DYLAN WILIAM, HANNAH BARTHOLOMEW
AND DIANE REAY

ASSESSMENT, LEARNING AND IDENTITY

For most of the last century, educational assessment derived its principal research paradigm from psychology, and while there was some acknowledgement of the differences in emphasis between psychological and educational measurement—as they were termed—there appeared to be reasonably broad consensus that the theoretical resources developed in psychology were appropriate for addressing problems and issues within education.

Within this paradigm, the creation of tests and other forms of assessment has been regarded as an essentially technical and objective undertaking although there has, during the last quarter-century, been an increasing acceptance that educational assessments have social consequences—people change what they do because of those assessments. In particular, the work of Samuel Messick has shown that any analysis of the validity of assessments that ignores the social context in which the assessments are used is necessarily impoverished. However, most analyses have tended to regard the social consequences of the use of assessments as separable from the technical issues involved in their construction, and even the analyses of assessments that do take into account the social consequences of educational assessments have tended to look at large-scale assessments, and at large-scale effects.

In this chapter, we will argue that there are no such things as an 'objective' assessments because their design is governed by considerations of how they will be used. We then use Messick's theoretical framework for validity argument to show that the meanings of educational assessments cannot be separated from their social consequences, and because of the high-stakes contexts in which they are used, assessments frequently come to shape the constructs they are designed to measure. By drawing on empirical work we have undertaken, we then illustrate the power of educational assessments to provide both constraints on, and affordances for, the way that students develop their identities in classrooms. In this way we show that assessment, learning and identity are inextricably linked.

PSYCHOLOGICAL AND EDUCATIONAL ASSESSMENT

The development of educational and psychological assessments has, by and large, been driven by the best of motives. The first systematic assessments were apparently conducted in China, in order to regulate access to the civil service. There was a

P. Valero and R. Zevenbergen (Eds.), Researching the socio-political dimensions of mathematics education: Issues of power in theory and methodology, 1—19.

concern that entrants to the civil service were almost exclusively drawn from the ruling classes, and formal testing was introduced in order to find ways of selecting the most talented applicants from all classes of society. Similarly, in the 1930s in the USA, the president of Harvard, J.B. Conant, was concerned that Harvard's students were predominantly the sons of those who had themselves been to elite universities (Lemann, 1999) and wanted to find a way of selecting students on the basis of their abilities.

Thus, although some authors have argued that the development of tests of this kind was linked to the eugenics movement, the concerns of the two projects — eugenics and aptitude testing— were in fact diametrically opposed. The eugenics movement was predicated on an assumption that ability is inherited (see Selden, 1999, for an excellent account). With such a view, aptitude testing is unnecessary, because one can select the most talented students by reference to their parents' achievements. In contrast, the desire for a system of aptitude testing is implicitly founded on a belief that ability is only weakly inherited—if at all—and is therefore, to all intents and purposes, randomly distributed throughout the population, irrespective of ethnicity, sex and social class. The problem is, of course, that 'ability' rapidly becomes conflated with access to educational opportunities, which makes its identification within a population extraordinarily difficult. For those who wish to use an assessment to identify talented individuals, the key concern is that the assessment does, indeed, identify ability, rather than irrelevant features such as the quality of education received.

In the USA, for example, with its highly devolved education system, setting a test that would fairly assess the scholastic achievements of students across the whole country would be impossible since each school is free to set its own goals. Furthermore, a measure of scholastic achievement would be an impure measure. While success in school depends on the capability of the student and on their perseverance, it is also dependent on the quality of teaching and specifically the opportunity to learn (see Bursten, 1992). In a system where it is believed that education should be funded locally, as is the case in the USA, then students from affluent areas are likely to be at an advantage. The response, therefore, was the creation of the Scholastic Aptitude Test—now termed just SAT—which was intended to measure aptitude for higher education irrespective of the quality of schooling experienced. The extent to which such a test does this, is, of course, a question of *validity*.

VALIDITY AND RELIABILITY

In the earliest days of educational and psychological testing, the validity of an assessment was defined as the extent to which it assesses what it purports to assess (Garrett, 1937). Initially, this was simply a requirement for content validity. In other words, the test should assess a relevant and representative sample of the content of interest. This was generally investigated through the use of a panel of experts who would be asked to look at each item on the test and rate its relevance and also, then, to comment on the overall balance of the items in the test. The important point here

is that even in these earliest days, validity was an essentially social construct, depending on the consensus of a panel of experts.

However, we also want tests to be reliable. After all, here is no point in having a 'good' test—that is, one which assesses all the relevant content—if the result of an individual depends as much on chance as on her or his skills and abilities. The reliability of a test can be thought of as a kind of 'signal' to 'noise' ratio, and if we want to maximise the reliability of an assessment, we can do this either by reducing the 'noise' or by *increasing the 'signal'*.

In the context of educational assessment, 'increasing the signal' entails creating a test that maximises the differences between individuals in the same way that in communications engineering, for example, increasing the signal would correspond to maximising the potential difference between presence and absence of signal in a wire. The fact that tests tend to distribute scores across the whole mark range is therefore in no sense 'natural'. It is the result of a decision to improve the reliability of a test not through reducing the error in scores but instead attempting to mask the errors by making the differences between students as great as possible. Nor is the fact that scores on tests tend to produce a 'bell-shaped' pattern in any way natural. It is, rather, the result of decisions about the kinds of items to include. By replacing items of moderate difficulty with harder and easier items, a test with a rectangular distribution of scores can easily be produced.

This process of test construction therefore *requires* the production of tests that maximise the differences between individuals. This, in turn, requires tests to place less emphasis on what is common between students, such as their experience of schooling, because this common experience would tend to reduce the differences between students. After all, if one sets a test in which students were asked to describe the activities in which they engaged during a school day, it is likely that everyone who went to school would pass, and those that did not, would not. The fact that our educational assessments find large differences between students in what they can do is therefore not natural at all, but a direct and immediate consequence of the need for reliability.

For example, many studies have found that schools have comparatively little effect on educational achievement (e.g., Rutter, Maughan, Mortimore & Ouston, 1979). How has this been established? By measuring educational achievement with a test that was designed to maximise differences between *individuals*. Items that assess the common experiences of all students are not used because they do not discriminate, so we should not be surprised to find that the differences between students at the same school (sometimes called the within-school variance) is greater than differences between schools (between-school variance). Such tests *create*, and *reify* the constructs they purport to assess.

By this, we do not mean to suggest that constructs such as 'mathematics' are completely capricious. Clearly there are limits to how far a construct can be distorted in the search for reliable means of assessment. But we hope from the technical arguments above that it is clear that a range of social factors intrude into the design of apparently 'objective' assessment instruments.

As an illustration of how ingrained these ideas are, imagine that we wanted to measure differences in students' achievement in schools using different mathematics

curricula. Instead of using conventional tests we could modify a standard test for our purpose by excluding all the items that show little difference between schools, and adapting existing items to place greater emphasis on aspects which differ systematically from school to school. To many people, this 'feels' wrong —it feels as if we are fixing the test to give us the result we want, and that somehow there must be a 'natural' test. But there is not. There is no neutral ground on which we can stand. As Cherryholmes (1989, p. 115) remarked, 'Constructs, measurements, discourses and practices are objects of history'. The meanings that we can legitimately attach to test results are also products of their history, and assessments cannot be understood outside the social context in which they are used, and do not make sense unless the history of that social context is also understood.

Although many, if not most, tests claim to be retrospective in that they purport to indicate the extent to which a student has acquired a certain body of knowledge, they are almost always also used prospectively—for example to select individuals for employment or further educational opportunities. This means that what a test 'purports' to assess is only part of the picture. The information from the test, once in the public domain, can be used in all kinds of ways not foreseen by the constructor of the test. Therefore, if it is to be at all useful, the concept of validity cannot be a property of a test, nor even of the results of a test. To be useful, validity must be, rather, a property of the inferences made on the basis of test results. As Cronbach (1971, p. 447, emphasis in original) noted, 'One validates, not a test, but an *interpretation of data arising from a specified procedure'*.

For example, those who wish to defend the use of an aptitude test such as the SAT for admission to higher education would compare the SAT scores of individuals with their grades in college or university (typically the grades achieved at the end of the first year). A strong correlation between these two measures (sometimes called the *predictor* and the *criterion* respectively) is taken as evidence of the utility of the predictor for selection. A difficulty with this kind of study is that we are getting only half the picture. We do have evidence about how well those actually admitted to higher education did, but we know nothing about how well those not admitted would have fared had they been admitted, and to address this simply by admitting more students into higher education, suspecting that many of them would fail, would be ethically questionable.

Furthermore, because all tests of achievement are, as noted above, impure measures, the relationship between predictor and criterion is likely to be different for different groups. Yet any differences of this kind will be masked if only the overall correlation between predictor and criterion is considered. For example, in the USA, it has been found that while many minority ethnic groups score less highly than whites on a given test, for a given score on that test the grades achieved by minority ethnic students at the end of the first year of college are *higher* than those of whites. In general it appears that the differences in *criterion* scores between minority ethnic groups are only half as great as those in the *predictor* scores (Hakel, 1997). In other words, using such a predictor for selection would systematically under-represent the potential for success in minority ethnic students. It is for this reason that many of the elite universities in the USA have replaced requirements for specific test scores with the requirement that a student is in the upper quartile of their age-cohort at their

school. Of course, this approach, too, is fraught with difficulties. In general, no test can give us access to an 'untainted' evaluation of an individual's capabilities.

Nonetheless, the requirement for a predictor to correlate well with a criterion seems unexceptionable. After all, if the correlation is not good, then one can hardly use the predictor as a predictor! Some have gone so far as to say that, for predictive validity, correlation is the only thing that is important:

> the information about validity is in the correlation coefficients [...] The nature of the measurement is not what is important to this statement. The important fact being reported is that these variables can be used to predict job performance within the limits of accuracy defined by the correlation coefficient (Guion, 1974, p. 288).

This approach to validity certainly seems more objective than simply asking panels of experts to comment on the balance and appropriateness of items in a test. Yet this definition of what makes a 'good test' has further unexpected consequences.

A key requirement to achieving a high correlation between predictor and criterion is that each of the items in the test must *discriminate* well. In other words, we want to make sure that each item on the test is more likely to be answered correctly by those who are good at whatever the test is measuring than by those who are not. Of course, the problem is that we do not yet have a measure of what the test is meant to be measuring —that is, why we are developing the test— so what we do is then to see how well each item correlates with the scores obtained on the other items in the test. Items with poor correlations are then removed, thus increasing the difference between students, so increasing both the reliability, and the predictive validity of the test.

This process, is, of course, the reverse of what is usually imagined as happening. Rather than taking a well-defined domain and devising a test to assess an individual's competence on that domain, the development of the test is driven by the requirements laid down by psychometricians for what makes a good test. When these tests are used in high-stakes settings, students are coached to produce the best possible results, and so the test comes to stand for the entire subject. In this way, tests come to define what they are supposed to measure (Hanson, 1993).

While social and political factors are crucial, if rarely acknowledged, influences on the apparently objective processes of developing both educational and psychological tests, there is an important distinction between educational and psychological tests that renders the former even more open to social influences. Most psychological tests are restricted. They are, for the most part, available only to specialists, who must generally receive specific training in their administration before they can be purchased. In contrast, educational tests are widely available, and even when they are restricted, such as is the case with large-scale 'aptitude' tests such as the SAT, they are used in high-stakes contexts and as a result a large number of publications giving students practice in similar items has been produced.

It was this realisation that the quality of assessments could not be understood independently of the social situations in which they are used that prompted Messick (1980) to argue that the ethical and value concerns implicit in educational tests, and the social consequences of their use, should also be considered as part of the key

concept of validity. This was encapsulated in a model of validity as the crossing of two facets: the function, and the basis of test interpretation (see Figure 1).

	Result interpretation	Result use
Evidential basis	Construct validity **A**	Construct validity and relevance/utility **B**
Consequential basis	Value implications **C**	Social consequences **D**

Figure 1: Messick's framework for validity enquiry

The two cells in the upper row (A and B) deal with traditional conceptions of validity. The evidential basis of result interpretation (cell A), encompasses those aspects of construct validity concerned with how well the assessment represents the domain being assessed—often called content validity—while predictive and concurrent validity—often grouped together as criterion-related validity—can be regarded as aspects of the evidential basis of result *use* (cell B). Construct validity was originally used as a kind of 'leftovers' box for the validation of tests where no agreed definition of test content existed and where there was no widely agreed criterion variable against which to validate the test—a typical example would be a test of 'math anxiety'. Over the last forty years, however, there has been increasing agreement that construct validity is 'the whole of validity from a scientific point of view' (Loevinger, 1957, p. 636). Put simply, construct validation is an enquiry into the evidence supporting inferences made on the basis of assessment outcomes.

Messick's contribution was to show that construct validity focused on the *evidential* basis of result interpretation and use, and that an understanding of how tests and other assessments actually function in society requires an investigation of the *consequential* basis as well. Furthermore, these are not separated activities, since the consequential basis of result interpretation and use can affect the evidential basis. For example, those who argue for the use of multiple choice tests in the assessment of mathematical performance use evidence from empirical studies to show that while such tests cannot assess all aspects of mathematics, the correlation between scores on extended-response and multiple choice tests in mathematics is very high. Therefore, they maintain, although multiple-choice tests do not assess all aspects of mathematics, the scores from such tests can be used as a good proxy for those aspects of mathematics not tested.

However, such a claim requires that the relationship of student performance on the tested and untested material remains unchanged. While this might be true in low-stakes settings, such as for the large-scale light-sampling testing such as the United States National Assessment of Educational Progress (NAEP) and international comparison studies such as the Third International Mathematics and Science Study (TIMSS), for those tests where life-affecting consequences accrue to students or

teachers, it is likely that the tests, or more precisely, the use made of information from the tests, will change the behaviour of those involved.

The use of multiple-choice tests as the sole or predominant method of assessment sends the message that the skills assessed in such tests are the ones that really matter. In other words, the tests embody *value implications* about the subject that come to define the subject (cell C). In this way tests that have already defined what they were designed to assess reinforce the idea that this definition of the subject is the one that matters.

These value messages will then influence the actions of teachers and students to place greater emphasis on multiple-choice items, with the social consequence that the kinds of mathematics assessed in constructed-response assessments are neglected (cell D). The strong correlation between the multiple-choice and constructed-response items, which meant that one could be used as a proxy for the other, is now weakened, so that scores on multiple-choice tests are no longer a good guide to performance on other items. The social consequence of relying on a multiple-choice test does not therefore impact just the consequential basis of the assessment's validity. It also fundamentally changes what the test is measuring, and its relationship with other assessments.

The political and social dimensions are important, therefore, not just to understand 'how we got here' in terms of assessment, but also to understanding 'where we are'. The political and social dimensions are immanent in our current assessment practices, not just in what they do but in what they are.

> Belief systems concerning the individual should not be construed as inhabiting a diffuse field of 'culture', but as embodied in institutional and technical practices—through which forms of individuality are specified and governed. The history of the self should be written at this 'technological' level, in terms of the techniques and evaluations for developing, evaluating, perfecting, managing the self, the way it is rendered into words, made visible, inspected, judged and reformed. (Rose, 1989, p. 218)

Objective assessment is therefore not just difficult to achieve, but by definition impossible. Our search must therefore be not to strive to free our assessments from their subjectivities—this can never be done—but rather to understand the origins of those subjectivities, how they arose, what purposes they serve, and perhaps most importantly, who benefits and who does not.

There are several good accounts of how these issues play out on a large scale (e.g., Broadfoot, 1995; Hanson, 1993). However, there are far fewer accounts of the role that assessment plays in shaping the identities of individual students in schools. Therefore, for the remainder of this chapter, we will illustrate some of the themes raised above by reference to the role that assessment plays in shaping the day-to-day reality of classrooms.

In many countries both in the 'North' and the 'South', greater and greater emphasis is being placed on the assessment of students; they are being tested more often, and their performance has important implications not only for the students themselves, but also for their teachers and for the schools they attend. The publication of 'league tables' of school performance and the widening of parental choice combine to exert considerable pressure on schools to maximise the

performance of their students on state-mandated tests and examinations. This pressure is felt by students, and has a major impact on what happens in classrooms. In the remainder of this chapter we will draw on data from two studies to examine the ways in which students' perceptions of themselves as learners are affected by the assessments to which they are subjected. Drawing first on data from a study into the impact of ability grouping in secondary mathematics classes, we consider some of the ways in which dominant images of mathematics as remote, abstract and very difficult are reinforced by assessment procedures which have profound repercussions in schools and dominate the mathematical identities that students are able to develop. However, the impact of assessment regimes is not limited to mathematical identities. In the subsequent section, we illustrate how the pressure to perform well in the national tests for 11-year-olds in one elementary school in England had profound implications for students beyond the subjects being assessed, extending to their potential careers and even raising questions about their moral worth. In this way, assessments influence what is to be learnt, how it is to be learnt, and even what it means to be a learner. Ultimately assessments even shape who you can be.

ABILITY-GROUPING AND ASSESSMENT IN SECONDARY MATHEMATICS

There is a widely-held concern, supported by a significant body of research, that grouping students by ability is divisive, and results in severely limited opportunities for many students, particularly those from working class and some minority ethnic backgrounds. Despite this, in Britain during the last ten years large numbers of schools have reintroduced or widened their use of ability grouping (Boaler, 1997c). The primary reason for this appears to be a desire to boost a school's standing in the 'league tables' of school performance by making the school attractive to middle-class parents, who tend to prefer a high-degree of ability grouping within schools so that the education of their children is not disrupted by less motivated students (Gewirtz, Ball & Bowe, 1995). High stakes assessments therefore have an impact on what takes place in schools, in terms of how grouping is structured and how resources are allocated. Our research suggests that the impact on individual students is also significant.

While in many school subjects, such intensive use of ability grouping is a relatively recent phenomenon, mathematics has long been widely considered to be particularly unsuited to mixed ability teaching. Twenty years ago, when heterogeneous ability grouping was at its most popular, a government report found that 80% of mathematics teachers in England, compared to only 3% of teachers of English, thought that mixed-ability groups were inappropriate for teaching their subject (Her Majesty's Inspectors of Schools, 1979). The most recent data collected by school inspectors suggest that 94% of students in England are taught mathematics in homogenous ability groups in the upper secondary years.

In order to investigate the impact of ability grouping on students' achievement in, and attitudes to, mathematics, we have followed a cohort of approximately 1000 students in six secondary schools as they moved from being taught in mixed-ability groups in their seventh and eighth years of compulsory schooling (Grades 6 and 7)

to homogenous ability groups or 'sets' in years 9, 10 and 11 (see Boaler, Wiliam & Brown, 2000 for a fuller account of the study).

While the current pressure to prioritise examination performance is clearly being felt in mathematics classrooms, the reasons for mathematics teachers' near wholesale rejection of mixed ability teaching also relate to the dominant model of mathematics as highly abstract and very difficult. Mathematical ability is seen to be a rare commodity, and most students are assumed to be incapable of making much progress in the subject. Such perceptions were evident among the students we interviewed, and clearly have an impact on their sense of themselves as learners of mathematics.

Interviewer:	You don't think you're very good at it [maths]?
Dean:	No I'm not, I don't really have a natural gift for it I don't think.
I:	But you're in the top set.
Dean:	I think the only reason I'm in there is because in the first year we had Mr Williams and he said he wanted to push me. He didn't really think I was up to the standard but with a little push I could.

(Dean, set 1, Alder School[1])

Fathima:	Also people find maths very hard. There is always a psychological thing in your mind that maths is hard. No matter what, everyone thinks maths is hard. So when you're trying to concentrate you're thinking, no, maths is hard, I don't want to do it.
I:	So where do you think that comes from?
Fathima:	I don't know, people all around. People —you don't see mathematicians being a normal person— they have to be really big and brainy

(Fathima, set 1, Cedar School)

These perceptions both feed off, and feed into, the prevailing model of mathematics education in British secondary schools, and they are reinforced by assessment practices which emphasise differences between students. The notion of 'ability' is seen to be particularly salient in relation to mathematics, and the gulf between those who do and those who do not possess ability is assumed to be enormous, although as we saw earlier, this gulf is in no sense natural but a product of the way that success is defined in mathematics, and the need for 'good' assessments to discriminate between individuals. Whereas many school subjects are taught in such a way that the same activities may be tackled by students in a range of different ways, and at different levels, mathematics is more often taught according to a hierarchical model whereby new learning depends on what has previously been learnt, and it is much more usual for students of different abilities to be set completely different work, thus resulting in curriculum polarisation and restriction of the opportunity to learn.

Furthermore, most mathematics assessments require students to answer closed questions in predictable forms, and so mathematics teaching tends to focus on teaching a range of 'standard' procedures. The emphasis is on learning a series of

[1] The names of students and schools are, of course, pseudonyms.

steps and becoming fluent at applying them so as to obtain correct answers to closed questions.

As discussed above, the influences of these sorts of pressures serve to define the subject as it is taught in schools, and fundamentally change students' experiences of mathematics. In particular, they enshrine particular models of what it means to be successful in the subject, with the result that it is very much easier for some students than for others to regard themselves as being good at maths. As Broadfoot (1995, p. 68) argues:

> In education, as in other areas of contemporary social life, the advent of 'normalizing judgement' makes possible the idea of fixed definitions of competence. This normalizing judgement combines with the idea of 'hierarchical observation' to provide the 'rational authority' for competition and selection. […] This Benthamite notion of 'panoptic' surveillance, in which individuals learn to judge themselves as if some external eye was constantly monitoring their performance, encourages the internalisation of the evaluative criteria of those in power, and hence provides a new basis for social control.

Consistent with Boaler's earlier study (Boaler, 1997a), our work in the six schools suggests that while these factors operate in all mathematics classrooms, they are particularly salient in the group containing the highest-achieving students. This 'top set culture' tends to marginalise many of the girls who are put in these groups (Bartholomew, 2001). This culture both draws on, and reinforces, notions of the elusive nature of 'mathematical brilliance', and of there being a clear hierarchy of mathematical ability among students, and is fuelled by the emphasis on the speed of working typical of top sets (Boaler, Wiliam & Brown, 2000). The top set environment lends itself to easy competition between students, but the climate is one in which success, and therefore notions of 'ability', is determined by a student's capacity to generate large numbers of correct answers quickly. This reinforces the idea that the students with 'real talent' in mathematics are those who can perform at a high level in lessons without appearing to have to work very hard and, in a reversal of the usual association of bad behaviour with low ability, in top set groups it is often the students who 'muck about' to some extent in lessons who are regarded as being the students with most ability in the subject (Bartholomew, 2001). This resonates with Walkerdine's finding that, while 'hard-working' is seen as a positive trait by teachers in working-class schools, it is viewed negatively in middle-class schools, where academic success is expected of all, and students who have to work hard to achieve that success are regarded as lacking ability. Walkerdine also found that, whereas boys are frequently seen to have mathematical 'flair' regardless of their actual attainment, high achieving girls routinely have their success dismissed as the product of plodding hard work (Walkerdine & Girls and Mathematics Unit, 1989).

In many schools at which students are grouped according to their ability, the composition of the different groups is sharply polarised along social class lines, with middle-class students concentrated in the higher sets and working-class students in the lower sets (Gillborn & Youdell, 2000; Hallam & Toutounji, 1996; Harlen & Malcolm, 1997; Sukhnandan & Lee, 1998). In mathematics, it is also the case that

boys are frequently over-represented in top set classes and this was certainly the case in our six schools. Although within this study, the composition of the top set groups varies considerably from school to school, they are all places where the collection of values promoted speaks to a particular middle-class masculinity. The rationality of mathematics, the image of the 'great mathematician' and the possibility of being regarded as particularly clever if you can do well in mathematics without being seen to take your work too seriously, seem to have a particular potency for middle-class boys. In most of the top-set classes we have observed during the course of this study, the students who are regarded as being the 'best' in the class are those who display most confidence in lessons, who are quickest to find answers, and who make sure everyone else in the group knows that they got there first—often a group of middle-class boys.

Yet these conceptions of success are riddled with contradictions, and it is important to recognise that the 'pecking orders' established in top set groups do not represent an absolute hierarchy of mathematical ability. Rather the 'recognition rules' (see Lerman and Zevenbergen, this volume) focus on particular aspects of mathematical competence, such as speed, memorisation of facts, etc. which are not regarded as important by professional mathematicians (Buxton, 1981). The students who cannot, or for whatever reason do not wish to, respond with appropriate realisations are denied access to the highest status positions.

Those who are seen to have 'real ability' are generally those who can respond with appropriate realisations, but in practice this is often likely to be those students who are best at suspending their disbelief for long enough to apply the realisation rules (see Lerman & Zevenbergen, this volume) they have been taught. Boaler found that, while both girls and boys feel that highly procedural top set lessons limit their opportunities for understanding mathematics, in general boys are more inclined than girls to 'play the game' (Boaler, 1997b). Thus, while many boys appear able to work through a set of exercises without questioning too much, and to derive some meaning and motivation from competing with their classmates, many girls —unable or unwilling to compete on these terms— withdraw in lessons and are perceived by their teachers —and often their peers— to be lacking ability (Bartholomew, 2001; Boaler, 1997d).

As Lerman and Zevenbergen (this volume) point out, these recognition and realisation rules are conveyed to students through the pedagogic discourse, but they have their origin in the assessments that, through the processes described above, come to define what counts as mathematics.

We are not, of course, proposing that the students who are widely regarded as being good at mathematics are in fact less good at the subject than those who are seen to be struggling in lessons. Rather we are suggesting that the culture of top-set mathematics groups, and of mathematics more generally, reinforced by the assessments that are used, makes it very much easier for some students to believe themselves to be good at the subject than for others. The case of one particular 'top-set' student, Tania, who was able to change significantly her perception of herself as a learner of mathematics, illustrates the extent to which the responses of individual students are bound up with the wider context that defines, and constantly reinforces, what it means to be good at maths. In a questionnaire that she completed at the end

of Year 10 (grade 9), in response to the question 'what do you think are the bad things about your maths lessons?' she wrote:

> We go though the topics very quickly, without having enough time on one. A lot of the people in the class are naturally very clever, and it is embarrassing to get something wrong in front of them. (Tania, set 1, Willow School)

This reponse was typical of many of the 'top-set' students in this study. However, when she was interviewed in Year 11, she began by stating that her approach to mathematics had changed completely since the previous year. It is interesting to think about the ways in which she was able to change her perceptions of herself as a learner of mathematics:

> I: So something must have changed.
> Tania: My attitude. More thinking about myself than what other people know. That instead of what other people know and what I don't know, it's more what I know. Now I'm concentrating on that.
>
> (Tania, set 1, Willow School)

This sounds like a small step to take. She realised that by focusing on her own progress, rather than worrying that she is performing less well than others, she could concentrate on the areas where she needed to improve. Yet this awareness demanded considerable changes in her understanding about mathematics, and how success at the subject is achieved.

Boaler, Wiliam and Zevenbergen (2000, p. 201) argue that useful insights into the nature of mathematics education can be gained from shifting the focus away from the question of 'ability' and rather to think in terms of students 'belonging' to the community of practice (Wenger, 1998) of those who are successful at mathematics. They continue:

> Changing the emphasis from 'ability' to 'belonging' […] demythologises the special status of mathematics. The idea of 'belonging' immediately raises the question of 'belonging to what?' allowing the possibility of multiple communities of practice, rather than a single monolithic edifice.

It is exactly this 'single monolithic edifice' that leads so many students to believe that there is only one way to do mathematics. The fact that they are unable to compete with the quickest in terms of producing realisations is evidence that they lack that special talent. The dominant image of success is one from which many students feel excluded. In order to incorporate being successful at mathematics as an aspect of her own identity Tania first had to reconceptualise what it meant to be successful—in other words to change the recognition and realisation rules. This involved dismantling the hegemonic male-dominated model of the brilliant mathematician, and the belief—so deeply ingrained in many of the students we interviewed—that mathematics is something you either can or can't do. In recognising that different students approach mathematics in different ways, she was able to demystify the performance of some of the other students in her class, and, at the same time, to begin to feel more confident in her own abilities:

Tania: There were a couple of lessons where it really sort of hit me as like I was really working hard and I really changed my attitude in maths. I found that the people I thought were so clever…I was getting better marks than them and I was more ahead of them in class, while they were just like chatting. So well I thought, you know…

[…]

Tania: I think that with some people, like the people in my class — the ones people feel threatened by—those kind of people, I find that they'll just stick to it like this is it, this is how you have to do it and you always have to do it like this. Whereas me, I can't do it like that. That's why I bring in old work because I won't be able to answer the question like how they do it. So I'll try and bring in everything I know and try and find an answer.

[…]

I: So, what do you think it is that they do?

Tania: It's like…imagine we're doing an equation or something and we're trying to find a solution to it, they'll say "Here's the formula, this is what you do." Where I would probably go "If I look back at this topic, I can use that to solve this bit" and then I'll do that and then I'll get an answer like that.

(Tania, set 1, Willow School)

NATIONAL TESTING IN PRIMARY SCHOOLS

All students in England and Wales are tested at the age of 11 in English, mathematics and science. Although these assessments are relatively 'low-stakes' for students, in that little in terms of their individual futures is contingent on the results, the stakes are very high for their teachers. Schools whose students are judged to be gaining insufficiently good scores in these tests are subject to 'special measures' involving visits by government inspectors as often as every three months, and if subsequent improvements are deemed too slow, the school can be closed down, and all the teachers lose their jobs. These tests therefore create huge pressure on teachers to improve their results. In our study of a class of 11-year-olds (Reay & Wiliam, 1999) at Windermere School, we found that the teacher in turn placed huge pressure on the students to improve their performance:

I was appalled by how most of you did on the science test. You don't know anything. I want to say that you are judged at the end of the day by what you get in the SATs[2] and some of you won't even get level two[3].

Some of the students understood that it was the teachers, rather than the students, who were really being assessed in the tests:

I:	So what are the SATs for?
Jackie:	To see if the teachers have taught us anything.
Terry:	If we don't know nothing then the teacher will get all the blame.
Jackie:	Yeah. It's the teacher's fault.
Tunde:	Yeah. They get blamed.
[...]	
Mary:	SATs are about how good the teachers have been teaching you and if everybody gets really low marks they think the teachers haven't been teaching you properly.

However, some felt that although the tests were really assessing the quality of teaching, they could nevertheless impact upon their own lives:

I:	So are they important, SATs?
Lily:	Depends
Tunde:	Yes
Terry:	No, definitely not.
Lewis:	It does affect your life
Ayse:	Yeah, it does affect your life
Terry:	No, as if it means you know I do badly then that means I'm gonna be a road sweeper.
I:	You mean, you think that if you do badly in SATs then you won't be able to do well or get good jobs?
Jackie:	Yeah, 'cause that's what David [the class's teacher] is saying.
I:	What is he saying?
Jackie:	He's saying if we don't like, get good things in our SATs, when we grow up we are not gonna get good jobs and…
Terry:	Be plumbers and road-sweepers…
Tunde:	But what if you wanted to do that?
I:	Instead of what?
Terry:	Footballers, singers, vets, archaeologists. We ain't gonna be nothing like that if we don't get high levels.
I:	And does that worry you about your future?
Jackie:	Yeah.
Lewis:	Yeah.
Ayse:	Yeah it worries me a lot

[2] National curriculum assessment at the ages of 7, 11 and 14 —the end of each of the first three 'key stages' of compulsory education— consists of two components: a series of judgements made by the school about a student's performance over the key stage, generally called 'teacher assessment', and an externally set standardised assessment. Originally, these were to take the form of a series of tasks, called 'standard assessment tasks' (SATs) rather than traditional tests. However, by the time the system was fully implemented in 1994, the government had replaced the tasks with formal timed written tests. The external components of national curriculum assessment for 11 year olds have therefore never been called 'SATs', but teachers, students and parents continue to refer to them in this way, and so, for simplicity of presentation, we have followed this usage.

[3] Level 2 is the average level of achievement of 7-year-olds.

Terry: No, because he [referring to the teacher] is telling fibs.

For some, not only were the tests seen as critical filters, affording or denying admission to key occupations, but also as having predictive powers that extended into the moral sphere:

Sharon: I think I'll get a two, only Stuart will get a six[4]
I: So if Stuart gets a six what will that say about him?
Sharon: He's heading for a good job and a good life and it shows he's not gonna be living on the streets and stuff like that.
I: And if you get a level two what will that say about you?
Sharon: Um, I might not have a good life in front of me and I might grow up and do something naughty or something like that.

Now for some students, particularly older ones, a natural response to such a regime might be to resist and to absent oneself from the entire assessment process. As Foucault (1977) has observed, being documented was once the prerogative only of society's elite and even for most of the last century, assessments were used primarily for the minority, for example for entrance to higher education. In such a climate, not to have been assessed was unremarkable, and so such resistance might be a sensible strategy. However, where assessment is universal, then not to be assessed is to be marked. Those who are not assessed are not just lacking in some desired attributes. They are beyond the pale.

This appears to have been realised by some of the students. The tests for 11-year-olds in mathematics were 'tiered' in order to improve their reliability, so that each tier gave access to only a restricted number of the available levels. In order to ensure that students were entered for an appropriate tier, students scoring below the minimum threshold for a particular tier would not be awarded a lower grade but would instead not be awarded a level at all. This was forcefully communicated to students in this school, as the following interview extracts makes clear:

Hannah: I'm really scared about the SATs. Ms. O'Brien [a teacher at the school] came and talked to us about our spelling and I'm no good at spelling and David [the class teacher] is giving us times tables tests every morning and I'm hopeless at times tables so I'm frightened I'll do the SATs and I'll be a nothing.
I: I don't understand Hannah. You can't be a nothing.
Hannah: Yes, you can 'cause you have to get a level like a level 4 or a level 5 and if you're no good at spellings and times tables you don't get those levels and so you're a nothing.
I: I'm sure that's not right.
Hannah: Yes it is 'cause that's what Ms. O'Brien was saying.

I: Norma, why are you worried about SATS now?
Norma: Well, it seems like I'll get no points or I won't be able to do it, too hard or something.
I: What would it mean to get no points?
Norma: Well instead of being level three I'll be a nothing and do badly —very badly
I: What makes you think that? Have you been practising?

[4] Level 6 is a standard equivalent to that achieved by above-average 14-year-olds.

Norma: No, like I analyse … I know I worry about loads of things.

These extracts, and others from other students in the class, show that a metonymic shift took place over the year leading up to the tests. From thinking of themselves as students who might *get* a particular level, the students changed to talking about themselves as *being* a level three, four, five or six. The causes of this shift are, of course, complex, but there can be little doubt that a major influence was the culture of the school, which had embraced the need to improve its test scores irrespective of the consequence for the students' achievement in wider terms. Students were increasingly valued not for their personal qualities, but rather for what they could contribute to the targets set for the school by the school district. For many of the students in the class, the results of these assessments came to be bound up with not just what kinds of careers might be open to them, but who they were now, who they could be, and even their moral worth. Resistance to this process was not considered an option even by those students who had some insight into the nature and purpose of the assessment, while for others, the prospect of not being given a level at all was clearly worse than getting a level of some kind, no matter how low. This is particularly interesting in that the tier of entry is the decision of the school, rather than the student, and yet, the responsibility for failure in this regard has been effectively passed to the student. Thus despite their insights into this situation, the students accept that this is the way things have to be, and even though they know that the purpose of the tests is to assess the quality of the teaching they receive, failure is taken to be the responsibility of the student.

CONCLUSIONS

In this chapter, we have tried to show how apparently neutral assessments are not objective at all, but rather are 'objects of history'—created to fulfil particular social functions, which have shaped the assessments in particular directions that are not readily apparent. The seemingly innocuous requirement for the results of a test to be reliable requires that the test disperses individuals along a continuum so having the effect of placing a magnifying glass over a very small aspect of human performance, and this is particularly marked in mathematics. It represents a process of 'making difference' where little difference existed before. These hidden biases become especially important when the assessments are used as outcome measures for schooling processes, since the processes used in their development have inbuilt tendencies to maximise the differences between individuals.

At the same time, this process of maximising difference does so in a uni-dimensional way. Rather than maximising difference in terms of the various ways in which students differ, one particular variable is elevated to the exclusion of others. This is then exacerbated even further if only limited forms of assessment—for example, multiple choice tests—are used, presenting a stark realisation of the Macnamara fallacy[5]:

[5] Robert Macnamara was the USA Secretary of Defense during the Vietnam war who argued that the ratio of Viet Cong/North Vietnamese Army losses to USA/Army of the Republic of Vietnam losses was an important measure of military effectiveness: 'Things you can count, you ought to count. Loss of life is one.'

> The first step is to measure whatever can be easily measured. This is OK as far as it goes. The second step is to disregard that which can't easily be measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide. (Handy, 1994, p. 219)

In our study of six secondary schools the huge difference perceived by students and their teachers between students who are successful in mathematics and those who are not is not natural, but again the result of historical forces. Different definitions of mathematics would lead to different assessments that might not distinguish so sharply between students, but might distinguish students who had followed different curricula instead, thus making success a multi-dimensional, rather than a uni-dimensional construct.

Wenger (1988) shows how learning to become a medical claims processor involves adopting the practices of the community of claims processors:

> They learn how not to learn and keep their shoulders bent and their fingers busy, to follow the rules and ignore the rules. They learn how to engage and disengage, accept and resist, as well as how to keep a sense of themselves in spite of the status of their occupation. They learn how to weave together their work and private lives. They learn how to find little joys and how to deal with being depressed. What they learn and don't learn makes sense only as part of an identity, which is as big as the world and as small as their computer screens, and which subsumes the skills they acquire and gives them meaning. They *become* claims processors. (pp. 40-41, emphasis in original)

In the same way, for most students in mathematics classrooms, there is only one way to become successful as mathematics students, and that is to take on the role, the *identity*, of mathematician that is laid out for them by the school. At Willow School, the nature and importance of mathematics assessments promoted—or at least was entirely consistent with—a highly procedural pedagogic discourse, and discouraged approaches which would lead to a more critical view of the nature of mathematical knowledge. For four out of her five years of secondary schooling, Tania believed that she could not be successful at mathematics, that she could not become a participant in this community of practice, because she could not identify with the hegemonic masculine image of mathematics that was communicated to her through the assessments to which she was subjected.

Fortunately for her, she was able to carve out for herself a distinctive mathematical identity, but for a variety of reasons, other students will not have her agency. For them, the choice is either to take on the one particular version of the role of mathematician that the school lays out, or to disengage and identify themselves as being unable or unwilling to do mathematics. Ironically, these may be exactly the people who would make good mathematicians because of their desire for deep understanding.

At Windermere School, although some students (Terry for example) were able to resist the simple equation of test success with career success, most of the students were not. They genuinely believed that their test results would determine not just the

quality of their future lives, but also, in some cases (Sharon for example), their moral worth. The magnification of difference, along a small component of the whole make-up of an individual, led to a labelling of students so that they came to *be* their levels of attainment

In each of the two examples presented in this chapter, the assessments used have had a very powerful influence on the identity of students. Assessments serve as the message systems of communities of practice, informing the students about the extent of their participation in the community, but assessments also have an indirect effect on identity through their impact on learning:

> Because learning transforms who we are and what we can do, it is an experience of identity. It is not just an accumulation of skills and information, but a process of becoming—to become a certain person or, conversely, to avoid becoming a certain person. Even the learning that we do entirely by ourselves contributes to making us into a specific kind of person. We accumulate skills and information, not in the abstract as ends in themselves, but in the service of an identity. (Wenger, 1998, p. 215)

Assessment, learning and identity are therefore inextricably related. Although they are often taken as unexceptionable, assessments come to define fields of enquiry, and yet apparently innocuous requirements for reliability and validity have profound consequences. Those who end up as 'winners' and 'losers' is in large measure the result of the choices made in creating these assessments. To understand what these assessments do, to understand who can, and cannot be successful, and what that means for them, one needs to investigate the historical and social forces that have shaped those assessments. When 'researching the social', nothing can be taken for granted.

REFERENCES

Bartholomew, H. (2001). Learning environments and student roles in individualised mathematics classrooms. London (UK), University of London, Ph.D. dissertation.

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

Boaler, J. (1997b). Reclaiming school mathematics: The girls fight back. *Gender and Education, 9*(1), 285-306.

Boaler, J. (1997c). Setting, social class and survival of the quickest. *British Educational Research Journal, 23*(5), 575-595.

Boaler, J. (1997d). When even the winners are losers: Evaluating the experiences of 'top set' students. *Journal of Curriculum Studies, 29*(2), 165-182.

Boaler, J., Wiliam, D. & Brown, M.L. (2000). Students' experiences of ability grouping—disaffection, polarisation and the construction of failure. *British Educational Research Journal, 27*(5), 631-648.

Boaler, J., Wiliam, D. & Zevenbergen, R. (2000). The construction of identity in secondary mathematics education. In J.F. Matos & M. Santos (Eds.), *Proceedings of the Second International Mathematics Education and Society Conference* (pp. 192-202). Lisbon: Centro de Investigação em Educação da Faculdade de Ciências da Universidade de Lisboa.

Broadfoot, P.M. (1995). *Education, assessment and society: A sociological analysis.* Buckingham, UK: Open University Press.

Bursten, L. (Ed.) (1992). *The IEA study of mathematics III: Student growth and classroom processes*. Oxford, UK: Pergamon.

Buxton, L. (1981). Do you panic about maths? Coping with maths anxiety. London: Heinemann Educational Books.

Cherryholmes, C.H. (1989). *Power and criticism: Poststructural investigations in education*. New York: Teachers College Press.
Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington: American Council on Education.
Foucault, M. (1977). *Discipline and punish* (A.M. Sheridan-Smith, Trans.). Harmondsworth, UK: Penguin.
Garrett, H.E. (1937). *Statistics in psychology and education.* New York: Longmans, Green.
Gewirtz, S., Ball, S.J. & Bowe, R. (1995). Markets, choice and equity in education. Buckingham, UK: Open University Press.
Gillborn, D. & Youdell, D. (2000). *Rationing education.* Buckingham, UK: Open University Press.
Guion, R.M. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist, 29*(5), 287-296.
Hakel, M.D. (Ed.) (1997). *Beyond multiple-choice: Evaluating alternatives to traditional testing for selection.* Mahwah, USA: Lawrence Erlbaum Associates.
Hallam, S. & Toutounji, I. (1996). *What do we know about the grouping of pupils by ability? A research review.* London: University of London Institute of Education.
Handy, C. (1994). *The empty raincoat*. London: Hutchinson.
Hanson, F.A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley, USA: University of California Press.
Harlen, W. & Malcolm, H. (1997). *Setting and streaming: A research review.* Edinburgh: Scottish Council for Research in Education.
Her Majesty's Inspectors of Schools (1979). *Aspects of secondary education in England.* London: Her Majesty's Stationery Office.
Lemann, N. (1999). *The big test: The secret history of the American meritocracy.* New York: Farrar, Straus & Giroux.
Loevinger, J. (1957). *Objective tests as instruments of psychological theory. Psychological Reports, 3* (Monograph Supplement 9), 635-694.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012-1027.
Reay, D. & Wiliam, D. (1999). I'll be a nothing: structure, agency and the construction of identity through assessment. *British Educational Research Journal, 25*(3), 343-354.
Rose, N. (1989). *Governing the soul: The shaping of the private self*. London: Routledge.
Rutter, M., Maughan, B., Mortimore, P. & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children.* Shepton Mallet, UK: Open Books.
Selden, S. (1999). *Inheriting shame: The story of eugenics and racism in America.* New York: Teachers College Press.
Sukhnandan, L. & Lee, B. (1998). *Streaming, setting and grouping by ability: A review of the literature.* Slough, UK: National Foundation for Educational Research in England and Wales.
Walkerdine, V. & Girls and Mathematics Unit (Eds.) (1989). *Counting girls out.* London: Virago.
Wenger, E. (1998*). Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.