OPEN UNIVERSITY PRESS

McGraw - Hill Education

ASSESSMENT FOR LEARNING
Putting it into practice

> "This is a surprising and welcome book… a heartening read that shows the power of assessment for learning and the potential for academics and teachers jointly to put into practice ideas that can improve classroom learning and teaching."
>
> *TES*

This is an expansion of the widely sold booklets *Inside the Black Box* and *Working Inside the Black Box*, by the same authors.

The starting point of this book was the realization that research studies worldwide provide hard evidence that development of formative assessment raises students' test scores. The significant improvement in the achievements of the students in this project confirms this research, while providing teachers, teacher trainers, school heads and other leaders with ideas and advice for improving formative assessment in the classroom.

Assessment for Learning is based on a two-year project involving thirty-six teachers in schools in Medway and Oxfordshire. After a brief review of the research background and of the project itself, successive chapters describe the specific practices which teachers found fruitful and the underlying ideas about learning that these developments illustrate. Later chapters discuss the problems that teachers encountered when implementing the new practices in their classroom and give guidance for school management and LEAs about promoting and supporting the changes.

This book offers valuable insights into assessment for learning as teachers describe in their own words how they turned the ideas into practical action in their schools.

Paul Black is Emeritus Professor at King's College. He is well known for his work in science education and in formative assessment.

Chris Harrison taught science in London schools before coming to King's College. She combines research in assessment and development work with primary and secondary teachers.

Clare Lee taught secondary mathematics before becoming research fellow for the King's College formative project. She is now Teacher Advisor for Assessment for the Warwickshire LEA.

Bethan Marshall taught secondary English, drama and media and worked as English advisor for primary and secondary schools. Her research is focussed on English and on assessment.

Dylan Wiliam is Director of the Learning and Teaching Research Center at ETS, Princeton, New Jersey, USA. Before that, he was a secondary school teacher in London, and worked at King's College London from 1984 to 2003.

Cover design: Barker/Hilsdon

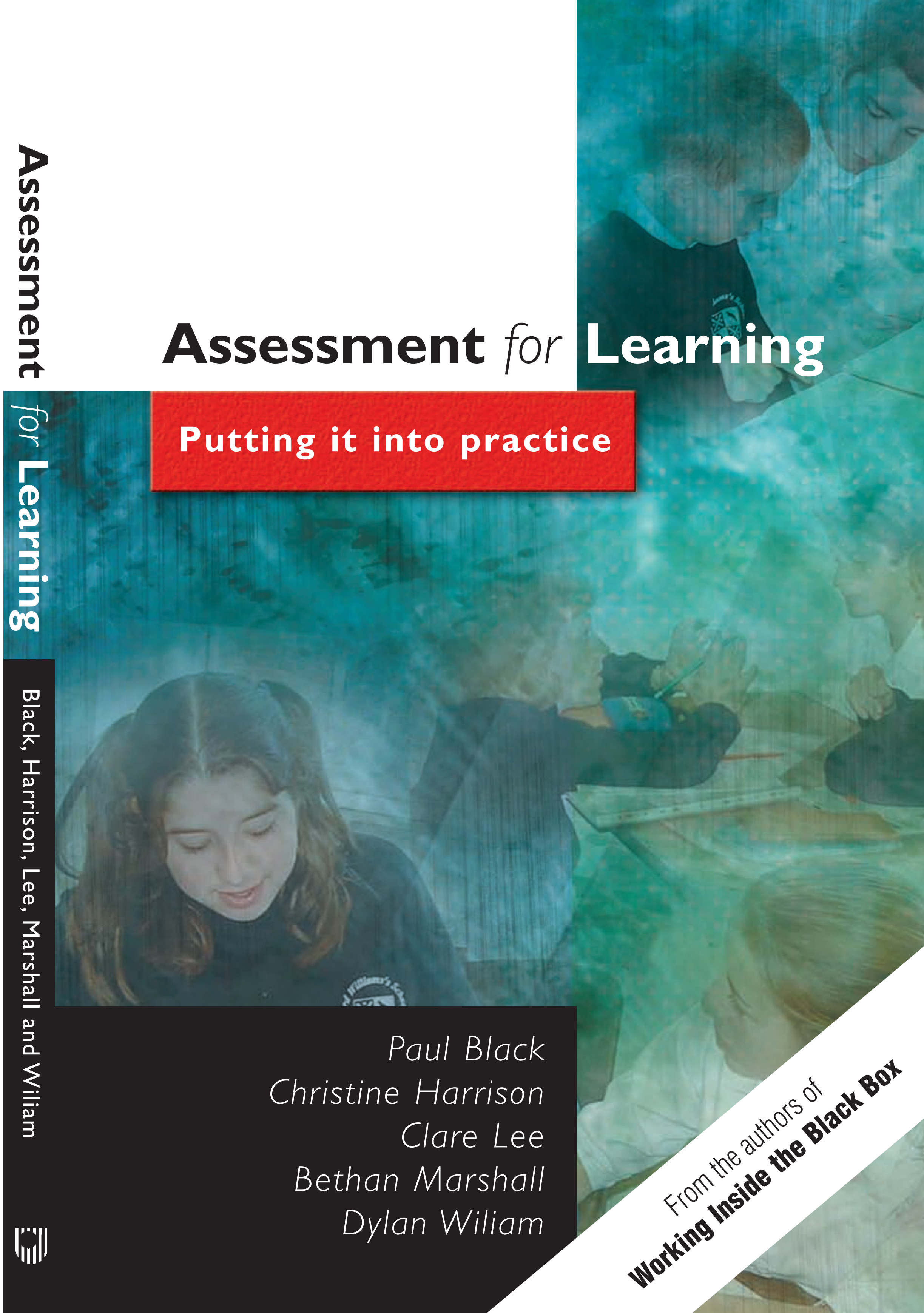
ISBN 0-335-21297-2

9 780335 212972

www.openup.co.uk

Assessment *for* **Learning**

Black, Harrison, Lee, Marshall and Wiliam



Assessment *for* Learning

Putting it into practice

Paul Black
Christine Harrison
Clare Lee
Bethan Marshall
Dylan Wiliam

From the authors of **Working Inside the Black Box**

Assessment for learning

Assessment for learning

Putting it into practice

Paul Black, Christine Harrison, Clare Lee, Bethan Marshall and Dylan Wiliam

Open University Press

Open University Press
McGraw-Hill Education
McGraw-Hill House
Shoppenhangers Road
Maidenhead
Berkshire
England
SL6 2QL

email: enquiries@openup.co.uk
world wide web: www.openup.co.uk

and

Two Penn Plaza, New York, NY 10121 - 2289, USA

A catalogue record of this book is available from the British Library

ISBN 0 335 21297 2 (pb)   0 335 21298 0 (hb)

Library of Congress Cataloging-in-Publication Data
CIP data has been applied for

Typeset by RefineCatch Limited, Bungay, Suffolk
Printed in Poland EU by

Contents

Acknowledgements

The origin of this work was an initiative of the Assessment Task Group of the British Educational Research Association (now known as the Assessment Reform Group). Their invitation to two of us (P.B. and D.W.) to review the research literature on formative assessment gave the initial impetus for the project with schools on which this book is based. We are grateful to them, but also to the Nuffield Foundation, who provided the support for the review and subsequently agreed to fund our project. We also acknowledge with pleasure the collaboration with Professor Myron Atkin and his colleagues at Stanford University and the funding provided by the US National Science Foundation through Stanford (NSF Grant REC-9909370), which enabled us to extend and broaden the range of our work.

We are indebted to the two local education authorities and their officers, notably Rose Collinson and Sue Swaffield from Medway and Dorothy Kavanagh from Oxfordshire, for their collaboration and support throughout the project. Through their efforts we were able to secure the commitment of the heads of six schools and to them we are also grateful. We acknowledge also the work of these officers and of representatives from the Department for Education and Skills and from the Qualifications and Assessment Authority, whose contributions in the project's steering group kept us in touch with many realities that we might otherwise have overlooked.

If this project has heroines and heroes, however, they are the forty-eight teachers of English, mathematics and science who took on the central and risky task of turning our ideas into practical working knowledge in their own classrooms. This project turned out to be far more productive and rewarding than we could ever have predicted: we are clear that the main reason for this lies in the commitment, to their educational ideals and professional values, that these teachers brought to their work with us.

The authors

Paul Black is Emeritus Professor at King's College. During his career he has been involved in a range of Nuffield curriculum projects and in many research projects, mainly in science education and in assessment. In 1987–88, he chaired the task group (TGAT) that advised ministers on the new national assessment and testing policy. Since his retirement he has concentrated on the study of formative assessment.

Chris Harrison taught science in several schools in the London area before she joined King's College in 1993. She now spends part of her time working with trainee teachers. Her work in formative assessment at King's has led to several professional development projects with teachers in the UK and abroad at both primary and secondary level. She has also worked on developing thinking skills in classrooms.

Clare Lee was the research fellow for the KMOFAP project. Previously she had taught secondary mathematics for over twenty years in several schools and researched, part-time, issues surrounding Teacher Action Research. Currently she is Teacher Adviser for Assessment for Warwickshire, leading the drive to establish Assessment for Learning within the LEA.

Bethan Marshall worked as an English teacher in London comprehensives for nine years before taking up her post at King's College. For five years she combined her post there with work as an English adviser. She is known for her writing and broadcasting on issues relating to English teaching and, latterly, on assessment issues.

Dylan Wiliam is Professor of Educational Assessment at King's. After several years teaching mathematics in London schools, he joined the college to work on a joint project with the ILEA to develop Graded Assessment in Mathematics as a new route to GCSE. He subsequently directed the first consortium to set the new key stage 3 national tests. He is well known for his work in both mathematics education and in assessment. He is currently an Assistant Principal of King's College.

1 Introduction

Why study this book?

This book is about changes in teachers' classroom practice that can make teaching and learning more effective. To be useful, such a book should be both practical, in giving concrete details and examples of classroom work, and principled, in giving a basis in both evidence and theory to underpin the practicalities. The authors of this book worked in a university and can lay claim to expertise in the areas of evidence and theory. The practical and concrete they have learnt in work with teachers, and the experiences, the evidence and the writing of these teachers is an important source for the work. So we are confident that teachers will benefit from taking our message seriously.

What is proposed?

What is proposed concerns assessment. This is not a simple or innocent term. Many take it to include all forms of testing. With this broad meaning, assessment can be seen to serve a range of purposes. One is to provide numerical results to publish in league tables – that is, the purpose is to help make schools *accountable*. Since this book is not concerned with assessments made for this purpose, we shall not discuss the controversies about the whole process (Black 1998; Wiliam 2001).

A second purpose is to provide students with *certificates*, such as GCSEs. The idea here is to give information about the students which they themselves, prospective employers and those controlling admission to further stages of education can use to make choices. This purpose calls for assessment methods which can be reliable, in that they are comparable across different schools, and indeed across the country as a whole, and also valid in that they give the users what they really need to know about each student. These are exacting requirements, but again assessments made for this purpose are not the main focus of this book.

As our title makes clear, this book is about a third purpose – *assessment for learning*. The focus here is on any assessment for which the first priority is to serve the purpose of promoting students' learning. It thus differs from the purposes described above. For the league tables or the GCSEs, the main assessment methods are formal tests: these usually, although not inevitably, involve tests that are infrequent, isolated from normal teaching and learning, carried out on special occasions with formal rituals, and often conducted by methods over which individual teachers have little or no control. Assessment for learning is not like this at all – it is usually informal, embedded in all aspects of teaching and learning, and conducted by different teachers as part of their own diverse and individual teaching styles.

An assessment activity can help learning if it provides information to be used as feedback by teachers, and by their students in assessing themselves and each other, to modify the teaching and learning activities in which they are engaged. Such assessment becomes *formative assessment* when the evidence is used to adapt the teaching work to meet learning needs.

Formative assessment can occur many times in every lesson. It can involve several different methods for encouraging students to express what they are thinking and several different ways of acting on such evidence. It has to be within the control of the individual teacher and, for this reason, change in formative assessment practice is an integral and intimate part of a teacher's daily work.

Why take formative assessment seriously?

Evidence of surveys of teacher practice shows that formative assessment is not at present a strong feature of classroom work. It follows that to establish good formative assessment practices in classrooms requires that most teachers make significant changes. Any non-trivial change in classroom teaching involves the teacher both in taking risks and, at least during the process of change, in extra work. We are confident, however, that teachers will find it worth their while to take on the changes that are involved in improving formative assessment, for the following reasons:

- There is ample evidence that the changes involved will raise the scores of their students on normal conventional tests.
- The changes involved have been shown to be feasible – that is, teachers have been able to incorporate them successfully in their normal classroom work.
- The work involved turns out to be a redistribution of effort: the message is not about working harder, but about working smarter.

- The changes can be made step by step – a big 'leap in the dark' is not necessary.
- Teachers come to enjoy their work more and to find it more satisfying because it resonates with their professional values.
- They also see that their students come to enjoy, understand and value their learning more as a result of the innovations.

These are bold claims. One purpose of this book is to explain the evidence and experience upon which they are based. Part of that evidence comes from existing books and research papers that recount the work of many groups from around the world. However, the major part comes from the experience of a group of teachers in six schools. In collaboration with us, these teachers have worked out, over the two and a half years of the project's work, how to implement reforms in their formative assessment. Their experience, our observations of their work and their reflections on the changes they have made are the bedrock for this book.

In particular, the claim that test scores can be raised by methods which teachers find professionally rewarding is based on the test results of this group of teachers, whether from their schools' normal tests, from key stage tests or from GCSE examinations. The teachers showed extraordinary commitment to this work, although this commitment was in part fuelled by the rewarding nature of the classroom work that they experienced. Apart from day release for twelve one-day meetings, they did not have any concessions to reduce their normal teaching load.

How this book tells its story

The content of this book is in part a story of our work in schools and in part a reflection on the lessons we infer from that story. As Figure 1.1 illustrates, there are three main themes, outlined as follows:

Overview

Chapter 2 describes lessons learned from an extensive survey of the relevant research literature which was completed in 1997, and published alongside the booklet *Inside the Black Box* (Black and Wiliam 1998a,b). The chapter sets out the principles, derived from this survey, on which the work with schools was based. Chapter 8 is a closing reflection. It will summarize the main lessons. It will also reflect on what we have learnt about making an impact, particularly in both highlighting and contributing to the most important debate of all, which is the debate about the improvement of the learning of students and of

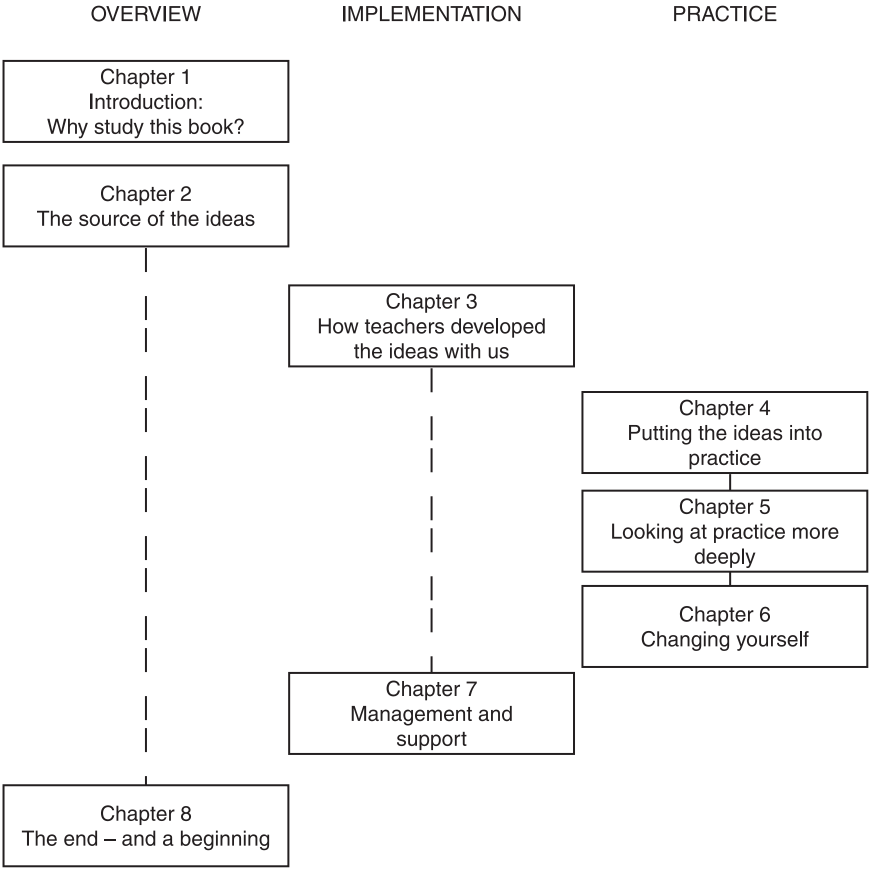

OVERVIEW
IMPLEMENTATION
PRACTICE
Chapter 1 Introduction: Why study this book?
Chapter 2 The source of the ideas
Chapter 3 How teachers developed the ideas with us
Chapter 4 Putting the ideas into practice
Chapter 5 Looking at practice more deeply
Chapter 6 Changing yourself
Chapter 7 Management and support
Chapter 8 The end – and a beginning

Figure 1.1 An outline of the book: three main strands.

their capacity to learn in the future. That is, it is about the core activity of all schools.

Implementation

Chapter 3 describes how we went about putting ideas into practice with six schools and the forty-eight teachers of English, mathematics and science involved. It also sets out the evidence about the significant learning gains that these teachers achieved. Chapter 7 returns to this theme by exploring ways in which, given the findings presented in this book, schools and those advising and supporting them might plan to implement formative practices.

Practice

Chapters 4, 5 and 6 are the heart of this book. They discuss the lessons learnt in practice from three perspectives. Chapter 4 sets out the concrete activities that the teachers developed as they transformed ideas about formative assessment into practical working knowledge. Chapter 5 looks at these activities from more fundamental perspectives, reflecting on them in terms of the principles of learning and of motivation that are entailed, and exploring also the similarities and differences between practices in different school subjects. Such reflections are prompted by our experience that changes in formative assessment practices, far from being just a set of useful tactical changes in classroom learning, have turned out to be far more significant and far more radical in their effects. Chapter 6 adopts a more personal and individual perspective, describing the experience of teachers as individuals as they worked to change both their approach to fundamentals of teaching and learning and their beliefs about these fundamentals. All three of these chapters draw upon the experiences and writing of the teachers involved. Apart from the three long pieces in Chapter 6, all references to teachers and their schools are pseudonyms.

Those mainly interested in practical application in their classrooms might concentrate on the three chapters on practice; those who look for ways to disseminate the practices will want to read Chapters 3 and 7 as well; while study of Chapters 1, 5 and 6 will help a reader to see the developments from a more fundamental and theoretical perspective. Chapters 2–6 draw mainly on our own work, especially our work with schools. In Chapter 7, we provide ideas from a wider range of sources.

2 The source of the ideas

Introduction

Although some of us have been interested in formative assessment for over 20 years, the origin of the work on formative assessment that is described here was the review by Black and Wiliam (1998a). This review covered a very wide range of published research and provided evidence that formative assessment raises standards and that current practices are weak. However, there was little to help teachers put the research findings into practice. This was followed by the booklet *Inside the Black Box* (Black and Wiliam 1998b), which served four aims:

- The first was to give a brief review of the research evidence.
- The second was to make a case for more attention to be paid to helping practice inside the classroom.
- The third was to draw out implications for practical action.
- The fourth was to discuss policy and practice.

This chapter will concentrate on the first two of these aims. The third has been the main aim of our work since 1998 and is the main theme of later chapters of this book. The fourth will be looked at only briefly in the closing chapter.

The research evidence

The review by Black and Wiliam (1998a) involved studying reviews of research published up to 1988 and then checking through the issues of over 160 research journals and books for the years 1988 to 1997. This process yielded about 681 articles or chapters to study. The seventy-page review drew on material from 250 of these sources. One of the priorities in evaluating the research reports was to identify and summarize studies that produced quantitative evidence that innovations in formative assessment can lead to improvement in the learning of students. Brief accounts of four such studies will serve here to give the flavour of the evidence.

The first was a project in which twenty-five Portuguese teachers of mathematics were trained in self-assessment methods on a 20-week part-time course, methods which they put into practice, as the course progressed, with their students – 246 of them aged 8 and 9 years (Fernandes and Fontana 1996). The students of a further twenty Portuguese teachers who were taking another course in education at the time served as a control group. Both the experimental and the control groups of students were given the same pre- and post-tests of mathematics achievement, and both spent the same amount of time in class on mathematics. Both groups showed significant gains over the period, but the experimental group's mean gain was about twice that of the control group – a clearly significant difference. Similar effects were obtained for some older students.

The focus of the assessment work was on regular – mainly daily – self-assessment by the students. However, this focus meant that the students also had to be taught to understand both the learning objectives and the assessment criteria; they were also given the opportunity to choose learning tasks and to use these in assessing their own learning outcomes. Thus the initiative involved far more than simply adding some assessment exercises to existing teaching. So this research raised a question: whether it is possible to introduce formative assessment without some radical change in classroom pedagogy because, of its nature, this type of assessment is an essential component of classroom learning.

The second example was itself a review of twenty-one different studies, of children ranging from pre-school to grade 12 (Fuchs and Fuchs 1986). The main focus was on work for children with mild disabilities, and on the use of the feedback to and by teachers. The studies were carefully selected – all involved comparison between experimental and control groups, and all involved assessment activities with frequencies of between two and five times per week. For each study, the authors first calculated the difference between the gain in scores of the experimental over the control group, and then divided this figure by a measure of the spread of the scores across the children of either group. They did this because they could use this ratio, which is known as the 'effect size', to compare different studies with one another. The overall mean of the effect sizes was 0.73 for handicapped children and 0.63 for the non-handicapped. Where teachers worked with systematic procedures to review the assessments and take action accordingly, the mean effect size was 0.92, whereas where action was not systematic it was 0.42.

Two features of this last example are of particular interest. The first is that the authors compared the striking success of the interactive (i.e. formative), approach with the unsatisfactory outcomes of projects which used diagnostic pre-tests only as a filter to assign children to pre-prepared individual learning programmes. The second feature was that the main learning gains from the formative work were only achieved when teachers were constrained to use the data in systematic ways, ways which were new to them.

The third example was undertaken with 5-year-old children (Bergan *et al.* 1991). It involved 838 children drawn mainly from disadvantaged home backgrounds in six different regions in the USA. The teachers of the experimental group were trained to implement a system that required an initial assessment to inform teaching at the individual pupil level, consultation on progress after 2 weeks, new assessments to give a further diagnostic review and new decisions about students' needs after 4 weeks, with the whole course lasting 8 weeks. There was emphasis in their training on observations of skills to assess progress, on a criterion-referenced model of the development of understanding and on diagnostic assessments designed to help locate each child at a point on this model. Progress in reading, in mathematics and in science in the experimental group was considerably greater than in the control group even though the tests used were multiple-choice and not well suited to the child-centred style of the experimental group. Furthermore, of the control group, on average 1 child in 3.7 was referred as having particular learning needs and 1 in 5 was placed in special education; the corresponding figures for the experimental group were 1 in 17 and 1 in 71.

The researchers concluded that the capacity of children is underdeveloped in conventional teaching so that many are 'put down' unnecessarily. One feature of the experiment's success was that teachers had enhanced confidence in their powers to make referral decisions wisely. This example illustrates again the embedding of a rigorous formative assessment routine within an innovative programme linked to a criterion-based scheme of diagnostic assessment.

The fourth example was a study of an inquiry-based middle-school science curriculum module (White and Frederiksen 1998) that was focused on a practical inquiry approach to learning. There were twelve classes of thirty students each in two schools. Each class was taught to the same curriculum plan and all students worked in peer groups. A control group of classes spent part of the classroom time on a general discussion of the module, while an experimental group spent the same length of time on discussion structured to promote reflective assessment, with both peer assessment of presentations to the class and self-assessment. All students were given the same basic skills test at the outset. On comparison of the scores gained on their projects, the experimental group showed a significant overall gain. However, when the students were divided into groups according to low, medium or high scores on the initial basic skills test, the low scorers were better than the control group by more than three standard deviations, the medium scorers by just over two standard deviations, and the high scorers by just over one standard deviation. A similar pattern, of superiority of the experimental group, was also found for scores on a test of the physics concepts. For students in the experimental group, those who showed the best understanding of the assessment process achieved the highest scores.

Here again the formative assessment was built into an innovation to change teaching and learning. Three features stand out: the use of 'reflective assessment' in peer groups, the use of several outcome measures all directly reflecting the aims of the teaching, and the fact that the intervention was most effective for the lowest attaining students.

In all, about twenty relevant studies were found: the second example described above (by Fuchs and Fuchs 1986) was one of the twenty and itself reviewed twenty-one studies, so in effect the body of evidence included over forty studies. All of these studies showed that innovations that include strengthening the practice of formative assessment produce significant, and often substantial, learning gains. The studies ranged over ages (from 5-year-olds to university undergraduates), across several school subjects and over several countries. The mean effect sizes for most of these studies were between 0.4 and 0.7: such effect sizes are among the largest ever reported for sustained educational interventions. The following examples illustrate some practical consequences of such large gains:

- An effect size of 0.4 would mean that the average (i.e. at the 50th percentile) pupil involved in an innovation would move up to the same achievement as a pupil at the 35th percentile (i.e. almost in the top third) of those not involved.
- A gain of effect size 0.5 would improve performances of students in GCSE by at least one grade.
- A gain of effect size 0.7, if realized in international comparative studies in mathematics (TIMSS; Beaton *et al.* 1996), would raise England from the middle of the forty-one countries involved into the top five.

Some, but not all, of the studies showed that improved formative assessment helped the (so-called) low attainers more than the rest, and so reduced the spread of attainment while also raising it overall. Any gains for such students could be particularly important: they show that the 'tail' of low educational achievement might be due, at least in part, to failure to develop the potential talents of the 'weaker' student.

It therefore seemed clear that very significant learning gains might be achievable. The fact that such gains had been achieved by a variety of methods, which had, as a common feature, enhanced formative assessment, indicated that it is this feature which accounted, at least in part, for the successes. It also showed that the positive outcomes might not depend on the fine details of any particular innovation. However, it did not follow that it would be an easy matter to achieve such gains on a wide scale in normal classrooms. The research reports did bring out, between and across them, other features that appeared to characterize many of the studies:

- All of them involved new ways to enhance feedback between those taught and the teacher, ways which required new modes of pedagogy and therefore significant changes in classroom practice.
- Underlying the various approaches were assumptions about what makes for effective learning – in particular that students have to be actively involved.
- For assessment to function formatively, the results had to be used to adjust teaching and learning – so a significant aspect of any programme would be the ways in which teachers do this.
- The ways in which assessment affected the motivation and self-esteem of students, and the benefits of engaging students in self-assessment, both deserved careful attention.

Current practice

The second feature of the research review was to look for research evidence about the quality of the everyday practice of assessment in classrooms. This evidence showed that such practice was beset with problems and shortcomings, as the following quotations indicate:

> Marking is usually conscientious but often fails to offer guidance on how work can be improved. In a significant minority of cases, marking reinforces under-achievement and under-expectation by being too generous or unfocused. Information about pupil performance received by the teacher is insufficiently used to inform subsequent work.
>
> (General report on secondary schools – OFSTED 1996)

> Why is the extent and nature of formative assessment in science so impoverished?
>
> (UK secondary science teachers – Daws and Singh 1996)

> The criteria used were ‘virtually invalid by external standards’.
>
> (French primary teachers – Grisay 1991)

> Indeed they pay lip service to it but consider that its practice is unrealistic in the present educational context.
>
> (Canadian secondary teachers – Dassa *et al.* 1993).

The most important difficulties, which were found in the UK but also elsewhere, could be briefly divided into three groups. The first was concerned with *effective learning*:

- Teachers' tests encourage rote and superficial learning; this is seen even where teachers say they want to develop understanding – and many appear unaware of the inconsistency.
- The questions and other methods used are not discussed with or shared between teachers in the same school, and they are not critically reviewed in relation to what they actually assess.
- For primary teachers in particular, there is a tendency to emphasize quantity and presentation of work and to neglect its quality in relation to learning.

The second group was concerned with *negative impact*:

- The giving of marks and the grading functions are over-emphasized, while the giving of useful advice and the learning function are under-emphasized.
- The use of approaches in which students are compared with one another, the prime purpose of which appears to them to be competition rather than personal improvement. In consequence, assessment feedback teaches students with low attainments that they lack 'ability', so they are de-motivated, believing that they are not able to learn.

The third group focused on the *managerial role* of assessments:

- Teachers' feedback to students often appears to serve social and managerial functions, often at the expense of the learning functions.
- Teachers are often able to predict students' results on external tests – because their own tests imitate them – but at the same time they know too little about their students' learning needs.
- The collection of marks to fill up records is given greater priority than the analysis of students' work to discern learning needs; furthermore, some teachers pay no attention to the assessment records of previous teachers of their students.

Of course, not all of these descriptions apply now to all classrooms and, indeed, there will be many schools and classrooms to which they do not apply. Nevertheless, these general conclusions were drawn by authors in several countries, including the UK, who had collected evidence by observation, interviews and questionnaires from many schools.

Planning further work

The need

Overall, the research review of Black and Wiliam (1998a) attracted widespread attention in the academic world. The whole of the journal issue in which it appeared was devoted to the topic, the review article being followed by six articles, each of about ten pages, in which experts in the field from the USA, Australia, Switzerland and South Africa commented on the review. Although these added many valuable insights, none of them challenged its main findings.

For the world of professional practice, the booklet *Inside the Black Box* and the article based on it published in a US journal, met its overall aim of attracting attention and raising debate. Its success in this respect is evidence of the importance of the issues raised, and of the fact that the message speaks to the basic professional concerns of very many teachers. It has been widely quoted in policy and professional circles and at the time of writing has sold over 40,000 copies.

The obvious question raised by the nature of the message and the evidence of positive impact was whether any follow-up action should be taken. Innovations in formative assessment had already been undertaken and reported in the literature before 1998. Apart from those directly concerned with the issue, formative feedback was also a feature in several other innovations, notably mastery learning, assessment by portfolios, curriculum-based assessment and cognitively guided instruction. It did not seem feasible, however, to attempt to replicate any of these in UK schools working within the constraints of the national curriculum and assessment. So any action would have to be based on a selection of those ideas in the research literature which appeared to be both feasible and potentially productive.

A key reservation about any such action was expressed in the following passage from *Inside the Black Box*:

> Teachers will not take up attractive sounding ideas, albeit based on extensive research, if these are presented as general principles which leave entirely to them the task of translating them into everyday practice – their classroom lives are too busy and too fragile for this to be possible for all but an outstanding few. What they need is a variety of living examples of implementation, by teachers with whom they can identify and from whom they can both derive conviction and confidence that they can do better, and see concrete examples of what doing better means in practice.
>
> (Black and Wiliam 1998b, pp. 15–16; emphasis in original)

Thus any ideas, however promising, derived from the research would have to be tried out in practice.

The basis

The research review did set out some of the issues that any programme of development might have to face. These were concerned with teacher change, students' perspectives and the central concept of feedback, respectively.

The problem of 'teacher change' was emphasized in several of the research reports. There was evidence of patchy implementation of reforms of teacher assessment in France (Broadfoot *et al.* 1996) and in Canada (Dassa 1990), while in the UK some changes had produced a diversity of practices, some of which appeared to be counter-productive and in conflict with the stated aims of the changes that triggered them (McCallum *et al.* 1993; Gipps *et al.* 1997). Where changes had been introduced with substantial training or as an intrinsic part of a project in which teachers had been closely involved, the pace of change was often slow – it was very difficult for teachers to change practices that were closely embedded within their whole pattern of pedagogy (Torrie 1989; Shepard *et al.* 1994, 1996; Shepard 1995) and many lacked the interpretive frameworks required to coordinate the many separate bits of assessment information in the light of broad learning purposes (Bachor and Anderson 1994). A project with teachers in the creative arts, which tried to train them to communicate with students to appreciate the students' view of their own work, found that despite the training many teachers stuck to their own agenda and failed to respond to cues or clues from the students that could have re-oriented that agenda (Radnor 1994). In a project aimed at enhancing the power of science teachers to observe their students' at work, teachers could not find time for observing because they were not prepared to change classroom practices to give students more free responsibility and give themselves a less closely demanding control – the authors interpreted this as a reluctance to break the existing symbiosis of mutual dependency between teachers and students (Cavendish *et al.* 1990).

The main issue that emerged from such studies is that there are close links between formative assessment practice, the other components of a teacher's own pedagogy and a teacher's conception of his or her role. It followed that implementation of changes in classroom assessment would call for rather deep changes both in teachers' perceptions of their own role in relation to their students and in their classroom practice. It was also evident that the context of national or local requirements for certification and accountability exerted a powerful – usually harmful – influence on assessment practice.

For 'students' perspectives', the central problem was clearly expressed by Perrenoud:

> A number of pupils do not aspire to learn as much as possible, but are content to 'get by', to get through the period, the day or the year without any major disaster, having made time for activities other than school work [. . .] Formative assessment invariably presupposes a shift in this equilibrium point towards more school work, a more serious attitude to learning [. . .] Every teacher who wants to practise formative assessment *must reconstruct the teaching contracts so as to counteract the habits acquired by his pupils*. Moreover, some of the children and adolescents with whom he is dealing are imprisoned in the identity of a bad pupil and an opponent.
>
> (Perrenoud 1991, p. 92; emphasis in the original)

This rather pessimistic view was supported by several research studies. Some students might be reluctant to participate in any change not only because of a wish to minimize effort, but also because of fear and insecurity. Another problem is that students might fail to recognize formative feedback as a helpful signal and guide (Tunstall and Gipps 1996). Overall, a student's beliefs about learning were important in forming these responses. Issues of confidence, in oneself as a learner, and of motivation would serve to trigger the adoption of a negative response to change.

That these considerations would be central was indicated by the theoretical position of Sadler (1989). He pointed out that the core of the activity of formative assessment lies in the sequence of two actions. The first is the perception by the learner of a gap between a desired goal and his or her present state (of knowledge and/or understanding and/or skill). The second is the action taken by the learner to close that gap to attain the desired goal. The learner first has to understand the evidence about this gap and then take action on the basis of that evidence. Although the teacher can stimulate and guide this process, the learning has to be done by the student. It would be a mistake to regard the student as the passive recipient of any call to action: there are complex links between the way in which a message is understood, the way in which that perception motivates a selection among different courses of action, and the learning activity that might follow. These arguments made it clear theoretically that the development of self-assessment by the student might have to be an important feature of any programme of formative assessment, a point that had already been illustrated in several of the research studies.

The third key idea was the concept of 'feedback'. This concept deals with a feature central to the operation of any system that has to adapt to manage change. The key components in the operation of feedback of any such system are:

- data on the actual level of some measurable attribute;
- data on the desirable level of that attribute;

- a mechanism for comparing the two levels and assessing the gap between them;
- a mechanism by which the information can be used to alter the gap.

With small changes of terminology, the above four steps could be a description of formative assessment. The last of these components is essential: if the information is not actually used in altering the gap, then there is no feedback. It is also clear that the quality of the feedback provided is a key feature in any procedure for formative assessment.

One of the most important reviews of the effectiveness of feedback was carried out by Kluger and DeNisi (1996). They reviewed numerous reports of the effects of feedback on performance and, after excluding those which did not meet their stringent criteria of quality, were left with 131 reports, yielding 607 effect sizes and involving 12,652 participants. They found an average effect size of 0.4, but the standard deviation of the effect sizes was almost 1 and about two effects in every five were negative. However, their definition of feedback required only the collection and reporting of the data. Where the procedures also involved ways of using the data to make improvements, the effects were all positive. The explanation of this difference is that when people are only told that they have done well or badly, it will affect their ego but it is not likely to improve their involvement with their tasks (this issue is discussed in detail in Chapters 4 and 5).

The prospects

The principles to do with teacher change, student change and feedback would clearly have to be borne in mind in any innovative development. While these pointed to several theoretical ideas that would be relevant, notably those concerned with theories of learning, theories of motivation and Sadler's analysis of the role of feedback, there was no comprehensive theory that could form a basis for action.

At the more directly relevant level of strategies and tactics for classroom work, the literature indicated that the choice of learning tasks, the quality of questioning, classroom discourse and the orientation of feedback on oral and written work, self- and peer-assessment and the use of tests were all issues that could demand attention. However, for none of these could one confidently set out a recipe for improvement, not least because their implementation within any comprehensive framework, and within UK classrooms, had not been studied.

Underlying these reservations was a more general caution about the 'transfer' of innovations, however well researched, into the daily practice of teachers. The mere collection and publication of research data is not enough, as was made clear in the following extract from *Inside the Black Box*:

> *Thus the improvement of formative assessment cannot be a simple matter. There is no 'quick fix' that can be added to existing practice with promise of rapid reward.* On the contrary, if the substantial rewards of which the evidence holds out promise are to be secured, this will only come about if each teacher finds his or her own ways of incorporating the lessons and ideas that are set out above into her or his own patterns of classroom work. This can only happen relatively slowly, and through sustained programmes of professional development and support. This does not weaken the message here – indeed, it should be a sign of its authenticity, for lasting and fundamental improvements in teaching and learning can only happen in this way.
>
> (Black and Wiliam 1998b, p. 15; emphasis in original)

Such arguments raised the wider issues of the transfer of research results into professional practice and of the nature of the involvement of teachers in such work. It was concluded that what was needed was to set up a group of schools, each committed to the development of formative assessment. In such a process, the teachers in their classrooms would be working out the answers to many of the practical questions that the research literature could not answer, and reformulating the issues, perhaps in relation to fundamental insights, but certainly in terms which could make sense to their peers in ordinary classrooms. It was envisaged that in such a programme the schools involved would need extra support, both to give their teachers time – to plan the initiative in the light of existing evidence, to reflect on their experience as it developed and to advise on training work for others in the future. In addition, there would be a need to support external evaluators to work with the teachers to help their development of the work and to collect evidence about its effectiveness. Such evidence would both help guide policy implementation and to disseminate findings to others.

In summary, it could be claimed that a firm case for a development programme had been made and that a basis for such a programme had been laid. However, it was equally clear that although the signpost had been set up on a road worth following, this work was only a first step along that road.

3 How teachers developed the ideas with us

The starting point

Given our commitment to undertaking development work to determine how formative assessment could be incorporated more effectively into professional practice, we had to find partners – local education authorities (LEAs), schools and teachers willing to take part in such a venture. We started by holding joint discussions with assessment advisory staff from Oxfordshire and Medway, chosen because we knew that key members of their LEA advisory staff would understand and support our approach and so might be willing to enter into a project with us. Thus Dorothy Kavanagh in Oxford and Rose Collinson and Sue Swaffield in Medway joined us in discussions which led to production of an agreed detailed plan. However, while they and a selection of their schools could take on the load of engaging in the work, and could commit some days of work of their advisory staff so that each authority would be fully involved, financial support was needed to release teachers for meetings and to make research staff available to observe and analyse the progress of the work.

So the next step was to make an application to the Nuffield Foundation for funding of a project. The proposal that we submitted to them set out the aims, the roles of the partners to be involved, the assumptions lying behind our approach and the timetable for the project.

The proposal and the partners

Our overall aims were to develop the implementation of formative assessment in the normal professional practices of teachers and to explore the advantages of such implementation. A related aim was to lay a basis for the design of programmes for wider dissemination of the findings and in particular to design in-service training (INSET) to replicate the implementation. The project method was to carry out an interactive INSET development programme involving three groups of partners: the teachers and their senior staff in the schools, staff at King's and the LEA advisory staff.

The role of the teachers was to plan and then to implement individual innovations in their classrooms, and then to help evaluate these, particularly by reflecting on their experience in developing formative assessment. The role envisaged for the staff at King's was, at the outset, to present ideas to the teachers and help them in designing and implementing their own innovations. Subsequently, they were to support and evaluate the processes of implementation and to use the findings in the design of dissemination and INSET work. Finally, the role envisaged for the advisory staff of the local authorities was, at the outset, to take the lead in the selection of the sample schools and in the negotiations with those selected. Subsequently, they were to share with the King's staff in the work of support, evaluation and dissemination

The proposal – assumptions

We thought it important to spell out, in our proposal, the main assumptions on which the project would be founded and organized. The first assumption was that existing research evidence had already established that development of formative assessment could produce substantial improvements in the learning of students and there was no need to repeat such work. However, there was a need to study how different teachers might realize such improvements within the normal constraints of curriculum and external testing requirements. We believed that any attempt to force adoption of a simple recipe by all teachers would not be effective, and that success would depend on how each could work out his or her own way of implementing change.

We judged nevertheless that existing research did provide some important guidance that would be helpful for all teachers. In particular, promotion of self-assessment by students, as a component of strategies to develop their capacity to take responsibility for their own learning, should be fundamental to the development of productive formative assessment. Part of our task would be to initiate the work by distilling and conveying such messages.

The Nuffield Foundation accepted the proposal with one reservation. We had not envisaged collecting any quantitative evidence of learning gains, since we judged that there was already adequate evidence that such gains could be achieved. However, the Foundation's referees advised that such quantitative evidence would be necessary to avoid the charge of 'but will it work here?', given that much (in fact, almost all) of the evidence cited by Black and Wiliam (1998a) was from outside the UK. So we agreed to collect quantitative evidence. This part of the project is described in the last section of this chapter.

The project started work in earnest in January 1999 and the funding supported work up to the end of the 1999–2000 school year. However, we subsequently entered into negotiation with colleagues at the School of Education in Stanford University who wished to develop a similar project with schools in California. They succeeded in obtaining funding for this from the US National Science Foundation. The King's team and the work in our schools were included in the project so that we were able to continue with full support until the summer of 2001. The contribution from King's was to inform the development at Stanford on the basis of our experience, and to develop further evidence in our schools about the dissemination of the project and about the interface between formative and summative assessment practices in our schools. We shall not in this book discuss the work in California.

The schools and teachers involved

Our proposal was that we would work with science and mathematics teachers. We believed that the detailed working out of new approaches would be different according to the nature of the school subjects. We therefore considered it likely that limited effort would best be invested in no more than two subjects – the two chosen were those in which the staff involved at King's had extensive experience. Earlier work at King's, notably in the Assessment of Performance in Science (APU; Black 1990), in the development of graded assessments in science (GASP project; Swain 1988) and in mathematics (GAIM projects; Brown 1989), and in work on assessment practices with science teachers (Fairbrother *et al.* 1994) provided important resources here. The experience with GASP and GAIM was that teachers achieved marked improvements in their assessment work, but that for many these were implemented only as frequent summative assessment. We thought that the obstacles that prevented many from achieving improvement in formative assessment could, in the light of our studies of existing research, be better understood so that we could foresee ways in which the new project could attempt to overcome these obstacles.

This restriction to areas in which we had extensive expertise also led us to work only in secondary schools. We could foresee that primary teachers, who guide the learning of students over a range of different subjects, would have problems and opportunities in formative assessment quite different from those of secondary teachers, and that these would have to be studied in a separate project.

For the secondary phase, we also envisaged that the work would be confined to years 7, 8 and 10 (i.e. to ages 11–12, 12–13 and 14–15 years). This was because it was likely that the pressure of external 'high-stakes' assessments would inhibit the development of formative assessment, and so the 'pressured' years 9 (key stage 3 testing) and 11 (GCSE) should be avoided. It was nevertheless clear that the interplay of the formative and summative assessments would be a factor in every year, but we judged that we might study this in those years when the summative aspects were largely within the control of