Assessment in Education
Vol. 12, No. 1, March 2005, pp. 21–38

'They don't give us our marks': the role of formative feedback in student progress

Emma Smith* and Stephen Gorard
University of York, UK

This paper presents the results of an experimental evaluation of a change in assessment practice in one comprehensive secondary school. The school divided 104 Year 7 pupils into four mixed-ability teaching groups. One of these was given enhanced formative feedback on their work for one year, but no marks or grades. The other three groups were given marks and grades with minimal comments, which was the usual prior practice in this school (and many others). Using data derived from assessment, prior attainment, pupil attitudes and background information, we conducted a contextualized analysis of progress in the four teaching groups for all subjects. This showed that progress in the treatment group (formative feedback only) was substantially inferior to that of the other three groups. In this paper, we also use data from observation of the process and from group interviews with the students involved, to help explain these results. Our findings are relevant to a consideration of the often lessened impact of research findings when 'rolled' out into wider practice, and what may be done about this.

Introduction

We were invited by a secondary school from the Welsh coalfield valleys to assist them with evaluating a small-scale trial of a different method of providing feedback to students. In particular, they were concerned to monitor the medium-term impact of changing their marking and assessment policy by removing the use of summative marks and grades but enhancing the use of formative feedback to students. Rather than introduce the practice across the entire school, they elected to pilot it in one year-group, parallel to a wider series of staff training sessions. The school's intention was that an intervention trial be carried out with what they termed 'mixed-ability' teaching groups. This paper describes a small-scale experiment to study the impacts of formative assessment techniques on the progress of students in Year 7. In particular, we

*Corresponding author. Department of Educational Studies, University of York, Heslington, York YO10 5DD, UK. Email: es25@york.ac.uk

ISSN 0969-594X (print)/ISSN 1465-329X (online)/05/010021–18

DOI: 10.1080/0969594042000333896

were keen to determine any differences the intervention might make in terms of pupil progress towards Key Stage 3, as well as to understand how the pupils reacted to the changed form of assessment. We start with a brief recapitulation of the argument for greater use of formative assessment, before explaining the methods used to evaluate its impact in one school.

The role of formative teacher-led assessment

In 1988, the National Curriculum Task Group on Assessment and Testing (TGAT) recommended that assessment should be an 'integral part of the education process, continually providing both "feedback" and "feedforward" and ought therefore to be systematically incorporated into teaching strategies and practices at all levels' (DES, 1988, paragraphs 3 and 4). The system that the group advocated combined both formative and summative approaches. However, ensuring parity of esteem for both teacher-led formative and summative assessment has been a contentious issue since the early days of TGAT. According to some commentators, this derived, at least in part, from 'political concerns for accountability, rather than by educational concerns for learning' (Torrance & Pryor, 1998, p. 10). Others have been concerned about the lack of attention and resources that had been given to developing the formative assessment process (Shorrocks-Taylor, 1999). Indeed, the first documents to provide any guidance on teacher assessment appeared in schools 15 months after teachers had started their assessments with students (Daugherty, 1995).

While the positive role of formative assessment has now apparently been widely accepted in mainstream education circles there has been, according to Black (2000), an absence of a coherent programme of research to underpin both the theoretical and practical development of the formative assessment process. For example, a review of research into formative techniques found that key work in this area showed little overlap and collaboration—'it seems that most researchers are not studying much of the literature that could inform their work' (Black, 2000, p. 409). This difficulty is compounded by different notions about what the term 'formative assessment' actually means (see, for example, Harlen *et al.*, 1992 or Stobart & Gipps, 1997). A conflation of both the terms and ideas underpinning 'formative' and 'teacher assessment' has led to the gradual replacement of the term 'formative assessment' in the lexicon of assessment policy. According to Daugherty, 'the number of references to formative assessment in official documentation about national assessment seems to have declined steadily from the TGAT report onwards' (Daugherty, 1995, p. 74). Indeed, in its attempt to clarify the role and status of both types of assessment, the 1993 Dearing Review (Dearing, 1993) identified only two types of assessment: those made by the teachers themselves (teacher assessment) and those provided by the results of national tests (Shorrocks-Taylor, 1999).

Thus, formative teacher assessment has had a somewhat limited and indirect role in the evolution of National Curriculum testing. Its original conception through TGAT was as a major part of a two-pronged national assessment system whose commitment was to 'formative assessment as the best way of achieving this raising of

standards, tempered by the recognized need for valid and robust summative approaches' (Shorrocks-Taylor, 1999, p. 165). Its current position is on the fringes of the national testing programme where, although schools are still required to report teacher assessed levels at the end of each Key Stage, its status is now far from what TGAT had envisaged.

One possible consequence of this shifting position of formative and summative techniques is that, on the one hand, we have a system of high stakes summative national testing that is unique among European countries; and on the other, a 'crisis account' of low standards and failing students (Gorard, 2001). This might lead one to consider that the current testing regime has proven unsuccessful in some respects, and that, ironically, the assessment method that *could* have produced the continuous improvement in student performance apparently demanded by Government might more properly be formative, rather than summative.

The success of formative assessment techniques in raising the achievement of students, particularly those of lower ability—who, we are told, constitute the UK's 'long tail' of underachieving students—has been demonstrated by recent research carried out, in particular, by the King's College Formative Assessment Group. The evidence that this research has produced to demonstrate the efficacy of formative techniques is compelling. Indeed Black (2000) cites research where the use of formative assessment techniques produced learning gains with effect sizes of between 0.4 and 0.7, larger than those produced by some other significant educational interventions. In addition to its potentially positive impact on learning, the research carried out by the Formative Assessment group also provides us with a valuable example of how academic research has been translated into practice and adopted by schools across the country. The popularity of formative assessment techniques among practitioners, as evidenced by its prevalence in in-service training courses and sales of the pamphlet *Inside the black box* (Black & Wiliam, 2001), should provide researchers with valuable lessons in how academic research *can* be translated into classroom practice (Black & Wiliam, 2003). Indeed, it was an in-service training course on formative assessment techniques that prompted the experimental evaluation described in this paper, the lessons from which are relevant to our understanding of what happens to complex interventions when they move from a research setting inhabited by pioneers and enthusiasts to wider practice.

The study

Students were allocated to one of four mixed-ability teaching groups of 26 pupils each, on the basis of prior attainment, teacher recommendation from primary school and friendship groupings. Although termed 'mixed-ability', the groups mainly comprised students of mid-range ability as, in this school, the most able students and those with Special Educational Needs were largely taught in separate groups. Because the groups were not allocated randomly, we needed to compare carefully the background, motivation and prior attainment of all pupils to provide a context for any later differences in outcome. Of these four groups, three continued with the existing school

policy of allocating marks and grades to pieces of assessed work, with minimal associated comments from the teacher, throughout the year. The fourth group, designated the treatment group, did not receive any summative marks or grades for any work at all (even where these were generated for other purposes). Instead, their teachers agreed to provide more careful, individual formative feedback of the kind that makes the assessment process also a learning process. Staff were given appropriate in-house training.

In order to assess the impact of the intervention on pupil progress, we collected a considerable amount of data about the students in each of the four mixed-ability teaching groups. This data included prior Key Stage 2 teacher assessments and test levels in the core subjects, as well as scores on standardized tests such as the Cognitive Ability Test (CAT) and Progress in English (PiE). In addition, the school provided a list of the students eligible for and taking free school meals. A questionnaire was administered to all groups during the second half of the autumn term. Responses were received from 104 pupils (100% response rate). The purpose of the questionnaire was to elicit further background information from the students in areas which recent research has suggested may have an impact on school performance, such as aspects of student motivation and attitudes towards school, and background variables such as parental employment and student's ethnic group (Smith, 2003). The variables included:

- sex
- ethnicity
- family type, and the number and order of siblings
- date of birth
- parental occupation
- attitudes towards school (including having a sense of belonging to school, teacher student relations, pressure to achieve and teacher support)
- home environment (such as family wealth, educational resources and communication with parents)
- attitudes towards learning (including interest and self-concept in maths and reading)
- strategies for learning (and the amount of effort they are prepared to put into work).

The findings from this questionnaire, and the background and prior examination data, give an indication of prior differences between the four groups and are used as context for our analysis of posterior differences in test scores via linear regression. The base-line data was used to try and control for raw-score differences between the treatment and control groups, in order to isolate the impact of the intervention on students' performance. The outcome measures against which the efficacy of the intervention was measured were student performance at the end of Year 7. These consisted of National Curriculum (NC) style levels which were awarded by the class teachers on the basis of assessments undertaken throughout the school year (but with results not fed back to the treatment group), and summatively at the end. Note that,

as all results refer to a year group in a self-selected school and the groups were not allocated randomly, probability-based tests of the significance of differences would be meaningless, and are avoided throughout the paper.

Unstructured group interviews were also carried out with students in the treatment group. The purpose of these interviews was to gain an understanding of the students' experiences of this intervention, particularly the impact of receiving comments rather than marks. The interviews were carried out with two groups of six students towards the end of the spring term. The students were selected at random from the treatment group. Although parental permission was received for the meetings, the students were not told the purpose of the interviews until afterwards. Our approach was to find out whether general discussion about their enjoyment of school would prompt the students to discuss the intervention, without the need to be directed to do so by the researcher.

Characteristics of the students

The tables below show that the treatment group (given formative feedback) and the control group (given marks and grades) were quite similar in terms of all variables relating to family background, attitudes towards school and prior attainment. The treatment group had fewer males, fewer unemployed parents, more fathers in social classes 1 and 2, and better KS2 results for English and maths. They were, in addition, seven days older on average than the control group. The control group had fewer students on free school meals, more mothers in social classes 1 and 2, and better KS2 results in Science (Table 1). Overall, the groups are fairly well balanced.

Table 2 also shows that there was very little overall difference between the groups for both prior attainment measures and responses to the questionnaire. See Gorard and Smith (2004) for a more detailed discussion of the student characteristics.

Table 1. Percentages of students in each group with relevant characteristics

	Control	Treatment
Male	47	42
FSM	28	31
Father unemployed	47	23
Mother unemployed	66	58
Father class 1 or 2	30	35
Father class 3	29	54
Mother class 1 or 2	24	19
Mother class 3	33	46
KS2 English level 3	22	25
KS2 Maths level 3	29	33
KS2 Science level 3	15	8

Note: All cases reported ethnicity as 'white'. The KS2 results are the highest level for either TA or TT, or the only level where students were unrecorded for one but not the other. Cells are percentages.

Table 2. Mean scores of selected characteristics for students in each group

	Control	Treatment
KS2 English mark	55	53
CAT verbal	87	85
CAT quantitative	93	91
CAT non-verbal	90	92
CAT average	90	89
Interest in maths*	5	5
Interest in reading*	5	6
Self-concept (academic)*	6	6

*items derived from student questionnaire.

The descriptive analysis of the many variables from the questionnaire and academic performance data, only a few of which are presented here, suggests that, overall, there is very little to distinguish between the treatment and control groups in this study. They provide a sound basis for further analysis of the relative progress made by these groups in light of the 'black-box' style intervention. These differences will, anyway, be taken into account in our value-added analysis of the progress made by both groups.

The summer results

Having considered the background and prior attainment of both groups, we would expect their summer results to be very similar, with the treatment group perhaps slightly ahead if formative assessment has a beneficial effect. Table 3 shows the percentage of each group obtaining each National Curriculum level for the four core subjects in Welsh schools. These are raw-score figures not yet contextualized by the findings above.

The control group has superior outcomes for the three core subjects of English, maths and Welsh. There is no clear overall difference in science. Given that there is little in the background or prior attainment of the two groups to explain these differences, we would also expect our value-added analysis to show greater progress overall in the control group. A similar pattern was present for the other nine foundation subjects (see Gorard & Smith, 2004 for details).

Contextualized results

Here the analysis will focus on measuring the efficacy of the intervention. By taking into account prior attainment and any contextual differences between the treatment and control groups, and comparing the predicted and actual performance in the summer assessments, we can describe the relative progress made by the students and therefore the effectiveness of the type of assessment used with each group.

Table 3. Percentages of students in each group obtaining each NC level (core subjects only)

	Control	Treatment
English level 3	28	35
English level 4	47	65
English level 5	20	–
English level 6	4	–
English level 7	1	–
Maths level 2	–	4
Maths level 3	29	28
Maths level 4	62	64
Maths level 5	8	–
Maths level 6	1	4
Science level 3	13	35
Science level 4	76	39
Science level 5	11	27
Welsh level 2	12	35
Welsh level 3	46	58
Welsh level 4	21	8
Welsh level 5	17	–
Welsh level 6	4	–

Note: The lowest reported level for each subject includes all students obtaining that level or less. The highest reported level for each subject includes all students obtaining that level or more.
Caution: While the scores are expressed here as percentages for convenience, there are in fact less than 100 cases per group.

A linear regression model was created for each school subject, using the summer results as a numeric dependent variable, and the available numeric background and prior attainment scores as predictors. In this paper, only the results for the core subjects are given. In the regression model, any missing values were replaced by the overall mean score where appropriate; this enabled the inclusion of the questionnaire items in the model. Where appropriate, also, categorical variables such as sex were coded as dummy binary variables.

The R-scores appear in Table 4. As can be seen, the correlations between the predictors and the summer results are relatively strong for English, maths, and science, explaining over 50% of the variation in outcomes.

The next section displays the standardized differences (residuals) between the best statistical prediction of the summer results for each student, and the levels actually 'awarded'. For the purposes of this analysis, we have treated the ordinal variable

Table 4. Multiple correlation scores between background, prior attainment and subject results

	English	maths	science	Welsh
R	0.70	0.73	0.75	0.64

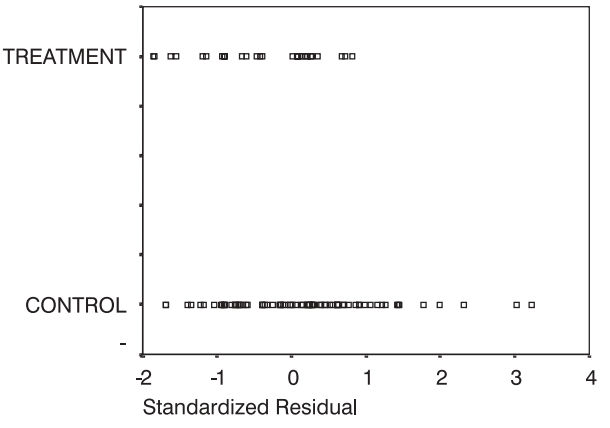

Figure 1. Comparison of standardized residuals in English between treatment and control groups (*Note*: the top line shows the scatter of residuals for the treatment group, and the bottom line shows the control group.)

representing NC levels and grades within levels as though the intervals were of equivalent size (they may not be). This should add to the caution with which the results are treated.

English results

Figure 1 reveals that the residuals for English are markedly superior for the control group. A far higher proportion of pupils in the control group have positive residuals (i.e. appear to have performed better than was predicted by the best model involving all prior data), and this group also has the highest residuals. The treatment group has the most negative residuals (i.e. appear to have performed worse than predicted). We have no evidence that the intervention was successful. On the contrary, we have some evidence that it was actually harmful to the results in English for this group.

Unsurprisingly, the key predictors of later performance in English include the score for Progress in English (PiE), academic self-concept and for interest in reading (both for the pupil questionnaire). More surprising is that it is those who score lowest on interest in reading and those with the lowest academic self-concept who show most progress. Those with an employed father also show more progress than those without.

Maths results

Figure 2 suggests a somewhat superior distribution of residuals for the control group in maths. This group has the highest residual and the treatment group the lowest, but the overall difference is not as clear as for English. What is clear is that there is, again, no indication that the intervention has been positively effective.

Key predictors of later performance in maths include Key Stage 2 scores in science and English, nonverbal and quantitative Cognitive Attainment Test (CAT) scores,

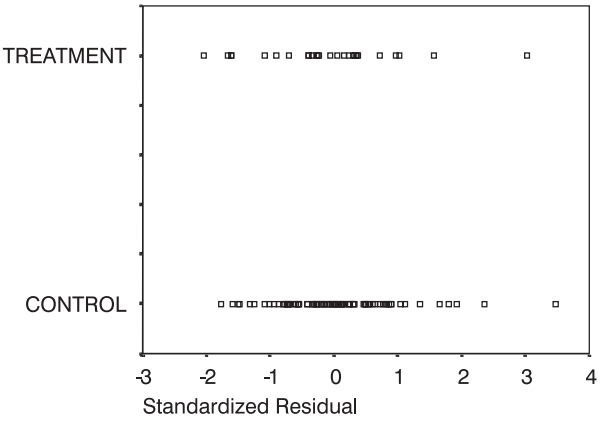

Figure 2. Comparison of standardized residuals in maths between treatment and control groups

and reading self-concept. Those from families with higher family wealth scores, and with fathers in professional or non-manual occupations also show greater progress.

Science results

There is no clear difference in the *quality* of the residuals for science (Figure 3), but there is a difference in their distribution. The control group are more homogeneous, while the extreme scores are in the treatment group. It is not clear which of these situations is educationally preferable, given that the proportions of pupils attaining the 'expected level' is the same for both groups. Perhaps the only conclusion to be drawn is that, again, we cannot say that the treatment has led to any clear improvement.

The key predictors for progress in science include Key Stage 2 English score, the quantitative CAT score, free school meals (those eligible have lower scores, on

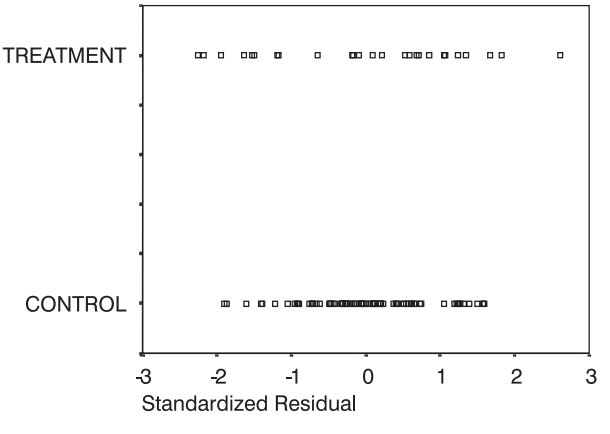

Figure 3. Comparison of standardized residuals in science between treatment and control groups

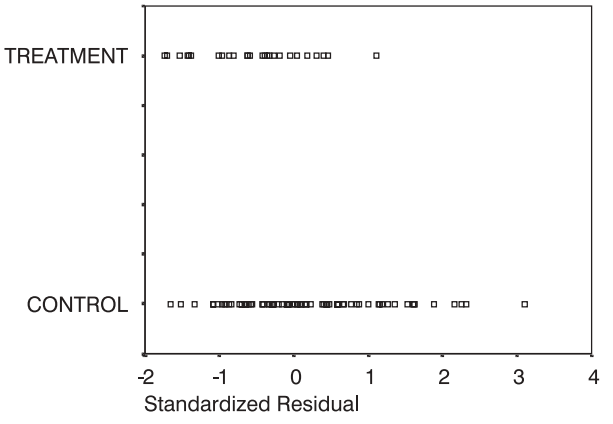

Figure 4. Comparison of standardized residuals in Welsh between treatment and control groups

average), and the instrumental motivation, teacher–student relations and belonging-to-school variables from the questionnaire.

Welsh results

As with English, the treatment group in Welsh have markedly inferior residuals, the lowest residual and a far lower average (Figure 4). It seems that were we to conclude that the treatment had been effective then its effect has been harmful to the students in this school, year and group.

Progress in Welsh is indicated by several variables including Key Stage 2 English scores, and the effort and persistence, teacher support, academic self-concept, and family wealth variables from the questionnaire. Perhaps more intriguingly, progress in Welsh is also predicted by lower Key Stage 2 science scores, and a lower mathematical self-concept.

Of the four core subjects, English, maths and Welsh show greater progress for the control group and science shows no clear difference. We would have to conclude, therefore, that the treatment has been ineffective (or worse) as a method for improving student learning. Overall, the value-added scores were better for the group receiving formative feedback in one of the thirteen subjects considered, the same in seven subjects, and worse in five (see Gorard & Smith, 2004). Even more worryingly, the subjects with the more negative residuals include three of the core subjects. The final section uses our own observations, and the student's experiences of the intervention to try and understand this somewhat surprising result.

Experiencing the intervention

We present here a commentary on three themes from the group interviews. Since the interviews were deliberately unstructured, there was only the occasional 'steer' rather

than a schedule, and the interviewee comments did not necessarily appear in this order:

- the students' general opinions about the project
- their opinions about the type of comments they receive on their work
- the wider implications of the study, for example the attitudes of their family and their friends in other classes towards the project, as well as recommendations the students would make for any changes to the strategy.

Being in the treatment group

The students were not told about our underlying interest in the nature of the intervention; the discussion was simply described as an opportunity to talk about their experiences in a new school. The discussion was initiated by the students being asked general questions about what they liked and disliked about their new school. They described the opportunities to make new friends, extra-curricular activities, as well as certain lessons, as aspects of their new school that they enjoyed. Smelly toilets, particular lessons and being 'pushed out of the way' by older students were commonly viewed less favourably. Asking about their experiences in their particular form group prompted the following responses:

> I would prefer to be in another form because we don't get our test marks back.
>
> They don't give us our marks. It's supposed to be some sort of project because…we don't have our marks…to see how we react and that.
>
> We only get comments like and we don't get our marks back
>
> Sometimes we do get our marks though… when like the teacher forgets. Then they ask for them back because we are not supposed to have our marks.

It is important to note a possible 'contamination' of the evaluation here, as we can have no real idea of the scale of errors such as teachers issuing marks to the treatment group inadvertently. The issue of *not* receiving marks was a concern for both interview groups, and was introduced into the discussion by the students themselves. When asked how they felt about being involved in this project there was a range of responses, but many students had fairly negative views, particularly because, in their opinion, not receiving marks gave them little idea about how to direct their efforts. For example:

> We feel different from the rest because they have their marks
>
> It is kind of good because it makes us feel special
>
> It doesn't make any difference, we feel normal…
>
> I disagree because we won't know what group we will be in next year.
>
> I disagree as well because you don't know whether to do better then.
>
> If they tell you your mark you can try to get higher the next time.

The class's involvement in the project had been explained fully to them by the assistant headteacher:

> We were in French and Mrs B came to us and said we were doing a project on you and we'll see how it goes and then come back later in the year.

However, several of the students also reported that some of their peers were confused by the new arrangements:

> They get confused because they don't get their mark and they ask the teacher, what's our mark? Because they forget and get confused that we just get comments.

It seems that although the students interviewed did not feel that they themselves had been 'picked upon' (i.e. disadvantaged) by their special treatment, there was some confusion within the group (and also apparently among teaching staff) about the aims and nature of the project.

Researchers who have studied formative assessment techniques and their impact on the learning process, recommend that schools and teachers should consider changing four aspects of their work: question-and-answer interactions, peer and self-assessment by pupils, marking homework, and involving the pupils in setting and marking their own tests (Wiliam & Black, 2002). Implementing these 'assessment for learning' strategies should be considered as a long-term initiative, which 'calls for changes in their (teachers') practice which are radical in scope and nature' (Black, 1998, p. 112). As such, it would seem that an initiative, which, in its very essence, transforms the learning and assessment process by transferring some of the control for learning from teacher to student, ought to be introduced with the full cooperation and understanding of all the students and teachers involved. The comments above suggest that perhaps this was not the case in the present study.

The formative feedback

According to Stobart and Gipps (1997), assessment can only be formative if it 'feeds back into the teaching-learning process' (p. 18). They argue that, in order for students to improve, effective feedback should enable the student to know exactly what they would have to do to close the gap between actual and desired performance—'the use of grades or 7/10 marking cannot do this' (p. 19). The use of 'rich' feedback is key to the interventions used by the King's College Formative Assessment Programme for helping teachers improve the day-to-day assessment of their students (Black & Wiliam, 2001). Their research suggests that 'feedback has been shown to improve learning where it gives pupils specific guidance on strengths and weaknesses, preferably without any overall mark' (p. 8). The use of written comments on students' work, rather than marks, was the key method of feedback that differentiated the treatment from the control groups in the present study. This is what the staff development associated with the intervention was intended to encourage.

When asked whether the comments they received were useful, the majority of students felt that the comments did not provide them with sufficient information so

that they would know how to improve. They did not believe that the simple act of awarding marks would stigmatize poor performers.

> Miss, I'd like to know my marks because comments don't tell us much.
>
> Well sometimes when you have marks, you have comments anyway. I would rather have both of them.
>
> I don't like it when we just have comments. I'd rather marks because you don't know how to beat your score. And I probably end up doing worse than what I am if I have my mark.
>
> I reckon we should all have our test marks because when we had our SATs all our class was supportive to us. Because like if someone had 544 and someone else had like 444, they say 'oh you did good anyway'.

One of the students produced his English book to demonstrate how the comments were being used:

> Interviewer: Can I read it out? It says 'an excellent effort, your poem tells us lots about your personality and is beautifully presented, a special effort'. Did that comment make you feel better than if you had just had 10/10?
>
> On that piece yes, but some pieces of my work I would rather have scores.
>
> Like stories…
>
> I think that like you put all your effort in like the write a story and then all you do is get told like 'oh very good', miss. But if you have marks then you know how good you really did.

This example of the 'formative' comments suggests that some, at least, were no more than would traditionally have been given as an adjunct to a mark or grade. The desire of many students to receive marks was such that several admitted to trying to work them out. This was particularly so in subjects like maths and languages when students admitted to adding up correct spellings in vocabulary tests in order to work out what marks they had received. In the words of one of their teachers: 'they were gagging for their marks'.

> When we are given back our tests in science, I do count the marks up so I know my marks.
>
> Sometimes in English we do get levels…
>
> …yes they are helpful.

A minority of the group did feel that having comments was helpful, because they would 'explain more than marks', but their support was somewhat lukewarm:

> I think it is… if it says try harder, I will just try harder and harder again until I get better
>
> You know when you have comments you can work it out, what is happening. If you like get 50%, you don't know whether it is good or bad but if you get a comment you can try harder next time.
>
> …they tell us what we have done wrong. I think we should have the mark like but if we have done say 7 out of 10, they should say how to improve it.

However, a consensus was reached between those who preferred marks and those who preferred comments when it was suggested that they might receive both marks and comments on their work:

> I think people would prefer that then because they would read the comments and look at their marks. And then they could understand the comment.
>
> I think that if we had a mark Miss, and we done like bad then they could give us a comment with it as well, how should we improve like. So marks and a comment, I think.
>
> If they write 'improve more' you don't know your score so you can improve.
>
> If you had like 4/10 and a comment then you would know how to improve.
>
> I think most of us in our class would like that.

Although this strategy of combining marks and comments might be welcome to the students, it does not, according to the King's College researchers, produce any clear improvement in performance: 'when students get marks and comments, they first look at their own mark and then at their neighbour's. They hardly ever read the comments' (Wiliam & Black, 2002, p. 8). What is crucial, however, is the *quality* of the comments and feedback given. As suggested earlier, this should provide students with guidance on how to improve, as well as opportunities and support to understand how to make the improvement. Asking these students about the types of comments their teachers write on their work prompted the following exchange, which was typical of the responses received from the students interviewed:

> …try harder next time.
>
> …try and improve.
>
> Interviewer: Do they tell you how you could try harder or improve?
>
> Yes, by writing the sentence in full.
>
> They just write 'very good' on the piece of work.
>
> They don't really explain it.
>
> They just say like 'very good' and that is it.
>
> Sometimes I get confused with it because some of my comments, well I just don't really get what they mean to say. They are confusing.
>
> Interviewer: Do you think the comments could be clearer?
>
> Yes, if they did it like that then I think more people would prefer that.
>
> Sometimes when they write you can't understand their writing.
>
> Yes it is a bit scrawly, like spider writing.
>
> In English if we have a test, a different English teacher marks them and last test we had most of us couldn't read her writing.
>
> Interviewer: Are some teachers better than others at writing these comments?
>
> No they just write 'very good' on the piece of work.

> They don't really explain it no.
>
> They just say like 'very good' and that is it.
>
> Sometimes I get confused with it because some of my comments, well I just don't really get what they mean to say. They are confusing.
>
> Interviewer: Would you go and ask the teacher to explain?
>
> Yeah, but then I still don't always get the comment

It is hardly surprising that the intervention produced no improvement in the treatment group if these examples really were typical. The pupils had simply had their marks and grades removed (most of the time). Substituting marks with comments on students' written work is the only one out of the four *'Black box'* assessment strategies adopted by the school, and recommended by the researchers. Although we cannot rule out the fact that individual teachers might have adopted self and peer assessment as part of their teaching and learning strategy, there was little to suggest that it was a co-ordinated part of the intervention undertaken by all of the teachers involved. Certainly, the students' views suggest that the defining difference between the treatment and the other groups was the fact that they received comments on their work rather than marks. That these comments were of little use, at least in the students' opinion, is evident from the exchanges above.

Wider implications

Many of the students' peers in other classes were aware of the project but, according to the students interviewed, they did not really understand its purpose. Some of the students mentioned that their friends thought it was a 'bit weird', but the situation was not highlighted as one that was of real concern to the group. This suggests that piloting the intervention in similar school settings might not adversely affect the peer relationships of any treatment group. On the other hand, it resulted, to some extent, in the treatment group being excluded from discussions in which their peers compared results in recent tests. Several of the students expressed frustration at not being able to communicate their progress in school to their parents:

> It is frustrating, when I was in primary school, I was used to going home and telling my mother the scores and now like it's harder because we can't talk about marks, it's like comments now so I try to read the comments out to my mother.
>
> I don't think we should have comments through the whole school I think we should just have marks because when your parents ask you what you have had in your test you can't really explain.

Although the students felt that not receiving marks prevented them from discussing their progress in school with their parents, research carried out in schools by the others has found that parents were often positive about the initiative, suggesting that comments gave them a clearer understanding about what they could do to help their children (Wiliam & Black, 2002). In this study, communication about progress in

school between student and parent appears to be largely a case of reporting results, rather than a situation where the parent was involved in diagnosing what the student would have to do to improve.

When asked what the school should do if it was to repeat this project with another year group, the students felt that it would work best if the project involved the whole of the year group, rather than just identifying a single class.

> I think that all the year 7s should do it because then no-one would feel like left out and miss, when they say 'what's your marks?' then they won't say 'ah like I had higher marks than you'. And they can't make fun of each other miss.

There was also a general feeling that the aims of the intervention should be explained clearly to all of the students involved. This included using these students to explain what happens to their peers.

Although the students interviewed were positive about their overall experience in their new school, they were acutely aware of their involvement in a project which, to their minds, involved replacing marks on work with comments—the only part of the recommended strategies for effective formative assessment that the school appears (at least to the students) to have adopted. It was suggested by the interview responses that any feedback that was provided was often poorly understood by the students and did little to enhance the learning process. Where comments were made, they appeared to focus upon enhancing self-esteem or self-image rather than on what needed to be done to improve and how the student might go about making any improvement. The researchers were not, however, able to analyse the quality of the comments received by the control and treatment groups—this was beyond the scope of the study. It is also important to remember that prior to this intervention, the students in the treatment group had always received, and expected to receive, marks at the end of a piece of work. This abrupt change further differentiates them from the control group and could be another factor in contributing to their present dissatisfaction with the new assessment technique.

Nevertheless, the confusion and possible lack of motivation that the intervention had on the students, coupled with the apparent negative impact on their assessment performance, suggest that there are important lessons to be learnt for schools and researchers about how research-based designs fare in the 'real world' and when control is passed from the designers to the practitioners alone.

Conclusion

We reiterate that this is a small study, set in one school, without complete isolation of treatment students and staff from the controls, without representing the full ability range in the Year 7 cohort, and without a standardized 'public' test as an outcome. Therefore, in no way does this work, of itself, seek to test the overall notion that formative feedback is to be preferred in the ways suggested by its advocates. Nevertheless, it does give an indication of what can happen when a scheme is 'rolled out' into wider practice. In the original study (Black & Wiliam, 2003), the six participating

schools were given greater support and attention by the researchers than was possible here. Despite the initial enthusiasm among the staff in our study school, the interviews with students portray misgivings about this approach or, at least, about how it was implemented in this instance. It is quite common for educational and other interventions to work better in the pioneering study than in more general practice. This is for a variety of reasons—experimenter and Hawthorne effects, regression towards the mean, motivational factors, and misunderstandings.

Nevertheless, we must conclude on the basis of the evidence here that, in this school, the approach adopted for the treatment group was ineffective overall, and somewhat unpopular with the students as well. In fact, given the contextualized and value-added nature of the analysis we can conclude that the treatment group were actually at a disadvantage compared to their peers. This throws up an important concern for advocates of evidence-based policy and practice. Even where, as here, there is reason to believe that that original research has produced an intervention capable of leading to student improvement (and that in itself is rare), its wider application by policy-makers or practitioners can lead, inadvertently, to student and school dis-improvement.

Notes on contributors

Emma Smith is a Research Fellow in the Department of Educational Studies at the University of York. Her research interests include underachievement and differential attainment in secondary schools, educational inequalities, school outcomes, educational assessment, gender and education, issues of teacher recruitment and retention.

Stephen Gorard is Anniversary Professor of Educational Studies, also at York. His research is focused on issues of equity, especially in educational opportunities and outcomes, and on the effectiveness of educational systems. Recent projects include widening participation in learning, market forces and school compositions, underachievement, teacher supply and retention and developing international indicators of inequality.

References

Black, P. (1998) *Testing: friend or foe? Theory and practice of assessment and testing* (London, Falmer Press).

Black, P. (2000) Research and the development of educational assessment, *Oxford Review of Education*, 26(3&4), 407–419.

Black, P. & Wiliam, D. (2001) *Inside the black box: raising standards through classroom assessment*, King's Assessment for Learning Group, King's College London. Available online at: http://www.kcl.ac.uk/depsta/education (accessed November 2003).

Black, P. & Wiliam, D. (2003) In praise of educational research: formative assessment, *British Educational Research Journal*, 29(5), 623–638.

Daugherty, R. (1995) *National Curriculum assessment: a review of policy 1987–1994* (London, Falmer Press).

Dearing, R. (1993) *The National Curriculum and its assessment* (London, SCAA).
DES (1988) *National Curriculum: Task Group on Assessment and Testing: a report* (London, DES/ Welsh Office).
Gorard, S. (2001) International comparisons of school effectiveness: a second component of the 'crisis account'? *Comparative Education,* 37(3), 279–296.
Gorard, S. & Smith, E. (2004) *The role of feedback in student progress: a small-scale experiment,* Occasional Paper 59 (Cardiff, Cardiff University School of Social Sciences).
Harlen, W., Gipps, C., Broadfoot, P. & Nuttall, D. (1992) Assessment and the improvement of education, *The Curriculum Journal,* 3(3), 215–230.
Shorrocks-Taylor, D. (1999) *National testing: past, present and future* (Leicester, The British Psychological Association).
Smith, E. (2003) Understanding underachievement: an investigation into the differential achievement of secondary school pupils, *British Journal of Sociology of Education,* 24(5), 575–586.
Stobart, G. & Gipps, C. (1997) *Assessment: a teacher's guide to the issues* (London, Hodder and Stoughton).
Torrance, H. & Pryor, J. (1998) *Investigating formative assessment: teaching, learning and assessment in the classroom* (Buckingham, Open University Press).
Wiliam, D. & Black, P. (2002, October 4) Feedback is the best nourishment, *Times Educational Supplement,* special supplement: 'Mind Measuring', pp. 8–9.