Journal for Research in Mathematics Education
2002, Vol. 33, No. 2, 78–110

The Interpretative Nature of Teachers' Assessment of Students' Mathematics: Issues for Equity

Candia Morgan
Institute of Education, University of London, United Kingdom

Anne Watson
University of Oxford, United Kingdom

This paper discusses fairness and equity in assessment of mathematics. The increased importance of teachers' interpretative judgments of students' performance in high-stakes assessments and in the classroom has prompted this exploration. Following a substantial theoretical overview of the field, the issues are illustrated by two studies that took place in the context of a reformed mathematics curriculum in England. One study is of teachers' informal classroom assessment practices; the other is of their interpretation and evaluation of students' formal written mathematical texts (i.e., responses to mathematics problems). Results from both studies found that broadly similar work could be interpreted differently by different teachers. The formation of teachers' views of students and evaluation of their mathematical attainments appeared to be influenced by surface features of students' work and behavior and by individual teachers' prior expectations. We discuss and critique some approaches to improving the quality and equity of teachers' assessment.

Key Words: Assessment; Communication; Equity/diversity; Reform in mathematics education; Social and cultural issues; Teaching practice

Reform of assessment methods and regimes is currently a concern in many countries around the world, and there is much discussion about the design of assessment tasks and systems that will provide valid and useful information for a variety of purposes, from immediate feedback to teachers and individual students to large-scale monitoring and evaluation of educational systems. In mathematics education, as in other subject areas, reformers have focused on the development of authentic or performance-based assessment of a broad spectrum of students' mathematical performance (e.g. Lesh & Lamon, 1992; Romberg, 1995; van den Heuvel-Panhuizen, 1996), often proposing substantial involvement of teachers as assessors of their own students (Black & Wiliam, 1998). The potentially positive influence of such assessment systems on the curriculum has also motivated reform (Bell, Burkhardt, & Swan, 1992; Stephens & Money, 1993). Although we are broadly in sympathy with the aims of such reform, we wish to raise some concerns about the nature of teachers' assessments of their students and the potential consequences for equity.

This paper arises from the United Kingdom[1] context in which alternative forms of assessing students' mathematical progress, including assessment by teachers, have been practiced alongside traditional methods for over twelve years. First, we review the ways that issues of equity and fairness are addressed in the literature on teachers' assessments in mathematics. We then develop the view that all assessment of mathematics is interpretative in nature and examine the classroom reality of informal assessment and the difficulties that arise in teachers' formal assessment of extended written mathematical tasks. Theoretical arguments are illustrated by reference to two independent empirical studies of teachers' assessment practices. Finally, we discuss some approaches to improving the quality of teachers' assessment practices.

Our concern with equity in assessment arises from the fact that the assessments made of students at all levels of education can have far-reaching consequences in their future lives. This is most obviously true in high-stakes assessments like the General Certificate of Secondary Education (GCSE) examinations in England. The results of such assessments are used explicitly for the differentiated allocation of curriculum tracks and further educational and employment opportunities, thus guiding students into different life paths. Similar uses are also made, however, of even the most fleeting and informal judgment made by a teacher of a student in the course of everyday work in the classroom. Although such assessments may not be formally recorded or even reflected upon, they nevertheless play a part in forming the educational future of individual students, often contributing to decisions about differentiation of the available curriculum. Such judgments affect the teacher's short-term behavior towards the student (e.g., the feedback provided, or the next task set) and influence the ways in which the teacher is likely to interpret the student's future performance (Walkerdine, 1988). Judgments that are based on unsound assessments can lead to unfair or unjust decisions and consequent inequitable treatment and opportunities for students. Such consequences of assessment operate in addition to systemic inequities in allocation of resources, an issue beyond the scope of this paper.

The issue of equity in assessment has generally been addressed from the point of view of attempting to ensure that all students are given equal opportunities to display their achievements and that assessment instruments do not have any systematic bias against particular social groups (see Gipps & Murphy, 1994, for a thorough review and discussion). Powerful critiques of some traditional methods of assessment in mathematics have identified the inequity inherent in them as well as the poor quality of the information they provide (Burton, 1994; Niss, 1993; Romberg, 1995). However, alternative assessment methods will not necessarily reduce inequity. Winfield (1995) identifies a number of aspects of alternative

[1] The curricula, assessment regimes, and approaches to teaching in the various countries that make up the United Kingdom are very similar. However, they are subject to different regulations. The studies reported in this article took place in England.

assessments that can affect equity, including the relationship between the assessment and the instructional conditions and possible mismatches between the expectations of assessment tasks and the ways in which learners draw on their cultural resources as they interpret them. Moreover, Baker and O'Neil (1994) report the concern of minority community groups about the potential for increased inequity with the introduction of performance assessments because of the unfamiliarity of the cultural contexts in which tasks may be set, a concern also raised by researchers in relation to social class (e.g., Cooper & Dunne, 2000). In response to these concerns, reformers have proposed multidimensional forms of assessment to allow all students to demonstrate what they know and can do as well as to ensure that the full range of mathematical objectives are addressed (de Lange, 1995; National Council of Teachers of Mathematics [NCTM], 2000). It has also been suggested that students' own teachers are in the best position to ensure that assessment is equitable because "probing what students are thinking, being sensitive to their experiences, and understanding how they perceive the assessment situation all contribute to making equitable decisions about students' learning." (NCTM, 1995, p. 15). Reformers argue that teachers can accumulate a multidimensional (and hence more valid) view of students' mathematics by assessing over time, using responses to a variety of tasks in several situations, and having knowledge of the context within which the student is working.

Our concern in this article is with inequity that arises not from the nature of assessment tasks but at the point of interpretation of student performance by teachers. This inequity may be comparative, in the sense that two students with similar mathematical understanding may be evaluated differently, or noncomparative, in the sense that a student who has achieved a particular form of mathematical understanding is not recognized to have done so. Although the focus of this view of inequity appears to be at the level of individual teachers' assessment of individual students, there is of course a strong possibility that some groups of students will be systematically disadvantaged because the behaviors they display to their teacher-assessors do not match those expected and valued as signs of achievement within the dominant culture. We present examples drawn from two empirical investigations into the practices of teachers who had been trained to make judgments in a criteria-referenced, multidimensional assessment system. These investigations lead us to question whether "reliable measures of mathematical achievement," as proposed by Baxter, Shavelson, Herman, Brown, & Valadez (1993, p. 213), are ever achievable and to identify some of the mechanisms of bona fide assessment practices that may lead to inequity.

RESEARCH ON TEACHERS' INVOLVEMENT IN ASSESSMENT

By moving away from so-called objective tests as a major form of assessment and introducing complex tasks that may be undertaken in a range of contexts, the traditional notions of reliability and objectivity may not be applicable. It is thus important to consider how fairness and equity may be conceptualized and achieved

in different forms of assessment. A number of potential sources of inequity arising from teacher involvement in assessment have been identified in the literature.

Inconsistent Application of Standards

For assessment to be equitable, different assessors should apply the same standards. The difficulty in achieving this is well recognized, especially where the assessment tasks are complex. In general, the response to this concern has been to emphasize the importance of training teachers to assess consistently. (Borko, Mayfield, Marion, Itexer, & Cumbo, 1997; Camp, 1993; Clarke, 1996). There is some evidence that, with training and experience, a substantial degree of agreement about standards of students' work can be achieved among teacher-assessors. In the field of English education, for example, teachers have had considerably more experience in making qualitative judgments about students' writing and acceptable levels of agreement are regularly achieved, though doubts are still expressed about the meaning of such judgments and the methods used to achieve them (see, for example, Filer, 1993; Wyatt-Smith, 1999). In mathematics, Baxter et al. (1993) report that different assessors can reach the same conclusion when assessing complex writing about special tasks done in research situations, though they point to the need for giving and marking (i.e., scoring) a wide range of tasks to allow for students' variable performance. However, assessor reliability found in research and training situations may be explained by the artificiality of the assessment activity, including the lack of accountability and the limited knowledge of the students, as well as the effects that judgments and decisions may have on them (Gitomer, 1993). Nevertheless, following five years of experience with large-scale portfolio assessment in the United States, Koretz (1998) reports that consistency among raters of eighth-grade mathematics portfolios improved "with refinements of rubrics and training" (p. 320) from a correlation of .53 to a correlation of .89.

Agreement among teacher-assessors may be explained by the development of shared constructs during the training and rating period (Roper & McNamara, 1993). Such constructs are not necessarily articulated by either the teachers or their trainers, but are manifested in "agreed marks" (i.e., high rates of agreement), a phenomenon described by Wiliam (1994) in relation to teachers assessing written work in mathematics. As Wiliam states, "To put it crudely, it is not necessary for the raters (or anybody else) to know what they are doing, only that they do it right" (p. 60). However, the development of shared expectations of students' work may result both in "pathologizing" unusual and creative approaches to assessment tasks and in stereotyping the tasks offered to students in order to maximize their chances of meeting assessment criteria (Morgan, 1998; Wiliam, 1994; Wolf, 1990).

Systematic Bias

The sort of training described above may address some of the inconsistencies between assessors' ratings but probably does not address systematic bias such as

may occur in the relationships between minority students and their teachers (Baker & O'Neil, 1994). For example, Cazden's study of Black and White teachers' responses to narratives produced by Black and White children suggests that the cultural and linguistic expectations of teachers belonging to one racial group may lead them to devalue the performance of students belonging to a different racial group (Cazden, 1988). Kraiger and Ford (1985) suggest that there are consistent differences in the ways that people assess members of different racial groups—differences that may not disappear even after training has increased general consistency. In a large-scale comparison between teacher assessments and standard tests of National Curriculum levels of students in England and Wales at Key Stage 1 (age 7), Thomas, Madaus, Raczek, and Smees (1998) found that students with special educational needs, those receiving free school meals (a surrogate measure of low socioeconomic group status), and those with English as a second language demonstrated lower achievement than others when assessed using written tests, but appeared to be even more disadvantaged when assessed by their teachers in "authentic" situations. Thomas et al. conclude that "certain aspects of how teachers judge student outcomes in the National Curriculum need to be examined in more detail. Indeed … findings suggest the possibility of systematic teacher bias" (p. 231). This finding was confirmed in a more recent study by Reeves, Boyle, & Christie (2001), in which they conclude that teachers consistently underestimate the attainment of those who have already been identified as having special educational needs. Thus, based on evidence from research studies, the presence of systematic bias demonstrates that increased reliability does not ensure equity.

Poorly Designed Tasks

The quality and consistency of teachers' informal assessment of their students are also dependent on the quality of the tasks used. This issue has been identified as a concern by several authors (e.g., Clarke, 1996). Senk, Beckman, and Thompson (1997) examined the kinds of assessment task used in 19 classrooms and found that, in general, teachers selected low-level abstract tasks that did not reflect the aims of reform curricula. Such a mismatch is likely to disadvantage those students whose achievements are stronger in reform aspects of the curriculum, restricting their opportunities to show what they can do. Advocates of recent reforms in curricula and assessment methods (e.g., Romberg & Wilson, 1995) propose that tasks designed to value students' demonstrations of what they know and can do are more in keeping with the aims of reform curricula than those that penalize them for what they do *not* know or *cannot* do. Such tasks are often practical, extended, and exploratory in nature, and are usually accompanied by specific and flexible marking (or scoring) criteria (Leder, 1992; Romberg, 1995; van den Heuvel-Panhuizen, 1996). Assessment methods based on practical or exploratory tasks have the potential to provide information that cannot be obtained from written tests. They may also serve to focus teachers' attention on specific aspects of mathematical performance rather than on general impressions, thus possibly avoiding inequitable judgments arising from teachers' preconceptions about their students.

Training teachers in the design and use of assessment tasks has also been proposed as a means of improving the quality of assessments (Clarke, 1996; Senk et al., 1997). In the United Kingdom, however, even after ten years of teaching and assessment reform and its associated training, Torrance and Pryor (1998) express concern about the methods used by teachers in formative assessment, including ambiguous questioning and tasks that are poorly focused on the subject matter.

There are also practical difficulties in giving specially designed assessment tasks to a whole class and monitoring each student's performance unless only written records are assessed (Barr & Cheong, 1995, p. 180). Administration of special tasks makes comparisons between different assessment environments difficult. These tasks do not take into account all of a student's mathematical achievements, nor are they useful for comparative purposes, because their meaning is dependent on interpretation and the specific context. None of these problems makes special tasks unjust in themselves, but such difficulties limit the uses to which the assessment outcomes can be put.

Studying Teachers' Assessment Practices

Although studies of the results of assessments can reveal inequity, they cannot tell us how this arises. Moreover, as Marshall and Thompson (1994) remark, results of assessments in research situations, with trained markers (raters), selected teachers, and in some cases artificial test situations, do not necessarily tell much about assessment in classrooms. In a more natural setting, Saxe, Gearhart, Franke, Howard, and Crockett (1999) have analyzed the changes in, and relationships between, the forms and functions of assessment used by teachers as they respond to pressures for reformed curriculum and assessment in the United States and suggest that teachers may not assess complex performance systematically. Most other studies of teachers' involvement in assessment (e.g., Gipps, Brown, McCallum, & McAlister, 1995; Senk et al., 1997; Stables, Tanner, & Parkinson, 1995) tend to be concerned with the assignment of grades as a result of formal, summative teachers' assessments, the ways teachers interpret procedures for assessing, recording and reporting, and their underlying beliefs about teaching and assessment rather than with the processes of interpreting student performance.

A notable exception is the comprehensive ethnography undertaken by Filer and Pollard (2000), which describes how different teachers formed different relationships with one student during her progress through primary school and formed correspondingly different assessments of her capabilities and achievement. In the context of secondary school mathematics, Rapaille's (1986) analysis of mathematics teachers' assessment practices suggests that the same teacher could allocate different marks to similar answers from students depending on the images of the individual students that the teacher had already formed. As Mavrommatis (1997) remarks in his discussion of classroom assessment in Greek primary schools, the implicit and "covert" nature of much teacher assessment makes it particularly difficult to research (p. 383).

THE INTERPRETATIVE NATURE OF ASSESSMENT

When teachers assess the texts produced by students, they read in an interpretative and contextualized way, relying on what Barr and Cheong (1995) term "professional judgement" (p. 177) to infer meaning in terms of the students' mathematical attainments. We are using the term "text" to refer to any verbal or nonverbal behavior that is taken by a teacher to indicate some form of mathematical attainment (or a lack of attainment). Thus, we include written and oral language in response to tasks set by the teacher or produced spontaneously by the student; we also include other signs such as drawings, gestures, and facial expressions, all of which may be read and interpreted by teachers as evidence of mathematical understanding. The teachers' professional judgment is formed not only from their knowledge of the current circumstances but also from the resources they bring to bear as they "read" the students' mathematical performance from these texts. These "reader resources" (Fairclough, 1989) arise from the teachers' personal, social, and cultural history and from their current positioning within a particular discourse. The professional enculturation of teachers seems likely to ensure a certain degree of common resource. Each individual act of assessment, however, takes place within a context that calls on teacher-assessors to make use of individual, as well as collective, resources. In teacher assessment of mathematics such resources include:

1. *Teachers' personal knowledge of mathematics and the curriculum, including affective aspects of their personal mathematics history*. For example, Watson (1999) describes a teacher who, having approached mathematics when at school in a way that had often been unusual compared to those of her classmates, was more willing than her colleagues to accept unusual methods from her students.

2. *Teachers' beliefs about the nature of mathematics, and how these relate to assessment*. Even when teachers are using the same method of assessment, its characteristics can vary substantially in practice among teachers who have different mathematical and curricular priorities (see, for example, Heid, Blume, Zbiek, & Edwards, 1999).

3. *Teachers' expectations about how mathematical knowledge can be communicated*. Individual teachers may also have particular preferences for particular modes of communication as indicators of understanding. Thus, what appears salient to one teacher may not to another (Morgan, 1998; Watson, 1999).

4. *Teachers' experience and expectations of students and classrooms in general*. Teachers' expectations about students' mathematical learning may be influenced by their existing notions of how a "good mathematics student" might behave, on the basis of evidence of nonmathematical aspects of behavior, social skills, gender, and social class background (McIntyre, Morrison, & Sutherland, 1966; Walkerdine, 1988).

5. *Teachers' experience, impressions, and expectations of individual students.* Early impressions are crucially important in the teacher's accumulation of information

about each student as these form a starting point for an ensuing cycle of interaction and interpretation (Nash, 1976).

6. *Teachers' linguistic skills and cultural background*. Mismatches between teachers' and students' language and culture are associated with lower evaluations of student performance (Bourdieu, Passeron, & Martin, 1994; Cazden, 1988; Kraiger & Ford, 1985).

Teachers may adopt a range of positions in relation to their students, other teachers and external authorities (Morgan, 1998; Morgan, Tsatsaroni, & Lerman, forthcoming). Different positionings (cf. Evans, 2000; Walkerdine, 1988) are likely to give rise to the use of different sets of reader resources and hence to different actions and judgments by different teachers or by a single teacher at different times in different circumstances.

Teachers' Construction of Knowledge About Students' Mathematics

What we have described above is a set of predispositions and experiences with which teachers approach the task of constructing their knowledge of students. The constructivist paradigm of much mathematics education research and curriculum development has led to recognition of students as active constructors of mathematical knowledge. There is also some recognition that teachers construct their knowledge of children's understanding (see Confrey, 1990; Simon, 1995) and that this does not necessarily coincide with some actual understanding. As von Glasersfeld (2000) says, "What we are talking about is but our construction of the child, and that this construction is made on the basis of our own experience and colored by our goals and expectations" (p.8). Cobb, Wood, and Yackel (1990) make clear that a teacher has to *interpret* classroom events and that the teacher is a learner, in general terms, about children's understandings. They found that teachers constructed knowledge about learning in general but made only tentative comments about individuals, treating these as working hypotheses rather than as assessments that could contribute to curriculum decisions or summative statements for an external audience. However, little research on assessment in mathematics education has focused on the teacher-assessor as an active constructor of knowledge about students' mathematical knowledge or acknowledged the essentially interpretative nature of the act of assessment.

INTRODUCTION TO THE TWO STUDIES

We present here examples from two studies of teachers' assessment practices, focusing on the question of how teachers interpret students' performance both in the classroom and in written exploratory assessment tasks. The two studies we describe and illustrate do not attempt to address the issue of systematic disadvantaging of social groups but focus on describing the mechanisms by which teachers' interpretations of student behavior are made and identifying sources of difference in those interpretations that have the potential to lead to inequity.

The two independently conceived but complementary research programs from which we draw our examples emerged from concerns about the consequences of particular methods of assessment for equity and with the quality of teachers' expertise as assessors when working with students in their own classrooms. Through close investigation of teachers' assessment practices—both formal and informal—and through interrogation of theories of interaction and interpersonal judgments, we conclude that a conventional notion of reliability is inappropriate when considering teacher assessment and, instead, we raise issues to be considered by the mathematics education community. Study A looks at teachers' assessment judgments made informally in the course of everyday teaching; Study B addresses the teacher-assessed component of a high-stakes examination. Both studies are located within interpretative paradigms on two levels: first, the teacher-assessor is conceived as interpreting the actions and utterances of the students; second, the researcher interprets the actions and utterances that the teachers make about the students (Eisenhart, 1988; Mellin-Olsen, 1993). Attention is focused on the possible meanings of the externally observable phenomena of students' mathematics and the ways in which teachers interpreted these. The researcher does not have privileged access to a so-called correct knowledge of the students.

Both studies originated within the national assessment system in England. This system is statutory for schools receiving public funding. Although the particular national context necessarily affects the detail of the teacher-assessor practices that we describe, our analyses are intended to illuminate much broader issues.

The Context of Assessment Systems in England

Teachers' roles as assessors in England are multiple; they assess in order to support the learning of their students, to help their students achieve good qualifications, to optimize summative results for their school, and to provide evidence for accountability. Though routine skills are largely assessed by written tests, teachers are also centrally involved in assessing the ways in which their students approach mathematical problems investigate and communicate mathematics, plan their work, and use and apply mathematics in nonmathematical, as well as mathematical, situations. High-stakes assessments are made by externally assessed examinations combined with teacher assessments. The results of these assessments are reported to parents and to the national Qualifications and Curriculum Authority. They are also used to compile "league tables," which are published each year and rank schools according to their assessment results.

The results of Key Stage Tests (scored by examiners employed directly by the national Qualifications and Curriculum Authority) are reported for students at ages 7, 11, and 14, alongside assessments provided by teachers, on the basis of the students' everyday work in class. The grade that students achieve at age 16 in the General Certificate of Secondary Examination (GCSE) combines scores on written examinations, assessed by examiners employed by national examination boards, and scores on reports of investigative problem solving, assessed by the students' own teachers. At all stages, teacher assessment is criteria-referenced, using descrip-

tors of levels of attainment in each of the four areas of the National Curriculum (Using and Applying Mathematics; Number and Algebra; Shape, Space, and Measures; and Handling Data). At GCSE, the parts of the descriptors of attainment in Using and Applying Mathematics relevant to investigative problem solving tasks are elaborated in greater detail (see Appendix A), and performance indicators related to specific tasks are also provided.

Open-ended explorations, with emphases on mathematical thinking, group work, discussion and extended written tasks, have been common and expected practice in mathematics teaching in England and other parts of the United Kingdom, in both primary and secondary phases, for many years. We are not claiming that all teachers teach in these ways, but we want to avoid drawing a distinction between traditional and reform methods because these coexist, even in the same classroom. For example, a teacher might use open-ended, student-centered methods yet assess learning through pencil-and-paper tests. Likewise, a teacher may teach predominately through teacher exposition and drill and practice but also include occasional investigative tasks in his or her repertoire. Such methods are now statutory as part of the National Curriculum introduced in 1989 and have been formally assessed as part of the GCSE examination since 1988, and they are also the product of gradual development over the last 30 years.

Training in assessment was given to all practicing secondary teachers when the GCSE started and to both primary and secondary teachers when the National Curriculum was introduced. In-service training, which has been in progress since the late 1980s, continues to be available. This takes the form of exercises in interpreting criteria, exercises in grading work of various kinds, agreement trials (in which teachers grade the same pieces of work independently and then meet in order to achieve agreement), and role-playing moderation procedures. New teachers learn about the statutory instruments and criteria in their preservice courses and are inducted into assessment practices in school through regular moderation discussions, in which a group of teachers examines and discusses samples of students' work in order to reach agreement about the grades that should be awarded. Teachers' assessment practices are expected to conform to national standards and are regularly inspected by government agencies. This well-established system therefore provides an important source of experience for countries whose curriculum changes are more recent than those in the United Kingdom.

The two studies reported here each took place toward the end of a period of adaptation to the new requirements, by which time all aspects of the GCSE and National Curriculum assessment systems were fully operational and teachers had all been trained in the assessment procedures and had been teaching and assessing in relevant ways for several years. All teachers in both studies had been trained in interpretation of criteria, continuous and authentic assessment, and moderation procedures. Although some had been involved in curriculum projects or other innovative work that may have led to, or presaged, national changes, others had not and may have been assessing in these ways largely because of the statutory requirements. If the teachers had been introduced to the assessment procedures

more recently, we might be inclined to conceive of any problems in their practice as the result of inexperience or inadequate training. Given the extent of the teachers' experience and training, however, we see the issues that arise from the studies as having a more general significance for systems involving assessment by teachers. For each study, we describe the methods used, provide and discuss an illustrative example, and outline further issues arising from the studies.

Study A: A Teacher's Constructions of Views of Students' Mathematics

The aim of Study A was to understand the ways in which teachers might accumulate and interpret their experience with students in mathematics classrooms, not only during formal assessment situations but also during normal day-to-day activity, and to critique the robustness of the ways in which the teachers formed their assessment judgments (Watson, 1995, 1998a). This study therefore goes some way towards dealing with the lack of research noted by Marshall and Thompson (1994) and provides an English perspective on the work of Senk et al. (1997).

In a previous study involving 30 teachers, Watson (1999) found that most of the teachers used a combination of record-keeping and personal recollection as the basis for their assessments. They spoke of "getting to know" their students as an essential contributor to assessment, and of using more formal assessment as one part of the process of "getting to know" students. They made summative assessments partly by personal recollection, a process which some of them, who were aware of the flaws of paper-and-pencil testing, regarded as a much fairer method.

Study A was designed to look in more depth at assessment processes in the classroom. Two teachers were each "getting to know" new students in Year 7, the first year of secondary education. The teachers each had in excess of 10 years' experience, both were well qualified to teach mathematics and had been using activity-based classroom approaches involving group work and discussion, extended work, and so on, for several years. Both teachers kept their repertoires up-to-date by attending out-of-school meetings and reading journals. They were fully involved in the study, knowing that its purpose was to find out more about how they developed their personal views of students and whether other views might be possible. Both teachers intended to accumulate knowledge of their students' mathematics performance over time, rather than substituting tests and special tasks for their intended holistic assessment style (Firestone, Winter, & Fitz, 2000).

In each class, the teacher and researcher together selected a small number of target students, none of whom initially appeared to represent extremes of attainment within the group. The target students were not told of their special role in the study, but both classes were aware that research was in progress about how mathematics was taught in their school.[2] The researcher observed one lesson a week with each

[2] At this time in the United Kingdom, getting informed consent to study individual students was not a requirement. The school authorities, the teacher, and the researcher agreed that it would be counterproductive to do so. It was decided that greater good would be achieved by pursuing the research this way than by informing the students and hence making the situation more artificial.

teacher during the first term with the new class and took detailed field notes. All public utterances by the target students in the observed lessons and some one-to-one interactions with them were noted; all written work produced by each student during the term was photocopied and kept as data; and behavior and actions were observed and noted using systematic observation punctuated with records of complete incidents. It was not possible to tape-record the target students, because they had to be unaware of their role, but it was possible to take written notes of all their public verbal comments and some of their other interactions. The researcher recorded what each target student was doing every few minutes. However, if one student was speaking or interacting with the teacher or doing some practical task, observations were made of the entire episode until the student returned to some continuous behavior such as doing exercises or listening to the teacher, when systematic observation would resume. After each lesson, the teacher and researcher would briefly discuss their views of incidents during the lesson, and field notes were taken during this discussion. Classroom field notes were copied for the teacher every three weeks, and there were two lengthy tape-recorded formal meetings during the term when all evidence was shared and discussed. The researcher, by virtue of watching the target students primarily, generally saw more of each one's actions, and saw these in more detail, during the observed lessons than the teacher did; however, the teacher had access to observational evidence from other lessons when the researcher was not present.

In the oral domain, the teacher's and researcher's experience overlapped but did not coincide. The researcher was aware of some utterances by students that were not noticed by the teacher, but the teacher again had access to evidence from other lessons. The students' written work was included in the discussion between teacher and researcher. In the written domain, teacher and researcher had the same evidence, including that from unobserved lessons, although in the case of the observed students, the researcher sometimes knew more about what the students had achieved but not written down, or what had preceded their written work. Both teachers in Study A were reluctant to use written work as a dominant source of evidence for students' mathematical prowess, particularly as they have to report on students' mathematical thinking, which might not be expressed—or expressible—in writing. In discussion, the teachers always contextualized written work by referring to what they knew about the circumstances of its production. The researcher also adopted this approach, accepting that the part that written work played in teachers' informal assessments appeared to be minor during the stage of the year when teachers were getting to know their students.

The regular discussions and formal meetings took the form of sharing data and impressions. At the formal meetings, participants gave a summary of their views of the students' mathematical strengths and weaknesses and described the data that they believed had led to these views. The researcher reconstructed episodes by juxtaposing the written work produced during or between the same lessons alongside records of what students said, did, and wrote. Then the complete data set was scrutinized, and alternative interpretations of the significance of different features

were discussed. The aim was neither to end up with the same view nor to generate a summative view but to explore differences and to attempt to explain why they arose. Both the researcher's and teachers' views would influence each other, and this was regarded as both inevitable and ethically desirable from the point of view of the students and the research participants.

Watson (1998a) gives further details of the collection and analysis of data, but for the purposes of this article, a summary of one case study highlights the kinds of problems that emerged. The following case of one 12-year-old student, Sandra,[3] illustrates ways in which the teachers' and researcher's observations and interpretations differed. Data about Sandra's behavior and utterances were extracted chronologically from the entire data set. The explicit aim of the discussions between teacher and researcher was to identify differences between what the two of them thought about Sandra and how these views had arisen. The key incidents arising from this discussion were those that supplied data suggesting different views or that stimulated the participants to talk about what may have influenced their views. The next sections of this article illustrate what emerged about Sandra during the formal meetings, focusing on two important areas of Sandra's mathematical achievement: mental arithmetic and mathematical thinking and problem solving.

Mental arithmetic. In verbal feedback sessions on arithmetic review questions done at home, Sandra frequently called out answers, waving excitedly and noisily when she wanted to contribute. Nearly all her enthusiastic contributions arose from work done at home (possibly with help from her parents; see episode below) or, very occasionally, from discussions with the teacher. The researcher noted that Sandra regularly changed her written answers in every lesson observed. The teacher was aware that she had altered some but did not know that she did so as frequently as was recorded in the field notes. As recorded in the researcher's field notes, the following episode, which took place in a whole-class session where students were giving answers to arithmetic homework, was typical of Sandra's behavior:

Teacher: What is the product of 125 and 100?
Sandra [*calls out*]*:* 225.
[*Teacher thanks her but explains she has given the sum.*]
Sandra [*audible to the researcher but not to the teacher*]:
But my mum says ... [*then inaudible*]
Teacher [*eventually, after getting answers from others*]:
12500.
Sandra [*loudly*]: Oh! that's what I had!

For later questions, Sandra calls out about half the answers correctly, but for the others, she changes the answers in her book. Maintaining a bouncy demeanor, she calls out at the end of the feedback, "I got nearly all right." After the lesson, the

[3] All students' and teachers' names in this article are pseudonyms.

teacher commented that Sandra and one other student had done better than the others on these questions, which was his view at that time.

The researcher noted that Sandra used her fingers when doing subtractions of small numbers in situations where a competent arithmetician would be likely to have known number bonds or to have developed some patterning strategies. The teacher had not seen this at all and said that his assessment of her competence in arithmetic—initially low based on results from a paper-and-pencil test—had risen to a relatively high level as a result of her oral contributions in class. The researcher saw a pattern to her oral contributions that the teacher did not, or could not, see. It was the pattern and circumstances of the contributions, rather than their quantity or nature, that indicated a contrast between Sandra's arithmetical abilities and her desire to be good with calculations. The teacher had a view that Sandra was mainly good at mental arithmetic, changing wrong answers in order to keep her confidence up, and assessed her as relatively strong in that area of mathematics. The researcher's view was that Sandra *wanted* to appear to be good but was regularly making and hiding errors that were not appearing in her final written work. If the teacher had been able to spend one-to-one time with Sandra, a situation that was impossible in a class of 30 students, these differences might have been explored further. Meanwhile, however, he had not assessed her as needing any particular help with arithmetic.

Mathematical thinking. In contrast to his view that Sandra was reasonably good at arithmetic, the teacher's perception was that Sandra was relatively weak in her ability to think mathematically while using and applying mathematics or when tackling new ideas, such as in exploratory work designed to allow the teacher to assess mathematical processes. However, the researcher observed several occasions when Sandra appeared to use mathematical thinking skills. For example, she was initially unsuccessful in making a rectangle with pentomino jigsaw pieces when using counting squares as a strategy. The teacher attempted to help her organize her counting method, but eventually she independently constructed another approach using an appropriate width and length as constraints within which to work successfully. During this time, she worked alone except when the teacher helped her to count. Nevertheless, later in the term, he said that she "doesn't see new ways of doing stuff."

In a lesson involving algebraic generalization of patterns made from strings of colored cubes, Sandra was working with a repeating sequence of three white cubes followed by one black cube. She called the teacher to her and said, "I'm trying to find out the reason why on a 2 it would be white, on 4 it would be black, on 6 it would be white and so on. So on 10 it would be white. So white occurs every other number. I don't know how to write that." Teacher said yes and then walked away. When asked about this later, he was unable to recall the incident but supposed that he had been happy with the verbal generalization and had not thought it worth pushing her further to an algebraic representation or that other students had needed him.

There were several other incidents in which Sandra appeared to do little except when the teacher was with her, and he commented to the researcher that "she always seems to have her hand up when I go past her." Nevertheless, some observations

suggested that she might be able to devise strategies, reuse strategies she had found effective in the past, describe patterns, and make conjectures resulting from pattern. The mismatch between the teacher's evaluation of Sandra's mathematical thinking as low level and dependent on his help and the researcher's interpretations that she had shown in some circumstances some features of mathematical thinking could suggest that the teacher had helped her to achieve progress in these areas but had not yet noticed this. Alternatively, it may be that he was always underestimating her thinking because the only times he had noticed it were in the context of her requests for help. It is also possible that the incidents observed by the researcher were atypical because not every lesson was observed. However, in Sandra's case, the teacher believed that discussing these incidents with the researcher had been useful, giving him further evidence for his evaluations, and he later observed and reported some similar incidents in which he, too, had noticed that she could think on a more abstract level. In some written work near the end of the term, Sandra had successfully drawn some complicated 3-D shapes constructed of interlocking cubes, a task which had involved imagery and reasoning as well as observation, and which few other students had managed. In this case, the teacher was able to say that he knew this work had been produced by Sandra on her own without direct help, but with frequent affirmation from him.

Relative to the researcher, therefore, the teacher initially appeared to overestimate Sandra's skills in mental arithmetic, the area of her mathematics achievement about which Sandra most wanted to impress him, and to underestimate her skills of reasoning, perhaps because she demonstrated less confidence about them or had less opportunity to articulate them, or perhaps because she created a negative impression by asking for help frequently. The teacher, seeing her work always in the context of what the rest of the class did and what his own expectations of her were, made a judgment that was *comparative* to what she had done before and to the rest of the class. But "what she had done before" included creating an impression in his mind; therefore, his judgments were relative to the picture already formed.[4] In this case, the teacher had already changed his mind once about Sandra because of a dramatic difference between a low entry test result and her confident demeanor in the classroom. He was slow to change his opinion again, even in the light of accumulating evidence, because what he saw could be explained in the context of his current view. Only when he had the chance to reflect on evidence presented as a sequence of similar incidents could he begin to review her strengths and weaknesses in mathematics. In general, we contend that initial impressions are likely to change only when it is not possible to explain later behavior in a way that is consistent with those impressions. Because any piece of evidence can be explained in multiple ways, the explanation most likely to be chosen is that which seems most natural in the light of existing theories and resources.

---

[4]The teacher has seen this analysis and accepts that it is not a criticism of his practice but rather a description of aspects of Sandra's work and the problems of informal judgment of which he was not aware.

This case illustrates several important features: the strong influence of and tendency to cling to impressions and the strong influence of obviously positive or negative behavior. To summarize, it suggests that teachers might also be influenced by: (a) seeing, or failing to see, patterns in responses and behavior, as shown in the differences between the teacher's and researcher's impressions; (b) types and patterns of behavior being overrepresented or underrepresented in the teacher's mental picture; (c) students' strong or weak social skills, such as Sandra's ability to attract the teacher's attention; (d) comparisons to early impressions, such as Sandra's enthusiasm about arithmetic; (e) time constraints that prevent a full exploration of a student's mathematics, as illustrated in the teacher's comment about having other students who needed him; (f) inability to see and use all the details that occur in classrooms, such as Sandra's finger-counting.

Similar influences were found in nearly all the ten cases of students studied in two classrooms (Watson, 1998a, 1998b), and similar results have also been reported by Ball (1997). In no sense are we suggesting that there is a true view of the student to be achieved or that the researcher is correct and the teacher is wrong. The outcomes of the study do not support these assumptions. On the contrary, the study suggests that informal assessments, including those that furnish information for summative assessments of performance, are inevitably and unavoidably influenced by a variety of factors that may have little to do with mathematical achievement. Of course, all these influences are acting also on the researcher, who is busy constructing her view of someone else's practice. So if "information about some aspects of mathematical performance as it is now conceived is best gained through observation" (Morony & Olssen, 1994, p. 397), then it is important to recognize that the partial and interpretative nature of classroom observation leads people to form legitimately different views even about standard tasks.

It may, however, be tempting to try to prevent bias in informal assessment from entering the formal process by making written work the only source of evidence. One may argue that such work, when it includes written reports of open-ended and extended problem solving and mathematical investigation, allows for the assessment of aspects of mathematical performance that cannot be assessed in timed tests (Black & Wiliam, 1998). Teachers' interpretation of written mathematics was the focus of Study B.

Study B: Teacher Assessment of Written Mathematics in a High-Stakes Context

The second study was set in the context of the GCSE examination for students aged 16+ in England. The 1988 reform of the public examination system introduced a component of *coursework*, completed in class and at home and assessed by students' own teachers, along with a more traditional timed examination assessed by external examiners employed by national examination boards. The coursework component, which at the time of its introduction could count for up to 100% of the GCSE but is now restricted to only 20%, most commonly takes the form of reports of one or more extended investigative tasks. These reports are intended to include

evidence of the mathematical processes that students have used (e.g., systematizing, observing, conjecturing, generalizing, and justifying) as well as the results of their investigation. These tasks thus involve students in producing sometimes lengthy written reports. It is, however, widely accepted that many students find it difficult to produce acceptable written evidence of their mathematical achievement (see, for example, Bloomfield, 1987).

Coursework tasks are assessed by students' own teachers, using a standard set of criteria of mathematical processes (see Appendix A), which are applied to all tasks. Because the set tasks allow students to choose the methods they will use and encourage some elements of original problem posing, it is difficult to provide more task-specific criteria (see Wiliam, 1994, for a discussion of assessment methods for this type of task). In practice, some examination boards publish task-specific "Performance Indicators" that describe the features of student work at various levels, but these are only meant to be illustrative of possible approaches to the task, and teachers are advised that "the Performance Indicators are only a guide and may need to be modified by the teacher in the light of specific individual cases" (Edexcel, 1999, p.8).

The original purpose of Study B was to investigate the forms of writing that students produced when presenting reports of extended investigative work for high-stakes assessments and to consider the match or mismatch between student writing and the forms of mathematical writing valued by their teacher-assessors (Morgan, 1995, 1998). A sample of three students' texts (written reports of their investigative work) was selected for each of two coursework tasks containing a range of linguistic characteristics (see Morgan, 1995, for details of the analysis and selection of these texts). Eleven experienced teachers from five secondary schools each read and evaluated the texts for one of the tasks during individual interviews. Before the interview, they were given time to work on the task themselves and to consider what they would look for when assessing it. The teachers were all experienced in using the general criteria for assessing such tasks, and these criteria were available for them to refer to if they wished. They were then prompted to think aloud as they read the students' texts and were encouraged to explain their judgments. Although we cannot assume that such an interview setting provides an authentic example of assessment practice, it does provide insight into the resources and strategies that teachers can use to make and justify evaluations of students' work and into the range of possible variation between them. Although some of the teachers expressed a lack of certainty about the "correctness" of the grades that they eventually allocated to the students' work, all tackled the task of assessment with confidence, appearing to be engaging in a familiar activity. Analysis of the interviews explored the teachers' assessment practices, identifying the features of the texts that the teachers attended to and the values that came into play as they formed judgments about the texts and about the student writers. Different teachers' readings of some sections of text were also compared in detail.

Most of the teachers had been promoted within their schools, holding posts as head or deputy head of a mathematics department or with other specified curric-

ular, pastoral,[5] or administrative responsibilities in addition to their teaching. Their positions and duties suggested that they were acknowledged as competent within their school communities. All had been trained in the use of the common set of criteria, were experienced in applying these criteria to their own students' work, and had participated in moderation processes both within their own schools and in relation to the decisions of external moderators employed by the examination boards. Their use of the language of the criteria during the interviews provided evidence of this familiarity.

The issue that we wish to consider here is the diversity that was discovered in the meanings and evaluations that different teachers constructed from the same texts. Given the small number of teachers and student texts involved, it is not appropriate to attempt to quantify the differences in the grades assigned to individual student texts, beyond commenting that, whereas the grades were consistent for some texts, they differed substantially for others. In one case, the same text was ranked highest (of a set of three texts) by some teachers and lowest by others and was assigned grades ranging from B to E where the possible grades ranged from A (highest) to G (lowest). We are more concerned here with how such differences may arise than with the differences themselves, and we present a single case study that illustrates one source of variation: the sense made by teachers of the mathematical content of a text. This example illustrates the ways in which teachers reading with different resources (including different prior experiences, knowledge, beliefs, and priorities) can arrive at very different judgments about the same student.

The task "Topples" involved investigating piles built of rods of increasing lengths, seeking a relationship between the length of the rod at the bottom of the pile and the length of the first rod that would make the pile topple over (see Figure 1). This task was one of a set provided by the official examination board for the

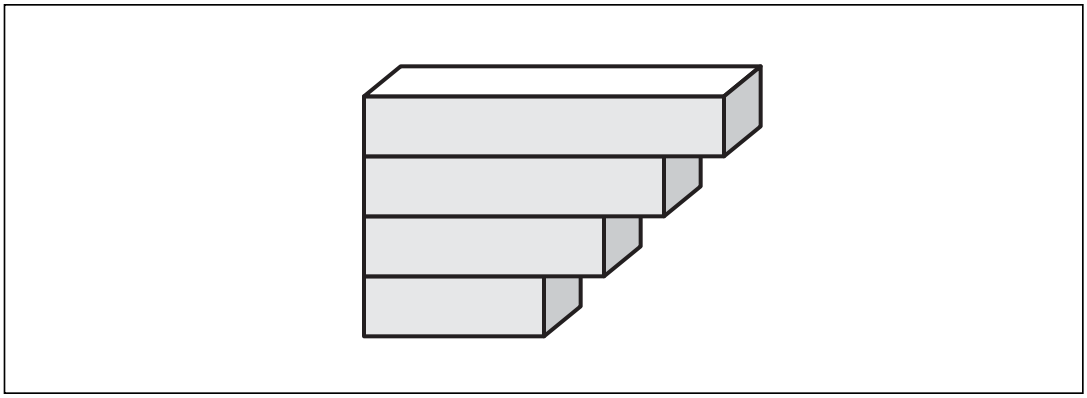


Figure 1. A pile of rods (the next rod might make the pile 'topple').

[5] *Pastoral* responsibilities include duties that American educators usually label as *guidance* or *counseling* activities. Teachers in the United Kingdom who have pastoral responsibilities may be called on to act as liaisons with parents, other caregivers, and social services; may play senior roles in disciplinary procedures within their schools; and may coordinate information from other teachers in order to monitor the overall behavior and performance of individual students.

specific purpose of allowing students to meet the general criteria for the coursework component of the GCSE (see Appendix A). The full text of the task is in Appendix B, and the performance indicators for this task are in Appendix C.

One student, Steven, completed the early parts of the task with results similar to those of many other students. He had built several piles of rods and found the length of the "topple rod" for piles of rods with bases of lengths from 1 to 10 units. He had tabulated his results and used the pattern in his empirical results to derive a formula relating the length of the "topple rod" to the length of the rod at the base of the pile. Having shown that he could use this formula to find the length of the "topple rod" for piles starting with rods longer than those he had available to build with, he then presented an alternative method for finding results for piles starting with long rods by scaling up his results for piles starting with short rods, illustrating this method by taking the result for a pile starting with a rod 10 units long and multiplying it by 10 to find the result for a pile starting with a rod 100 units long (Figure 2). The original formula and the work preceding this were similar to those produced by other students and their validity was not questioned by any of the teacher-readers. We are interested here only in the alternative method (which begins with "An alternative way to do this ..." in Figure 2), which was unique to Steven's report.

Steven had found the formula $(A + A) + \left(\frac{A}{2}\right) + b$, where A = the length of the rod at the base of the pile, and used it to calculate the length of the rod that would topple a pile that started with a rod of length 100 units:

$$(100 + 100) = 200 \qquad \left(\frac{100}{2}\right) = 50$$

$$200 + 50 = 250$$

250 would be the one at which the pile would topple.

An alternative way to do this would be to take the result of a pile starting at 10 and multiply it by 10.

$$(10 + 10) = 20 \qquad \left(\frac{10}{2}\right) = 5$$

e.g. $20 + 5 = 25$

or you could even take the basic result 1 without rounding it up, and you could multiply it by 100:

$$(1 + 1) = 2 \qquad \left(\frac{1}{2}\right) = 0.5$$

e.g. $2 + 0.5 = 2.5$

$$2.5 \times 100 = 250$$

Figure 2. Steven's alternative method.

No further justification of this method appeared in the text, which, in particular, gave no indication of how Steven had derived it. In order to evaluate Steven's work, each teacher-reader constructed his or her own understanding not only of the method itself but also of the means by which Steven might have derived it and of his level of mathematical achievement. The following extracts from interviews with three teachers (called here Charles, Grant, and Harry) who read this section of Steven's work illustrate how different these understandings can be.

Charles: Um OK, so I mean he's found the rule and he's quite successfully used it from what I can see to make predictions about what's going to happen for things that he obviously can't set up. So that shows that he understands the formula which he's come up with quite well, I think. There's also found some sort of linearity in the results whereby he can just multiply up numbers. Which again shows quite a good understanding of the problem I think.

Charles recognizes the mathematical validity of the alternative method, relating it to the linearity of the relationship between the variables. He takes this as a sign that the student has "come up with" the formula as a result of understanding the linearity of the situation. This results in a positive evaluation of Steven's mathematical understanding.

Grant: It's interesting that the next part works. I don't know if it works for everything or it just works for this, but he's spotted it and again he hasn't really looked into it any further. He's done it for one case but whether it would work for any other case is, er, I don't know. He hasn't looked into it … And he's used it in the next part, er, used th- this multiplying section in the next part, and it's just a knowledge of number that's got him there, I think, intuition whatever. He may have guessed at a few and found one that works for it.

Grant appears less confident about the mathematical validity of the alternative formula, expressing uncertainty about whether the method would work in general. Perhaps because of this uncertainty, his narrative explaining how Steven might have arrived at the method devalues this student's achievement, suggesting that the processes involved were not really mathematical: "spotting" the method, not looking into it properly, guessing, using "just a knowledge of number" or intuition. Steven is clearly not being given credit either for the result itself or for the processes that he may have gone through in order to arrive at it.

Harry: And he's got another formula here…. I don't really understand what he's done here…. So he's produced another formula where…. He's taken the result of a pile starting at ten and multiplying by ten and I don't understand what he's done there…. I would have asked him to explain a bit further. He's – the initial formula with two hundred and fifty is proved to be correct and he's trying to extend it. He's trying to look for other ways. Maybe he has realized that two hundred and fifty could be the exact answer or maybe not. So he's trying other ways to explain some of the inconsistencies that he's seen, but I think greater explanation [is] needed here.

Like Grant, Harry seems to have some difficulty making sense of the mathematics and does not appear to recognize the equivalence between the original formula and the alternative method. In spite of this, he is able to compose yet another narrative

to explain the student's intentions, stressing by repetition the suggestion that Steven has been "trying" (possibly with the implication that he has not succeeded). Harry, although appearing willing to applaud Steven's perseverance, also has a low evaluation of his mathematical understanding, even questioning whether he knows that his first method had already yielded a correct answer. Moreover, Harry locates the responsibility for his own failure to understand in the inadequacies of the student's text.

In this case, the three teachers' different interpretations of Steven's level of understanding and their different hypotheses about the methods he might have used seem to be connected to their personal mathematical resources. It is Charles, expressing the clearest understanding of the mathematics of the situation, who makes the most positive evaluation of Steven's understanding, whereas Grant and Harry, apparently uncertain of the general validity of the method, construct pictures of the student working in relatively unstructured or experimental ways.

Such major differences between teachers in their interpretations and evaluations of students' texts occurred primarily where the student's text diverged from the norm in some way—either in its mathematical content or in the form in which it was expressed. In this context, the norm appeared to be a text that followed a linear route from gathering data, through pattern spotting, to forming, verifying, and applying a generalization in a standard form. Our identification of this as a norm arises both from analysis of the teachers' interview data as a whole and from the literature on investigational work in the United Kingdom—for example, Wells (1993) and Hewitt (1992). Steven's presentation of an alternative method, though likely to be valid in other contexts, appeared as a deviation from the norm, not only because its mathematical form was nonstandard but also because it was unusual to see more than one method presented.

In order to make sense of a text (in this case a report of investigative work) and to use it to evaluate the student writer's achievement, each teacher must compose an explanatory narrative, drawing on the resources available to him or her. These resources include common expectations of the general nature of mathematical problem solving and reports of such work as well as more personal mathematical understanding and experiences. When a student text diverges from the usual to the extent that it is not covered by the established common expectations, each teacher must resort to his or her more personal resources, thus creating the possibility of divergence in the narratives they compose.

There is not space in this article to provide additional detailed examples, but we will briefly indicate some of the other ways in which teachers' interpretations and approaches to evaluation were found to differ in the broader study from which this example has been taken:

1. Teachers formed different hypotheses about work the student might have done in the classroom that was not indicated in the written report, and they expressed different attitudes towards valuing such unrecorded achievement. For example, an algebraic generalization unsupported by an earlier verbal generalization might lead

one teacher to suspect that the student had cheated whereas another would be "confident that … he's done it" (Morgan, 1998, p. 161).

2. Teachers made different judgments about some factual aspects of the text—for example, whether the wording of the problem given by the student had been copied from the original source or had been paraphrased in the student's own words (Morgan, 1998, p. 164).

3. Some teachers appeared to focus on building up an overall picture of the characteristics of the student, some were interested in making mathematical sense of what the student had done, and others focused solely on finding evidence in the text to meet specific criteria. A case study exemplifying this point can be found in Morgan (1996).

4. When there were tensions between the teachers' own value systems and their perceptions of the demands of externally imposed sets of assessment criteria, some teachers resolved this by submitting to the external authority whereas others were prepared to apply their own, unofficial criteria instead (Morgan, 1998, pp. 158–159).

Even when reading the same student's text, teachers may understand it in different ways. Even when they have formed similar understandings of what the student has done, they may assign it different values. The greatest discrepancies in final evaluations occurred in those cases where the student's work, like Steven's alternative formula, was unusual in some way. In cases where the student had produced a more conventional solution in a conventional form, teachers using different approaches to the assessment process and drawing on different resources seemed more likely to coincide in their overall evaluations. Students who are creative or produce unusual work—qualities that are officially endorsed by the aims and values of the curriculum reforms—are thus at risk because the value to be placed on their work depends crucially on the idiosyncratic resources of the teacher assessing it.

APPROACHES TO IMPROVED AND MORE EQUITABLE ASSESSMENT

Both the studies described above detail aspects of assessment practices that have the potential to lead to inequitable decisions about students' futures either through assignment of particular summative grades, thus affecting employment or life choices or through missing out on particular kinds of recognition, curricular support, or other opportunities, thus affecting future mathematical learning. In this section, we consider some approaches suggested within the mathematics education community that attempt to improve the quality of assessment, and their viability in the light of issues raised by the two studies. As we have argued, assessment is essentially interpretative. This means that, when we talk of the *quality* of assessment, we do not wish to suggest that *accuracy* is an appropriate criterion to apply. We are concerned rather, on the one hand, with the match between the stated aims of the curriculum and assessment systems and the outcomes of the assessment

and, on the other hand, with reducing the potential sources of inequitable judgments identified in the two case studies.

Could Specially Designed Tasks Ensure Equity?

Well-designed assessment tasks may go some way toward focusing teachers' attention on specific aspects of mathematical performance rather than the sort of general impressions found in Study A. However, they do not solve the problem of inequity in teacher assessment for the following reasons.

First, assessment still depends on interpreting particular actions of students. Some actions will be seen as significant and others not. Interpretations may not match a student's intentions but will depend, as we have seen in Study B, on the personal resources each teacher-assessor brought to the assessment. It is not merely a problem of task design, since many of the tasks used in Studies A and B were designed specifically for assessment of students' mathematical thinking and knowledge. Van den Heuvel-Panhuizen (1996) shows how tasks may be redesigned to leave less scope for alternative inferences, but this inevitably limits the opportunities for students to use higher-level problem solving processes, including specifying problems and choosing methods, that show their thinking skills.

Second, choice of method is a feature of mathematical problem solving that can be influenced by context, size of given numbers, form of presentation, available equipment, and so on. Hence, what is done is situationally specific and not necessarily generalizable for a student. Equally, absence of expected evidence may not indicate inability to generate such evidence in another task. Sandra, in Study A above, independently shifted from a low-level counting method of problem solving to one that was structural. At some other times, she remained stuck in a lower-level mode of working, even when helped. In attempting to make use of the results of such assessment tasks, teachers will interpret the meaning of the evidence (or the lack thereof) according to their preexisting views of the student.

Third, avoiding the effects of teachers' casual judgments is not possible, because the day-to-day interactions of the classroom are influenced by teachers' constructions of the students' mathematical attainment in the context of tasks designed primarily for teaching—for developing cognitive and metacognitive aspects of mathematics understanding—rather than for assessment. Reactions to these tasks influence the teacher's and students' views and expectations of what might happen in special assessment tasks. Teachers in Study B drew on their knowledge of the sorts of things that students generally did in classrooms in order to make sense of the texts they were assessing.

How Can Better Specification and Use of Assessment Criteria Avoid Inequity?

It is clear from the literature reviewed earlier that, in some circumstances, the clear specification of assessment criteria and training in the use of such criteria can improve the reliability of formal assessments made by groups of teachers. This by itself, however, is not enough to eliminate the possibility of differences among

teachers' interpretations. In the Study B, for example, the three teachers constructed very different understandings of the mathematical content of Steven's writing and of the mathematical processes that he might have gone through to achieve his results. The particular form of the mathematics produced by this student (and all the possible alternative forms) could not have been predicted by those designing the task in order to provide these teachers with guidance for all possible cases. If we are to make use of open-ended problems and investigations, encouraging students to be original and creative in mathematics (as advocated by curriculum reform movements in many countries), then criteria cannot be made specific enough to ensure that all teachers understand in identical ways the potential diversity of the mathematics produced by students. The experience of assessment of investigative work in England and Wales suggests that the existence of detailed criteria drives teachers towards setting less open, more stereotyped tasks (Wolf, 1990) and that, even where teachers themselves are committed to the ideals of investigative work and have been involved in the development of criteria, they become less appreciative of students' work that strays from predictable paths (Morgan, 1998; Wiliam, 1994).

There is some evidence to suggest that when students are aware of the criteria by which their work is to be assessed and are involved in the assessment process through peer- and self-assessment, they become "acculturated" (Tanner & Jones, 1994) and produce work that is closer to their teachers' expectations. The assessment of work produced by such students may thus be more likely to receive consistent evaluations. On the other hand, Love and Shiu's investigation of students' perceptions of assessment criteria indicated that some students also had a "sceptical awareness of the routinisation of producing work for assessment" (1991, p. 356). Again, there is the risk that students and teachers will direct their efforts towards producing work that is "safe," in that it matches routine norms, rather than taking possibly creative risks (cf. Gilbert, 1989, in the context of creative writing in English).

What Is the Role of Professional Dialogue in Improving Assessment Practice?

Studies of teachers working together over time suggest that, through the development of shared constructs, teachers can achieve greater reliability in the standards that they apply to examples of students' work. It has been suggested that such professional dialogue is the key to ensuring quality (and, by implication, equity) in teachers' assessments (Clarke, 1996). During the introduction of the National Curriculum in England and Wales, group moderation by teachers both within and between schools was recommended as a means of achieving reliability and as professional development for the teachers involved (Task Group on Assessment and Testing, 1987).[6] Moderation both during training sessions and as part of each

[6] In practice, the complex model of group moderation recommended by the National Curriculum Task Group on Assessment and Testing has never been implemented. It is necessarily time consuming and expensive—and hence unattractive to those who are eager for simple solutions to educational problems.

school's regular assessment practice is common for formal assessments such as GCSE coursework. The achievement of shared constructs and reliability in application of criteria and standards, however, does not address all the issues raised in the two studies we have presented here.

The three teachers in Study B had been trained and had participated in moderation of similar work in the past. That experience can be said to have been successful to the extent that they did not generally differ substantially in the final grades they allocated to most of the student texts. Their differences of opinion over Steven's work arose not from differences in the standards they were applying but from differences in their reading and interpretation of the idiosyncratic mathematical content of his text. Such differences might have been resolved if the teachers had had the opportunity to share and negotiate their interpretations, but it is surely not feasible to make every student text subject to such negotiated assessment.

The teachers in Study A were carrying out their assessment in the relative privacy of their own classrooms. The issue here is not so much the standards that the teachers were applying but the ways in which they became aware of and made use of the evidence available to them. The opportunity to share evidence and discuss impressions with the researcher may be seen as a form of professional dialogue. In order to examine one's own perceptions critically, it is necessary to become aware of possible alternative perceptions and interpretations and to engage with these.

One purpose of professional dialogue is to establish shared language with which to think about students' achievement as well as to communicate about it to others. Such a language should provide specialized resources particularly suited to the task of interpreting students' mathematical behavior, while helping to avoid the reliance on general resources such as the notion of ability that characterizes much classroom assessment practice (Dunne, 1999; Ruthven, 1987). However, professional dialogue and teacher education should not be seen simply as vehicles for teachers to acquire and apply similar understandings of criteria and standards. Indeed, the notion that there is one correct way of assessing a student or a piece of work underlies many of the potential problems we have identified. In contrast, we would suggest that one of the main benefits of professional dialogue among teachers is that it can raise doubts in teachers' minds about the certainty of their own interpretations of their students' behaviors. Resolving differences of opinion in assessment contexts by reaching a consensus is not necessarily the best course. Fully recognizing such differences may be a more useful approach, leading to an awareness of the possibility of alternative interpretations, of the inadequacy of the evidence available, and of the need to seek other perspectives. Dialogue with colleagues can thus provide teachers with questions with which to interrogate their own judgments.

CONCLUSION

The two case studies that we have presented not only illustrate the interpretative nature of assessment in both informal and formal contexts but also provide some

insight into the details of how different interpretations of students' achievements can occur. There are potential sources of inequity deeply embedded in traditional and reform processes for formative and summative assessment. Though closed-answer assessment tasks are known to pose problems in relation to equal access for students of different gender, ethnic group, and class, whenever evaluation of students relies on observation and interpretation of their behavior and their oral or written production, it will be influenced by the resources individual teachers bring to the assessment task. Such evaluation brings with it the possible sources of difference between teachers and differences in the ways individual students may be evaluated that we have illustrated in the two studies presented. The scope of the two studies has not allowed us to address directly the issue of systematic bias in relation to the assessment of students belonging to different social groups, but interpretation and evaluation of behavior are likely to be influenced by cultural expectations (Dunne, 1999; Walkerdine, 1988), and mismatches between the cultural and linguistic resources of teachers and students are likely to lead to evaluations that disadvantage students from nondominant social groups (Bourdieu, Passeron, & Martin, 1994; Cazden, 1988). We have argued that tighter specification of tasks or of criteria cannot remove such sources of inequity entirely and may indeed introduce other undesirable consequences for the quality of assessment. Interpretation is an essential characteristic of assessment activity and cannot be eliminated. However, insight into the details of how differences in interpretation occur may enable teachers to engage in critical reflection on their practice and, through such critical awareness, to lessen the likelihood of resultant inequity.

In day-to-day classroom interactions, we have identified the necessary incompleteness of teachers' awareness of their students' behavior, the potential for alternative interpretations of what is observed, the ways in which early judgments about individual students may influence subsequent interpretations of their behavior, and the consequent potential for inequitable evaluations. We certainly do not wish to question the integrity or competence of teachers as they make these judgments about their students, though raised awareness of processes of formative assessment (as described, for example, by Clarke & Clarke, 1998, and Torrance & Pryor, 1998) may at least increase the sources of evidence on which teachers base their judgments. Rather, the potential for inequity is a necessary consequence of the interpretative nature of assessment taking place in human interaction.

The concern for reliability in summative assessments has led to calls for tighter specification of criteria and training of teachers to develop shared ways of applying the criteria. We have argued that reliability is not a concept that can be applied simplistically to assessments of students' mathematics. Indeed, where students are given opportunities to respond to assessment tasks in open-ended ways, reliability may be an impossible goal. Moreover, it seems likely that tighter specification of criteria will lead to stereotyped responses from both teachers and students – in opposition to the value ascribed to creativity, openness and authenticity within 'reform' discourse.

Nevertheless, tentative judgments have to be made so that teachers can make pedagogical decisions. Teachers' own recognition that these determinations are situ-

ated and temporary indicates that such judgments should not be used as a basis for discriminatory action or high-stakes decisions. The issue of which teachers are less aware—the effects of their own beliefs and practices on equity—is harder to tackle, given that it affects students on many levels. Leaving the discussion of possible bias until summative judgments are made is too late; the model offered above of self-doubt and regular critical collegial discussion is rare, but it could be an important step towards equity.

REFERENCES

Baker, E. L., & O'Neil, H. F. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education: Principles, Policy, and Practice, 1*(1), 11–26.

Ball, D. L. (1997). From the general to the particular: Knowing our own students as learners of mathematics. *Mathematics Teacher*, *90*, 732–737

Barr, M. A., & Cheong, J. (1995). Achieving equity: Counting on the classroom. In M. T. Nettles and A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 161–184). Dordrecht, Netherlands: Kluwer Academic.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education*, *24*, 190–216

Bell, A., Burkhardt, H., & Swan, M. (1992). Balanced assessment of mathematical performance. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 119–144). Washington DC: American Association for the Advancement of Science.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5(1), 7–74.

Bloomfield, A. (1987). Assessing investigations. *Mathematics Teaching*, *118*, 48–49.

Borko, H., Mayfield, V., Marion, S., Itexer, R., & Cumbo, K. (1997) Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education*, *13*, 259–78.

Bourdieu, P., Passeron, J. C., & de Saint Martin, M. (1994). *Academic discourse: Linguistic misunderstanding and professorial power* (R. Teese, Trans.). Cambridge, UK: Polity Press.

Burton, L. (Ed.). (1994). *Who counts? Assessing mathematics in Europe*. Stoke-on-Trent, UK: Trentham Books.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*. Portsmouth, NH: Heinemann.

Clarke, D. (1996). Assessment. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International Handbook of Mathematics Education* (Vol. 1, pp. 327–370). Dordrecht, Netherlands: Kluwer Academic.

Clarke, D., & Clarke, B. (1998). Rich assessment tasks for Years 9 and 10. In J. Gough & J. Mousley (Eds.), *Mathematics: Exploring all angles* (pp. 93–96). Brunswick, Victoria, Australia: Mathematics Association of Victoria.

Cobb, P., Wood, T., & Yackel, E. (1990). Classrooms as learning environments for teachers and researchers. In R. B. Davis, C. A. Maher, & N. Noddings (Eds.), *Constructivist views on the teaching and learning of mathematics* (pp. 125–146). Reston, VA: National Council of Teachers of Mathematics.

Confrey, J. (1990). What constructivism implies for teaching. In R. B. Davis, C. A. Maher, & N. Noddings (Eds.), *Constructivist views on the teaching and learning of mathematics* (pp. 107–122). Reston, VA: National Council of Teachers of Mathematics.

Cooper, B., & Dunne, M. (2000). *Assessing children's mathematical knowledge: Social class, sex, and problem-solving*. Buckingham, UK: Open University Press.

de Lange, J. (1995). Assessment: No change without problems. In T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 87–172). New York: SUNY Press.

Dunne, M. (1999). Positioned neutrality: Mathematics teachers and the cultural politics of their classrooms. *Educational Review*, *51*, 117–128.

Edexcel (1999). *Mathematics 1385 and 1386: The assessment of MA1 1999*. London: Author.

Eisenhart, M. A. (1988). The ethnographic tradition and mathematics education research. *Journal for Research in Mathematics Education*, *19,* 99–114.

Evans, J. (2000). *Adults' mathematical thinking and emotions: A study of numerate practices.* London: Routledge.

Fairclough, N. (1989). *Language and power*. Harlow, UK: Longman.

Filer, A. (1993). The assessment of classroom language: Challenging the rhetoric of 'objectivity.' *International Studies in Sociology of Education*, *3*, 193–212.

Filer, A., & Pollard, A. (2000) *The social world of pupil assessment*. London: Continuum.

Firestone, W. A., Winter, J., & Fitz, J. (2000). Different assessments, common practice? *Assessment in Education: Principles, Policy, and Practice*, *7*(1), 13–37.

Gilbert, P. (1989). *Writing, schooling, and deconstruction: From voice to text in the classroom*. London: Routledge.

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995). *Intuition or evidence? Teachers and National Assessment of seven-year-olds*. Buckingham, UK: Open University Press.

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement, and equity*. Buckingham, UK: Open University Press.

Gitomer, D. H. (1993). Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 241–264). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heid, M. K., Blume, G. W., Zbiek, R. M., & Edwards, B. S. (1999). Factors that influence teachers learning to do interviews to understand students' mathematical understanding. *Educational Studies in Mathematics*, *37*, 223–49.

Hewitt, D. (1992). Train spotters' paradise. *Mathematics Teaching*, *140*, 6–8.

Koretz, D. (1998). Large-scale portfolio assessments in the U.S.: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy, and Practice*, *5*, 309–334.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, *70*, 56–65.

Leder, G. (Ed.) (1992). *Assessment and learning mathematics*. Victoria, Australia: Australian Council for Educational Research.

Lesh, R., & Lamon, S. J. (Eds.). (1992). *Assessment of authentic performance in school mathematics*. Washington, DC: American Association for the Advancement of Science.

London East Anglian Group. (1991). *Mathematics coursework tasks and performance indicators (1988–1991)*. London: Author.

Love, E., & Shiu, C. (1991). Students' perceptions of assessment criteria in an innovative mathematics project. In F. Furinghetti (Ed.), *Proceedings of the 15th Conference of the International Group for the Psychology of Mathematics Education* (Vol. II, pp. 350–357). Assisi, Italy: Program Committee of the 15th PME Conference.

Marshall, S. P., & Thompson, A. G. (1994). Assessment: What's new—and not so new—A review of six recent books. *Journal for Research in Mathematics Education*, *25*, 209–218.

Mavrommatis, Y. (1997). Understanding assessment in the classroom: Phases of the assessment process—The assessment episode. *Assessment in Education: Principles, Policy, and Practice*, *4,* 381–399.

McIntyre, D., Morrison, A., & Sutherland, J. (1966). Social and educational variables relating to teachers' assessments of primary students. *British Journal of Educational Psychology*, *36*, 272–279.

Mellin-Olsen, S. (1993). A critical view of assessment in mathematics education: Where is the student as a subject? In M. Niss (Ed.), *Investigations into assessment in mathematics education: An ICMI study* (pp. 143–156). Dordrecht, Netherlands: Kluwer Academic.

Morgan, C. (1995). *An analysis of the discourse of written reports of investigative work in GCSE mathematics*. Unpublished doctoral dissertation, Institute of Education, University of London, London.

Morgan, C. (1996). Teacher as examiner: The case of mathematics coursework. *Assessment in Education: Principles, Policy, and Practice*, *3*, 353–375.

Morgan, C. (1998). *Writing mathematically: The discourse of investigation*. London: Falmer Press.

Morgan, C., Tsatsaroni, A., & Lerman, S. (forthcoming). Mathematics teachers' positions and practices in discourses of assessment. *British Journal of Sociology of Education*.

Morony, W., & Olssen. K. (1994). Support for informal assessment in mathematics in the context of standards referenced reporting. *Educational Studies in Mathematics*, *27*, 387–399.

Nash, R. (1976). *Teacher expectations and student learning*. London: Routledge and Kegan Paul.

National Council of Teachers of Mathematics (NCTM). (1995). *Assessment standards for school mathematics*. Reston VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Niss, M. (Ed.) (1993). *Investigations into assessment in mathematics education*: *An ICME Study*. Dordecht, Netherlands: Kluwer Academic.

Rapaille, J. P. (1986). Research on assessment process in 'natural' conditions. In M. Ben-Peretz, R. Bromme, & R. Halkes (Eds.), *Advances of research on teacher thinking* (pp. 122–132). Lisse, Netherlands: Swets and Zeitlinger.

Reeves, D. J., Boyle, W. K., & Christie, T. (2001). The relationship between teacher assessment and pupil attainments in standard test tasks at Key Stage 2, 1996–8. *British Educational Research Journal*, *27*, 141–160.

Romberg, T. A. (Ed.). (1995). *Reform in school mathematics and authentic assessment*. New York: SUNY Press.

Romberg, T. A. & Wilson, L. D. (1995). Issues related to the development of an authentic assessment system for school mathematics. In T. A. Romberg (Ed.). *Reform in school mathematics and authentic assessment* (pp. 1–18). New York: SUNY Press.

Roper, T., & MacNamara, A. (1993). Teacher assessment of mathematics Attainment Target 1 (MA1). *British Journal of Curriculum and Assessment*, *4*(1), 16–19.

Ruthven, K. (1987). Ability stereotyping in mathematics. *Educational Studies in Mathematics*, *18*, 243–253.

Saxe, G. B., Gearhart, M., Franke, M. L., Howard, S., & Crockett, M. (1999). Teachers' shifting assessment practices in the context of educational reform in mathematics. *Teaching and Teacher Education*, *15*, 85–105.

Senk, S. L., Beckmann, C. B., & Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education*, *28*, 187–215.

Simon, M. (1995) Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, *26*, 114–145.

Stables, A., Tanner, H., & Parkinson, J. (1995). Teacher assessment at Key Stage 3: A case study of teachers' responses to imposed curricular change. *Welsh Journal of Education*, *4*(2), 69–80.

Stephens, M., & Money, R. (1993). New developments in senior secondary assessment in Australia. In M. Niss (Ed.), *Cases of assessment in mathematics education: An ICMI Study* (pp. 155-171). Dordrecht, Netherlands: Kluwer Academic.

Tanner, H., & Jones, S. (1994). Using peer and self-assessment to develop modeling skills with students aged 11 to 16: A socio-constructivist view. *Educational Studies in Mathematics*, *27*(4), 413–431.

Task Group on Assessment and Testing. (1987). Report of the task group on assessment and testing. London: Department of Education and Science and the Welsh Office.

Thomas, S., Madaus, G. F., Raczek, A. E., & Smees, R. (1998). Comparing teacher assessment and the standard task results in England: The relationship between students' characteristics and attainment. *Assessment in Education: Principles, Policy, and Practice*, *5*, 213–254.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment*. Buckingham, UK: Open University Press.

van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht, Netherlands: CD-ß Press.

von Glasersfeld, E. (2000). Problems of constructivism. In L. P. Steffe & P. W. Thompson (Eds.) *Radical constructivism in action: Building on the pioneering work of Ernst von Glasersfeld* (pp. 3–9). London: Routledge.

Walkerdine, V. (1988). *The mastery of reason*. London: Routledge.

Watson, A. (1995). Evidence for students' mathematical achievements. *For the Learning of Mathematics*, *15*(1), 16–20.

Watson, A. (1998a). *An investigation into how teachers make judgements about what students know and can do in mathematics*. Unpublished doctoral dissertation, University of Oxford, Oxford.

Watson, A. (1998b). What makes a mathematical performance noteworthy in informal teacher assessment? In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22nd conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 169–176). Stellenbosch, South Africa: University of Stellenbosch.

Watson, A. (1999). Paradigmatic conflicts in informal mathematics assessment as sources of social inequity. *Educational Review*, *51*, 105–115.

Wells, D. (1993). *Problem solving and investigations* (3rd ed.). Bristol, UK: Rain Press.

Wiliam, D. (1994). Assessing authentic tasks: Alternatives to mark-schemes. *Nordic Studies in Mathematics Education*, *2*(1), 48–68.

Winfield, L. F. (1995). Performance-based assessments: Contributor or detractor to equity? In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in educational testing and assessment* (pp. 221–241). Dordrecht, Netherlands: Kluwer Academic.

Wolf, A. (1990). Testing investigations. In P. Dowling & R. Noss (Eds.), *Mathematics versus the National Curriculum* (pp. 137–153). London: Falmer Press.

Wyatt-Smith, C. (1999). Reading for assessment: How teachers ascribe meaning and value to student writing. *Assessment in Education: Principles, Policy, and Practice*, *6*, 195–223.

Authors

Candia Morgan, Mathematical Sciences, Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, United Kingdom; c_morgan@ioe.ac.uk

Anne Watson, University of Oxford Department of Educational Studies, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom; anne.watson@educational-studies.oxford.ac.uk

Assessment Criteria for *Using and Applying Mathematics*

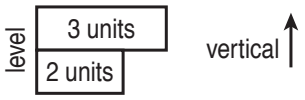
	Strand (i): Making and monitoring decisions to solve problems	**Strand (ii): Communicating mathematically**	**Strand (iii): Developing skills of mathematical reasoning**
1	Candidates try different approaches and find ways of overcoming difficulties that arise when they are solving problems. They are beginning to organise their work and check results.	Candidates discuss their mathematical work and are beginning to explain their thinking. They use and interpret mathematical symbols and diagrams.	Candidates show that they understand a general statement by finding particular examples that match it.
2	Candidates are developing their own strategies for solving problems and are using these strategies both in working within mathematics and in applying mathematics to practical contexts.	Candidates present information and results in a clear and organised way, explaining the reasons for their presentation.	Candidates search for a pattern by trying out ideas of their own.
3	In order to carry through tasks and solve mathematical problems, candidates identify and obtain necessary information; they check their results, considering whether these are sensible.	Candidates show understanding of situations by describing them mathematically using symbols, words and diagrams.	Candidates make general statements of their own, based on evidence they have produced, and give an explanation of their reasoning.
4	Candidates carry through substantial tasks and solve quite complex problems by breaking them down into smaller, more manageable tasks.	Candidates interpret, discuss and synthesise information presented in a variety of mathematical forms. Their writing explains and informs their use of diagrams.	Candidates are beginning to give a mathematical justification for their generalisations; they test them by checking particular cases.
5	Starting from problems or contexts that have been presented to them, candidates introduce questions of their own, which generate fuller solutions.	Candidates examine critically and justify their choice of mathematical presentation, considering alternative approaches and explaining improvements they have made.	Candidates justify their generalisations or solutions, showing some insight into the mathematical structure of the situation being investigated. They appreciate the difference between mathematical explanation and experimental evidence.
6	Candidates develop and follow alternative approaches. They reflect on their own lines of enquiry when exploring mathematical tasks; in doing so they introduce and use a range of mathematical techniques.	Candidates convey mathematical meaning through consistent use of symbols.	Candidates examine generalisations or solutions reached in an activity, commenting constructively on the reasoning and logic employed, and make further progress in the activity as a result.
7	Candidates analyse alternative approaches to problems involving a number of features or variables. They give detailed reasons for following or rejecting particular lines of enquiry.	Candidates use mathematical language and symbols accurately in presenting a convincing reasoned argument.	Candidates' reports include mathematical justifications, explaining their solutions to problems involving a number of features or variables.
8	Candidates consider and evaluate a number of approaches to a substantial task. They explore extensively a context or area of mathematics with which they are unfamiliar. They apply independently a range of appropriate mathematical techniques.	Candidates use mathematical language and symbols efficiently in presenting a concise reasoned argument.	Candidates provide a mathematically rigorous justification or proof of their solution to a complex problem, considering the conditions under which it remains valid.

APPENDIX B:
The "Topples" Task

TOPPLES

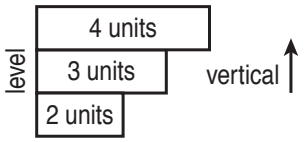
In this task you will be asked to balance some rods of different lengths on top of each other, until the pile topples.

The diagrams below are given as illustrations.



We start the pile with the 2 unit rod on the bottom and balance the three unit one on top of it, being careful that the left hand edges are level.

Then we balance the 4 unit rod on top of the three unit rod.



We continue building the pile, progressing through the sequence of rods, until the pile topples.

You should find that this pile of rods topples when we get to the 5 unit rod.

So the pile that starts with the 2 unit rod at the base eventually topples when we get to the 5 unit rod.

Your task is to investigate the relationship between the length of the rod at the bottom of the pile and the rod which first makes the pile topple.

1. Starting with rods of different lengths at the base, build up your piles until each one topples. Make sure that the rods increase by one unit of length at a time.
   (a) Record the length of the rod at the base and the length of the rod that makes the pile topple.
   (b) Tabulate your results.
   (c) Make any observations that you can.
   (d) GENERALISE.
   (e) Explain your result. (Well argued explanations based on intuition and insight will gain at least as much credit as those based on the principles of Physics.)
2. Imagine that you start with a rod of length 100 units and build up the pile using rods of lengths 101, 102, 103, … units.
   What will be the length of the rod that first makes the pile topple?
3. A pile topples when we place a rod of length 50 units on the top.
   (a) What will be the length of the rod on the bottom of the pile?
   (b) Explain your working.

OPTIONAL EXTENSION

Extend this investigation in any way of your own choosing.

APPENDIX C:

Performance Indicators for "Topples"

[From LEAG (1991). Mathematics coursework tasks and performance indicators (1988-1991). London: London East Anglian Group.]

The generalisation for this task is well within the syllabus at intermediate and higher level; it is a simple linear function. We would therefore expect to see an algebraic (symbolic) representation for this generalisation from candidates at grade C and above. From the candidates at grades B and A we would expect to see use of this algebraic form in the two specific cases given and to offer, certainly at grade A, some explanation of why the pile topples.

For the award of a grade D, candidates should be expected to do much of that for a grade C but to lack an element of sophistication. They might, for instance, fail to generalise in an algebraic form but be able to state generalisation in words or through a specific examples. They might even, at the lower levels of grade D or top grade E, answer the specific examples by extending their table of results.

The candidates who score a grade B will certainly have tried to undertake the task in an ordered, strategic manner, looking at well selected specific cases. They will also obtain a good table of results.

At the two lower levels, we would expect to see some reasonable attempts at the investigation but not handled in a any strategic fashion. The results at grade G are likely to be flimsy, few and not particularly accurate. For a grade F we would expect to see at least a couple, and preferably a trio, of correct results. The grade F and G candidates will not have been able to handle either of the specific examples.

As before, the optional extension should be used to enhance grades at all levels.