +0.23	-
Mechanical/electronics	4	-0.93	-1.25 to -0.78	0.18
Total	75			

Table 1: values of standard difference found in 75 grade 12 tests in the United States

First, the effect sizes are very small—as noted above, effect sizes below 0.20 are regarded as negligible by many authors—and the differences that are found between males and females are much smaller than the variation within males and within females. One way to look at this issue is to consider the 'overlap' between the populations. If it were the case that all females outscored all males, then the overlap would be 0% and if the standard difference were zero, the overlap would be 100%. Table 2 shows how the overlap in populations varies with the size of the standard difference, D. A standard difference of +0.2 means that 92% of students are in the region of overlap and in fact the proportion of the data in the overlapping region does not fall below 50% until D reaches a value of 1.4. A graph of the overlap for values of D up to 4.0 is shown in figure 1.

D	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Overlap	100%	96%	92%	88%	84%	80%	76%	73%	69%	65%	62%

Table 2: variation of overlap with standard difference (D)

Figure 1: variation of overlap with standard difference (D)

Second, the range of values of D in the studies is large, even within each of the 15 categories—for example, in the 14 studies on mathematics concepts, the average effect was -0.11 (ie in favour of males), but the values of D ranged from -0.28 to +0.15. This variation is puzzling, because if there were a difference between males and females in mathematics, the estimates of the size of these differences should be consistent. There are three possible sources of the variation in effect sizes.

The first is differences in the sample—the students who elect to take a particular test might not be representative of the population they are meant to represent. The obvious example here would be the Armed Services Vocational Aptitude Battery which is designed to provide vocational guidance for Armed Services Personnel. While it is taken by both males and females, far more males took the test, and for this reason the ASVAB data used by Willingham and Cole were adjusted to be representative of a national sample of young adults. However such weighting can do nothing about the fact that the females taking this test are not likely to be as representative of the whole female population as is the case for males.

The second source of difference is that of differences between populations. Even if a sample of students is representative of a population, such as the cohort of 18-year-olds in school, it might be that the population of males differs systematically from that of females (eg if more females stay on at school) . Willingham and Cole, however, that the data they analysed were based on representative samples and populations, so we are led to the conclusion that it is a third source of difference—the definition of mathematics employed in the construction of the test (what Willingham and Cole term ‘construct differences’)—that is the most important determinant of the size (and even the direction) of any sex differences.

It would be tempting to conclude that it was the type of mathematical concepts that were being tested that was causing this difference, but in fact the most extreme values of D found in the ‘mathematics: concepts’ category are for two tests that would appear to be measuring the same thing: the Armed Services Vocational Aptitude Battery(ASVAB) test of Arithmetical reasoning (D=-0.28) and the Differential Aptitude Test (DAT) test of Numerical reasoning (D=+0.15). We are led to the conclusion that the size of the sex-differences found on mathematics tests tells us more about the kind of mathematics being tested than it does about the respective capabilities of males and females.

While the overlap between males and females is considerable even for quite large values of D, it is important to bear in mind that this will not be the case at the extremes. If we look only at the top so-many percent of the population, quite small differences between males and females become very pronounced. Table 3 shows the variation in the proportion of males in the top 50, 20, 10, 5, 1 and 0.1 percent of the population according to the value of the standardised difference, D. For example, as we saw above in table 1, a standardised difference of -0.8 means that the overlap between males and females is 69% of the two populations. If we look only at the top 50% of the combined population of males and females, then we can see from table 3 that 65.5% will be male, and so 34.5% will be female—in other words, there are almost 2 males for every female. If we look, however, at the top 1%, of the same population, we can see that 90.6% will be male and only 9.4 percent will be female, resulting in nearly ten males for every female.

	Proportion of population					
D	**top 50%**	**top 20%**	**top 10%**	**top 5%**	**top 1%**	**top 0.1%**
-0.0	50.0	50.0	50.0	50.0	50.0	50.0
-0.1	52.0	53.5	54.4	55.2	56.6	58.4
-0.2	54.0	57.0	58.7	60.2	63.1	66.3
-0.3	56.0	60.4	63.0	65.1	69.2	73.6
-0.4	57.9	63.8	67.1	69.8	74.7	79.8
-0.5	59.9	67.1	71.0	74.2	79.7	84.9
-0.6	61.8	70.3	74.7	78.2	84.0	89.1
-0.7	63.7	73.3	78.2	81.9	87.6	92.2
-0.8	65.5	76.3	81.4	85.1	90.6	94.6
-0.9	67.4	79.0	84.3	87.9	92.9	96.2
-1.0	69.2	81.6	86.9	90.4	95.0	97.5

Table 3: variation in the proportion of males in selected proportions of population, according to standardised difference (D)

The situation is further complicated by evidence that there may be differences in the variability of the performance of males and females. If the scores of females are less variable than the scores of males, then this will tend to increase the proportion of males in the highest-achieving segments of the population even if there is no overall difference in the means of males and females. The effect of this can be seen clearly in table 4, which reproduces the same analysis as that in table 3, but based on a population in which the variability (measured as standard deviation) of females is 90% that of males —this was the figure found by Cleary(1992) on “almost all tests” (p54), although the National Assessment of Educational Progress (NAEP) found that the standard deviation of female scores was around 95% of those for males in mathematics. In contrast with table 3, the top 1% of the population contains approximately two males for every female even when there is no overall sex difference, and when D is -0.8 the top 1% of the combined population now contains approximately 27 males for every female.

	Proportion of population					
D	**top 50%**	**top 20%**	**top 10%**	**top 5%**	**top 1%**	**top 0.1%**
-0.0	50.0	53.1	55.9	58.9	66.0	75.5
-0.1	52.1	56.7	60.6	64.2	72.4	81.9
-0.2	54.2	60.5	65.0	69.3	77.9	87.1
-0.3	56.3	64.0	69.2	74.0	82.8	90.9
-0.4	58.3	67.5	73.4	78.3	86.9	93.9
-0.5	60.4	70.9	77.2	82.3	90.2	95.9
-0.6	62.4	74.1	80.7	85.7	92.9	97.4

-0.7	64.4	77.2	83.9	88.6	94.9	98.3
-0.8	66.3	80.1	86.8	91.2	96.4	98.9
-0.9	68.2	82.8	89.3	93.2	97.5	99.3
-1.0	70.1	85.3	91.4	94.9	98.3	99.6

Table 4: variation in the proportion of males in selected proportions of population, according to standardised difference (D) when female variance is 90% that of males

Whether the performance of males really is more variable than that of females is far from clear. It is certainly the case that on some tests (eg the Iowa Test of Basic Skills) the scores of males are more variable than those of females. However, there are many tests that have found little or no evidence of greater male variability. It therefore seems that, just like the sex differences discussed above, differences in variability are the result of differences in construct definition. Define mathematics one way, and you get males outscoring females at the mean, and with greater variability in their scores, so that the more selective the sample being taken, the more extreme male superiority appears to be. However, define mathematics in another way, and you will find no difference in either the means or the variability of male and female scores.

Causal explanations?

Although the differences between male and female performance in mathematics are small, they are nonetheless widespread, and so females are systematically disadvantaged where particular levels of performance or competence are required. What is rather surprising is that while many practitioners have been concerned over the last thirty or so years to do something about these differences, it appears that academics have been more concerned with finding their causes. While at first sight, this might seem sensible, in fact little progress has been made in understanding what causes these differences.

Perhaps nowhere else in educational research have correlations been pressed into service as causal explanations as readily as they have in the field of sex differences. It is almost as if the normal rules of caution, rigour, and parsimony are suspended, and ‘anything goes’. Dozens of more or less unsupported claims are made about the origin of such sex differences as have been found, even though as noted above, these differences appear to be more to do with the instruments used than in the nature of mathematics itself.

The earliest explanations were biological, arguing that male brains were more suitable for academic work (and particularly mathematics) than those of females. There are undoubtedly differences between males and females in how the brain is organised, particularly in the way the cognitive functions are distributed between the left and right sides of the brain (hemispherical specialisation) but because we understand so little about how the brain does what it does, these explanations cannot count as causes. For example, damage caused by stroke or trauma to particular parts of the brain have predictable consequence, and from this it is tempting to conclude that the part of the brain affected is responsible for the affected capabilities. But this does not follow. If we have trained a frog to leap in response to a particular sound, the fact that it cannot do so when we cut its legs off does not mean that the frog hears through its legs!

These biological explanations have been given a renewed lease of life by the upsurge of interest in evolutionary psychology. While it is probably unexceptionable to claim that the development of human brains, and some aspects of behaviour, have been influenced by evolutionary pressures, there are some who believe that such things as our sexual behaviour, moral behaviour and psychiatric disorders (Ridley, 1993; Stevens & Price, 1996; Wright, 1995 respectively) are all the result of evolutionary adaptation to situations pertaining hundreds of thousands of years ago, and are therefore, in some sense, ‘hard-wired’ into the brain. One of the most far-fetched explanations of this type is that superior male performance on spatial tasks is the result of evolutionary selection, whereby males, who went hunting for animals, needed to develop greater spatial capability than the females who remained behind and raised crops. In the first place, the anthropological evidence for this division of roles is very thin, but secondly it has been shown that the need to forage for edible plants, which do not move, can place greater demands on spatial processing and memory than does hunting (Eals & Silverman, 1994). In fact, the evidence on sex-differences in spatial abilities is very complex. There is a body of research that shows that tasks involving the mental rotation of three-dimensional objects show marked sex differences in favour of males. Linn & Peterson (1985) found a mean effect, D = -0.73 across a number of studies, although most of these studies were conducted on small, academically advanced, self-selected samples of students. However, females outperform males on recall of spatial arrays (McBurney, Gaulin, Devineni, & Adams, 1997), and in some cultures, such as the Inuit, women out-perform men on most spatial tasks (Berry, 1966; Berry, 1971). Most importantly, what is lacking is any kind of plausible mechanism that explains why, apart from in the polar regions, running around after wild animals might provide an evolutionary pressure to select for people who can do mental rotation of three-dimensional objects. Since we do not understand how these component skills are deployed in advanced mathematical reasoning, these 'explanations' are no more than fanciful speculation.

The strongest evidence against the idea that sex differences in mathematical abilities are largely genetic or innate, however, is their mutability. Alan Feingold examined the performance of males and females on the SAT from 1947 to 1983, and while the differences apparent in table 1 above (ie female superiority in language, male superiority in mathematics) were found , the size of the differences declined markedly over the period studied (Feingold, 1988). Hyde, Fennema and Lamon (1990) and Linn (1992) found that sex differences in a range of achievement tests had halved over the latter half of the twentieth century. A meta-analysis of 98 studies on sex differences in mathematical tasks conducted between 1974 and 1987 (Friedman, 1989) also found decreasing sex differences over time, and although the average effect favoured males, the size of the effect was very small, and a 95% confidence interval for the effect size included zero.

However, even these small differences have to be viewed with suspicion. As noted above, studies of mathematics achievement have used a range of different tasks and have found that males outperform females on some, that females outperform females on others, but that 'on average' males outperform females. However, the fact that the 'average' effect favours males just means that more of the tasks that favour males were used. In other words, the generally perceived superiority of males in mathematics is just the result of the preferences of researchers in their choice of what is to count as mathematics. We are rightly suspicious of researchers who use 'opportunity samples' of subjects in their research, because of the potential for bias. Accepting at face value, and weighting equally all research studies into sex differences in mathematics, would be to accept a poorly understood opportunity sample of *studies*.

As with the biological side of the debate, there are many possible explanations of sex differences in mathematics education. When the same child is dressed as a boy, s/he is encouraged to play with masculine toys such as tools and trucks and is left alone, whereas when dressed as a girl, s/he is encouraged to play with dolls and is spoken to frequently (Rogers, 1999 p96-97). The superiority of females with language may also, therefore, be simply the result of early experience rather than being innate. Valerie Walkerdine has suggested that the crucial influence is that of the tendency of young girls to be more compliant and obedient than young boys (Walden & Walkerdine, 1985; Walkerdine, 1988; Walkerdine & Girls and Mathematics Unit, 1989). This leads to greater success in the short term, in the rote aspects of mathematics common in the elementary school curriculum such as computation , but in the long-term leads to problems, when the ability to create one's own strategies become important, as is arguably the case for more advanced mathematics. This theme was taken up by Fennema and Carpenter (1998) in a study of students learning mathematics in grades 1, 2 and 3. They found that 42 of the 44 boys had used at least one self-invented algorithm by the end of the teaching experiment, compared with 30 of the 38 girls which just reaches statistical significance (Fisher's exact test, required because two of the expected cell counts are around 5, gives a probability of 0.039, while chi-square gives a value of 0.023). Now while this result is statistically significant, it must be borne in mind that had just three more girls invented their own algorithms and three fewer boys done so, then the proportion of boys and girls inventing their own algorithms would actually have been equal. While the statistical calculations tell us that the observed result is unlikely if there is no difference between girls and boys, this requires us to assume that the sample of boys and girls was truly random which is difficult to justify. From this, Fennema and Carpenter go on to argue that girls' adherence to taught strategies indicates a lower level of conceptual understanding. However, if, as seems to be widely accepted, that girls are more compliant than boys, then it could easily be that the greater use of teacher-taught strategies could be a matter of choice by the girls, rather than a matter of inability (Boaler, 2002).

I have explored this particular study in some depth not because it is a particularly egregious example, but because it illustrates the complexities of interpreting findings, and of disentangling the influences of the many forces at work on learners developing competence in mathematics.

Competing imperatives

The foregoing has shown that sex differences in mathematics achievement are small, diminishing, and, most importantly, socially constructed. Different definitions of mathematics lead to differences in the relative success of males and females. Even the assumed advantage that multiple-choice items confer on males depends on the kinds of items. For example, Beller and Gafni (1996) found that the male superiority over females was greater on constructed response items than multiple-choice items, although this appears to have been an artefact of the fact that the CR items were harder than the MC items, and the males performed better overall (which was, of course, in turn, an artefact of the definition of mathematics employed). Constructed-response items do generally favour females, but then it would be astounding if they did not, because of their superior linguistic abilities. Moreover, the size of the advantage that constructed-response items confer on females is highly variable. For example, in the United States, both Kansas and Kentucky have state-wide assessment programmes for grade 12 (ie 18-19 year-old) students that involve both multiple-choice and constructed-response items. In Kansas, the multiple-choice items favour males (D = -0.22) while the constructed-response items favour females (D = +0.14), so that the difference between the two formats represents a swing of 0.36 standard deviations. In Kentucky, the same pattern is found—multiple-choice items favour males and constructed-response items favour females—but the sizes of the effects are much smaller; -0.02 and +0.04 respectively. Even the size of the advantage that constructed-response items confer on females, therefore, depends on the definition of mathematics used.

All this shows that there cannot be a 'mathematics degree zero' in the sense that Barthes (1984) talked about 'writing degree zero'—there is no 'natural' choice of mathematics that we can use for our assessments. So how can we choose? In this final part of the chapter, I want to argue that all the issues discussed above can be subsumed within the key concept of validity, and that the choices about definitions of mathematics can be viewed in terms of conflicting concerns about the meanings and consequences of test results.

The validity of educational assessments

Validity was seen originally as a property of a test—a test was valid to the extent that it tested what it purported to test. However, a test may be more valid for one group than another—a particular mathematics test may involve complex language, and so would tell us little of the mathematical abilities of poor readers, but might work quite well for fluent readers. A further complication is that while a test may have been developed and administered for one purpose (eg to determine the achievement of a student), the results are often used in other ways (eg to determine how good the teacher of a particular class is). For this reason, it is widely agreed now that validity is not a property of a test, but a property of the use to which the information arising from the test is put.

Within this, of course, there are competing priorities. Professional mathematicians have a legitimate interest in determining what is to count as mathematics at school, but this influence has not always been productive. For example, at the famous Royaumont Seminar in 1959, a group of mathematicians decided that the best way to organise and teach school mathematics was with the structures that professional pure mathematicians used, such as set theory(Organisation for European Economic Co-operation Office for Scientific and Technical Personnel, 1961). While the resulting 'modern mathematics' may have been a useful platform for the ablest students, there can be little doubt that the abstract, de-contextualised axiomatic approach used was alienating for many students. It also seems likely that a disproportionate number of students alienated by a decontextualised approach would have been female (see, for example, Gilligan, 1982, Boaler, 1997), and so the move away from such approaches will have been a contributory factor in the reduction of sex-differences over recent years.

Other stake-holders will have other priorities for the curriculum. The 'industrial trainers' (Williams, 1960) will want school mathematics to have a strong vocational strand, while others may stress the increased control over one's life that mathematical competence may bring. These are concerns with the content of the assessment (what is sometimes called 'content validity'), where the primary aim of the assessment is to certify the achievement of an individual in mathematics (and so is, essentially, retrospective).

The other main use that is made of the results of mathematics assessments is prospective: using the results of one assessment to predict how well an individual will do in the future. Success in mathematics at one level of the educational system does correlate positively with success at the next level, although nowhere near as strongly as is often supposed (correlations as low as 0.3 are common). It might be supposed that choosing items that predict future performance well might offer the possibility of an objective choice of mathematical content (in other words, a 'mathematics degree zero'), but since the content of the mathematical curriculum at the next level is just as arbitrary as that at the prior level, then this is really no more than a self-fulfilling prophecy. Here, I should make clear that the term 'arbitrary' does not mean capricious but rather that the mathematics curriculum is the result of competing and conflicting forces, and that different resolutions of these conflicts would result in different curricula.

In response to this, our search for a 'mathematics degree zero' might focus not on the ability to predict success on harder *mathematical* content, but so that we can predict other, *non-mathematical*, performance. In fact, success at mathematics does predict future performance levels in a range of other activities and occupations, but there is little evidence that this is causal. Our current measures of mathematics performance tend to correlate better with measures of general intellectual aptitude than any other school subject. However, this again leads us to define mathematics in a particular way. If we include mathematical content that helps us to predict general job performance, we are likely to pick out those aspects of mathematics that are strongly related to general intellectual functioning—in which case, why not use an IQ test? Indeed, while many authors seem to regard the high correlation of mathematics scores with IQ scores as appropriate (Newton, Tymms, & Carrick, 1995), it could be argued that, in order to be really useful, mathematics assessments should *not* correlate well with IQ.

These debates can be best understood not as arguments about the technical merits of rival assessments but as arguments over the value implications of different definitions of what is to count as mathematics. Mathematics has high status as a subject precisely because it correlates so well with IQ test scores, and because it is a subject in which it is easy to create large differences in performance between individuals—see Wiliam, Bartholomew and Reay(2002) for an extended discussion of this point. It could even be argued that the pre-eminence of mathematics in the curriculum as a high-status subject arises because it is the subject with the greatest sex-differences in favour of males. After all being able to use a typewriter was a high-status occupation when typewriters were first developed and at a time when almost all typewriter operators were male, but declined in status as typewriter operation became feminized. Similarly, medicine, while still a high-status profession in the West, declined in status in the USSR as the profession became dominated by women. In the same way I would argue that the superior performance of males in mathematical subjects is not a *cause* of male supremacy in politics, employment and elsewhere, but a *consequence*. In other words, our definitions of mathematics are precisely those that keep males outperforming females . Changes in the definition of mathematics that reduce, remove or even reverse sex-differences are criticised as 'dumbing down' the mathematics curriculum, but these criticisms can equally well be viewed as laments for the loss of power (witness the arguments about the 'dumbing down' of mathematics brought about by the introduction of school-based performance assessments in the national school-leaving examination in England and Wales).

The struggles over what is to count as mathematics (in other words, its value implications) are important because assessments send messages about what is and is not important in mathematics education. The result is that teachers and students change what they do.

These four aspects of validity—content considerations, predictive utility, value implications and social consequences—have been integrated by Samuel Messick (1989) into a unified model of validity argument. The important feature of this model is that validating an assessment requires attention to all aspects:

> Unified validity does not imply answering only one overarching validity question or even several questions separately or one at a time. Rather, it implies an integration of multiple supplementary convergent and discriminant forms of evidence to answer an interdependent set of questions. (Messick, 1997 p59)

In validating an assessment, therefore, as well as asking the traditional questions about the relevance and representativeness of the content, and its utility for prediction, we must explore the value implications of adopting the assessment, and assess the likely social consequences. We must then attempt to reconcile the competing imperatives suggested by these different aspects.

For example, as we saw above, the ability to rotate three-dimensional shapes mentally shows a considerable sex-difference in favour of males. Should such items be used in mathematics tests?[4] Professional mathematicians might posit an *epistemological imperative*, arguing that omitting such items under-represents the construct of mathematical competence and such items must be included. Such items might also be favoured by those with a *utilitarian imperative*, since they are likely to discriminate strongly between candidates, and thus contribute to predictive validity. However, such items could well have a disproportionate impact on females. Including such items would lower the scores of females relative to males, and in the absence of evidence that the particular skill of mentally rotating three-dimensional shapes is relevant to future performance, even in mathematics, then a *moral imperative* would require omitting the item.

Random justice[5]

The tensions between the conflicting imperative discussed above are exacerbated when tests are required to fulfil many functions. Odd though it may seem, it is possible to ameliorate the tension by introducing a degree of randomness into the use that is made of test results. The principle is illustrated by means of an example. In almost all countries, entry into medical education is highly competitive. In most countries selection is based on scores on school examinations or some other assessment, more or less well designed to assess aptitude for medical education. Let us assume, therefore, that 10% of the applicants would be able successfully to train as a doctor, but that only 1% of the applicants can be admitted . How should we use the results of our test? In the anglophone countries, the typical approach is to accept the highest scoring 1% of the applicants on the assessment. The effect of this is to magnify any sex-differences in selection. To see why, look back at table 3. If there is a standardised mean difference (D)of 0.5 in favour (say) of males (assuming equal variances)then males will form 71%of the top 10% but just under 80% of the top 1% (and of course these trends are even more marked for larger values of D, or if the variance of the lower-scoring group is smaller, as in table 4). Restricting our selection to the top 1% has increased the proportion of males accepted for medical education, even though we have no evidence that they will do better. In contrast, in many countries in continental Europe, the selection procedure discards any 'surplus achievement' beyond the minimum threshold, and instead proceeds by randomly selecting from the 10%, whom, after all, we know could all train successfully as doctors.

In the assessment literature, it is conventional to distinguish between recruitment and selection. Recruitment is the process of identifying all suitable applicants (ie our medical example, this would mean identifying the 10% of applicants who could train successfully as doctors). Selection refers to the process of selecting from these appropriate individuals which one we will actually train as doctors. It may be that certain sub-groups (eg males, females, ethnic minorities, students with disabilities) are under-represented in the recruitment group, and we must be sure that such under-representation cannot be ameliorated by suitable adjustment of the training and working environment. For example, students requiring wheel-chair access might be denied a place at medical school not because they could not become good doctors, but because they would not be able to satisfy the particular requirements of a medical training programme. It is also necessary to ensure that any adjustments in the nature of the work for which students are being trained are made that would reduce under-representation. However, given these conditions, and once the recruitment group is well defined, we can frame the conclusion of the argument above in terms of the following principle:

the representation of sub-groups in the sample selected must be the same as that in the recruitment group.

Any selection process that increases the under-representation of any sub-group (when compared with the recruitment group) is unjust, and unjustifiable.

Another example here is that of the use of aptitude tests in selection to higher education in the US. The scores obtained on the SAT by students from some ethnic minorities is below that of whites, which is cited by some as evidence of bias in the test. But it is also the case that the same ethnic minorities do less well in higher education. However, the crucial evidence of bias is that the standardised difference in SAT scores is approximately twice that of scores in higher education, which implies that using the SAT for selection increases the under-representation of those ethnic minorities (Hakel, 1997). A fairer approach would be to identify the minimum score required on the SAT in order to succeed in higher education (using difference cut-score for different sub-groups), and if this process identifies more suitable candidates than can be admitted, making any further selection by lottery.

This seems counter-intuitive—we seem to be throwing away information. But if we do not have specific evidence that surplus achievement beyond the threshold required (ie to be counted in the recruitment group) produces greater performance, then it makes sense to disregard that evidence, even when working with an utilitarian imperative. When the importance of having doctors (to choose just one example) who adequately represent the communities they serve is factored in then the moral imperative is compelling. It is just an application of the general principle that if you can't make a selection without introducing bias (in the sense of unjustifiable difference), then random selection is the only fair way.

Conclusion

In this chapter, I have argued that differences between the scores of males and females on mathematics tasks have always been small when compared to the variability of males and of females (ie that the variation between each group is small compared to the variation within the group). The small differences have been regarded as much larger than they are primarily because of an over-reliance on statistical significance as a measure of effect size. When the empirical evidence on sex-differences in mathematics is viewed in terms of the overlap between populations, even the largest differences found in practice mean that 90% of the population is in the region of overlap, and only 1% of the variation (variance) in scores is attributable to sex. More importantly, the size, and even the direction, of these sex-differences depends on the mathematical tasks and items used. Even the 'pernicious hypothesis' (Noddings, 1992) that males are more differentiated than females can be seen to be the result of choices made about what is to count as mathematics.

Given these competing definitions of mathematics, it is tempting to try to arbitrate between them, and to attempt to locate 'mathematics degree zero', but the search is doomed to failure since there can be no such thing. Mathematics curricula are 'objects of history' (Cherryholmes, 1989, p116)—the outcomes of social contestations about what is to be included—and even the high status of mathematics in many contemporary societies may be simply a reflection of previous male dominance.

However, even if there were some absolute definition of mathematics, and even if we accepted the essentialist arguments that attempt to show that males are 'naturally' better than females at certain things, this would tell us nothing about how to use assessments of mathematics for selection, because one cannot deduce an 'ought' from an 'is'. An epistemological imperative encourages us to shape our assessments according to existing notions of what mathematics 'should' be while a utilitarian perspective would require us to look at the utility of these assessments for prediction. However, a moral imperative requires us to consider the social consequences of such actions. For example, even if it were agreed that mental rotation was an essential part of mathematics, and even if it were essential to the task for which we are selecting, we should not accept male superiority in this task. Instead we should do whatever needs to be done to mitigate its effect.

Finally, in moving from recruitment to selection, we should introduce no additional discrimination. Surplus achievement beyond the minimum required for whatever is being selected for should be used only when there is no difference in representation in sub-groups. If there is, then random selection from the recruitment group is the only fair way. Social justice sometimes requires random justice.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Barthes, R. (1984). *Writing degree zero & Elements of semiology* (Annette Lavers & Colin Smith, Trans.). London, UK: Jonathan Cape.

Beller, M. & Gafni, N. (1996). *Can item format (multiple-choice vs open-ended) account for gender differences in mathematics achievement?* Jerusalem, Israel: National Institute for Testing and Evaluation.

Berry, J. W. (1966). Temne and Eskimo perceptual skills. *International Journal of Psychology,* **1**, 207-229.

Berry, J. W. (1971). Ecological and social factors in spatial perceptual development. *Canadian Journal of Behavioral Science,* **3**, 324-336.

Boaler, J. (1997). Experiencing school mathematics: teaching styles, sex and setting. Buckingham, UK: Open University Press.

Boaler, J. (2002). Paying the price for 'sugar and spice': shifting the analytical lens in equity research. *Mathematical Thinking and Learning,* **4**.

Cherryholmes, C. H. (1989). *Power and criticism: poststructural investigations in education*. New York, NY: Teachers College Press.

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In Educational Testing Service (Ed.) *Sex equity in educational opportunity, achievement, and testing: proceedings of a 1991 ETS Invitational Conference* (pp. 51-90). Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Duxbury, N. (1999). *Random justice: on lotteries and legal decision making*. Oxford, UK: Oxford University Press.

Eals, M. & Silverman, I. (1994). The hunter-gatherer theory of spatial sex difference: proximate factors mediating the female advantage in recall of object arrays. *Ethology and Sociobiology,* **15**, 95-105.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist,* **43**, 95-103.

Fennema, E. & Carpenter, T. P. (1998). New perspectives on gender differences in mathematics: an introduction. *Educational Researcher,* **27**(5), 4-5.

Friedman, L. (1989). Mathematics and the gender gap: a meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research,* **59**(2), 185-213.

Gilligan, C. (1982). *In a different voice*. Cambridge, MA: Harvard University Press.

Glass, G. V.; McGaw, B. & Smith, M. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hakel, M. D. (Ed.) (1997). *Beyond multiple-choice: evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hyde, J. S.; Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin,* **107**, 139-155.

Linn, M. C. (1992). Gender differences in educational achievement. In Educational Testing Service (Ed.) *Sex equity in educational opportunity, achievement, and testing: proceedings of a 1991 ETS Invitational Conference* (pp. 11-50). Princeton, NJ: Educational Testing Service.

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

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

McBurney, D. H.; Gaulin, S. C. G.; Devineni, T. & Adams, C. (1997). Superior spatial memory of women: stronger evidence of the gathering hypothesis. *Evolution and Human Behaviour,* **18**, 165-174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education/Macmillan.

Messick, S. J. (Ed.) (1997). *Alternative modes of assessment, uniform standards of validity*. Mahwah, NJ: Lawrence Erlbaum Associates.

Newton, D.; Tymms, P. B. & Carrick, N. (1995). Is the GCSE fair? Examination success and cognitive style. *British Journal of Curriculum and Assessment,* **6**(1), 21-25.

Noddings, N. (1992). Variability: a pernicious hypothesis. *Review of Educational Research,* **62**(1), 85-88.

Organisation for European Economic Co-operation Office for Scientific and Technical Personnel (Ed.) (1961). *New thinking in school mathematics*. Paris, France: Organisation for European Economic Co-operation.

Ridley, M. (1993). *The red queen: sex and the evolution of human nature*. London, UK: Penguin.

Rogers, L. (1999). *Sexing the brain*. London, UK: Weidenfeld and Nicholson.

Stevens, A. & Price, J. (1996). *Evolutionary psychiatry: a new beginning*. London, UK: Routledge.

Walden, R. & Walkerdine, V. (1985). *Girls and mathematics: from primary to secondary schooling*. London: University of London Institute of Education.

Walkerdine, V. (1988). *The mastery of reason: cognitive development and the production of rationality*. London: Routledge.

Walkerdine, V. & Girls and Mathematics Unit (Eds.). (1989). *Counting girls out*. London, UK: Virago.

Wiliam, D.; Bartholomew, H. & Reay, D. (2002). Assessment, learning and identity. In R. Zevenbergen & P. Valero (Eds.), *Researching the socio-political dimensions of mathematics education: issues of power in theory and methodology* Dordrecht, Netherlands: Kluwer Academic Publishers.

Williams, R. (1960). *The long revolution*. London, UK: Chatto & Windus

Willingham, W. S. & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, R. (1995). *The moral animal: why we are the way we are*. London, UK: Little, Brown.

1 I am grateful to Jo Boaler and an anonymous reviewer for helpful comments on an earlier draft of this chapter, although of course, the responsibility for any remaining errors lies with me.

2 Specifically that the mean of the population is 50 and its standard deviation is 15.

3 The hinge-spread or H-spread of the data is the difference between lower and upper hinges of a set of data, where the hinges are the median values of the lower and upper halves respectively of the data (including the median where there are an odd number of data values) when the data are arranged in order. As the number of data increases, the H-spread approaches the value of the interquartile range, and thus provides a measure of the dispersion of the data values.

4 Note that this is not the same question as whether these items should be taught to students.

5 This phrase is taken from Duxbury(1999).