'Feedback', the Toolkit and teachers' workload

by Dr Terry Wrigley, Visiting Professor, University of Northumbria

England's schools are governed by a unique mix of surveillance and numerical performance indicators. Statistical calculations compare achievement and schools, often unfairly, while teachers are kept in a permanent state of anxiety by the risk of inspection and performance review.

Within this system built on fear, headteachers made anxious about the possibility of a poor year's results or a negative Ofsted feel compelled to create the impression that everything possible is being done. The head's attempt to Ofsted-proof the school can lead to impossible workload demands.

This climate of fear has become known as *performativity*:

> the uncertainty and instability of being judged in different ways, by different means, through different agents; the 'bringing off' of performances – the flow of changing demands, expectations and indicators that make us continually accountable and constantly recorded…. It is a recipe for ontological insecurity… how will we measure up?
>
> …Here then is guilt, uncertainty, instability, and the emergence of a new subjectivity – a new kind of teacher. [1]

It is no surprise, then, that teachers' working hours have gone through the roof because of the time spent preparing and assessing – or rather *showing* that you are. In some schools teachers have been expected to write out and file every lesson plan, provide proof of marking in a particular way, and keep copious records.

The latest government survey[2] reveals that primary school teachers spend nearly 11 hours preparing and 10 hours assessing and reporting; in secondary schools, 9 hours on each. An extra 3-4 hours are shown as 'general administration', mostly linked to preparation or assessment (eg organising resources, record keeping).

This has nothing to do with real school improvement: the increase is due to the regime of fear and the performativity pressures – creating the impression that everything is being perfectly managed.

Ofsted have at last responded to union pressure and clarified[3] that they no longer expect to see individual lesson plans, nor 'unnecessary written dialogue between teachers and pupils in exercise books'. This is a relief, but it is Ofsted itself which generates the fear, and primary teacher Jack Marwood[4] claims that practices such as 'triple marking' have continued. The spectre of Ofsted can render even the most positive interventions toxic, let alone those which are intrinsically problematic.

The toolkit

One of the recent sources of workload pressure arose, strangely, from the prominence given to *Feedback* in the Education Endowment Foundation (EEF)'s *Toolkit*. In fact many effective kinds of feedback do not add to teachers' work, for example a few words of guidance to a student during a classroom activity, but fearful heads have felt the need to have *visible* evidence of feedback in case the inspectors call.[5]

This is a classic case of fear leading to knee-jerk responses based on inadequate understanding. The EEF Toolkit is actually a crude device designed to offer quick-fix remedies to low attainment, based on

inappropriate statistical procedures.[6] It claims to identify the cheapest and most effective interventions to help 'close the gap'. It is the latest addition to England's complex system of 'governance by numbers' (league tables, RAISEonline, Ofsted grades, SATs and 5 A*-C scores, floor targets etc.)

Its methodology is based on throwing together and averaging hundreds of *meta-analyses*summarising many thousands of pieces of research to measure the effectiveness of interventions. This is like claiming that a hammer is the best way to crack a nut, but without distinguishing coconuts from peanuts, or sledgehammers from the inflatable plastic hammers you win at the fair.

Top of the list comes something loosely called *Feedback*, which appears particularly cost-effective because schools don't have to pay for it: teachers simply stay up later marking books.

Meta-analyses are used in Medicine to enable researchers to complement the reading of other research, though not to substitute for it; for example, if experiments have been based on small samples, averaging the results can suggest a general trend.

But the medical literature contains serious warnings against the misuse of meta-analysis. Statisticians are warned not to mix together different treatments, types of patient or outcome measures – the 'apples and pears' problem. If the original results differ strongly, they are advised to highlight the difference, not provide a confusing average.[7] This is exactly what has **not** happened in the *Toolkit*, which should never have provided an average score for "Feedback" since the word has so many meanings:

'A teacher or parent can provide corrective information, a peer can provide an alternative strategy, a book can provide information to clarify ideas, a parent can provide encouragement, and a learner can look up the answer to evaluate the correctrness of a response.'[8]

What research really says about Feedback

Even where it is legitimate to provide an average 'effect size', all decent meta-analyses provide a clear and accessible explanation of the studies this is based upon. The Toolkit provides some more specific references but many are over 20 years old or currently unobtainable. Some of the sources are very critical of particular types of feedback.

The EEF itself has funded one experiment, but with *zero* effect!

Clicking the References link reveals a list of 15 sources without any clear indication what each is supposed to demonstrate; 6 are more than 25 years old. This leads to 7 more detailed references (mostly from the first list), each with an 'effect size'. These range from .97 to .20. Which is to be believed? Summaries of six follow, in highly technical language, mostly without indicating which stage or subject, what kind of learning, what kind of feedback, which countries the research took place in, and so on. None of the abstracts are comprehensible without reading the original research report, and 4 of these 7 are from the 1980s summarising even earlier work. If you do gain access to the original studies, you find a bundle of problems. Some refer only to simple kinds of learning (memorisation, the acquisition of factual knowledge), and

feedback that is about accuracy. One shows that a third of the interventions had a negative impact. Another, claiming a high effect size, is restricted to students with learning difficulties.

Assessment for Learning

Two of the best known advocates of feedback and formative assessment, Paul Black and Dylan Wiliam, refuse to provide an average effect size precisely because the original studies are so disparate. A section of their key paper *Assessment and classroom learning*[9] has the title *No Meta-Analysis*.

This paper points to the decisive role of students: the feedback has to be *used*! They argue that the benefits of feedback often depend on giving greater responsibility to students, who must not be seen as 'passive recipients'. Successful feedback often involves rethinking teaching and learning, so that students have to think more about how to solve problems and teachers are more alert to what students are thinking.

Black and Wiliam argue that feedback cannot be evaluated outside of its relation to particular kinds of learning, and it is pointless making claims about feedback in general.

What is surprising from reviewing the literature is how little attention has been paid to task characteristics in looking at the effectiveness of feedback. The quality of the feedback intervention, and in particular, how it relates to the task in hand, is crucial. (Black and Wiliam 1998, p39)

Conclusion

*None of the research suggests the need to overload teachers with mountains of marking.*Comments on written work is one way to provide feedback, but other forms include monitoring students' learning during a practical process, providing judicious hints on tackling a problem, steering them to think about alternative solutions and methods, and so on.

There is nothing straightforward about implementing an intervention called 'Feedback'. There is no generic entity called 'feedback' which on average adds 8 months progress.

Improving learning and enhancing achievement requires a thoughtful process of staff development, not the quick-fix solutions implied by a 'Toolkit'. This quality of staff development is unlikely to happen in a climate of exhaustion and fear.

[1] Stephen Ball (2001) *Performativities and fabrications in the education economy – towards the performative society*.

[2] DfE (2014) *Teachers' workload diary survey 2013: research report*.

[3] Ofsted inspections – clarification for schools, Oct 2014, no. 140169.https://www.teachers.org.uk/files/ofsted-inspections-clarification-for-schools.pdf

[4] http://icingonthecakeblog.weebly.com/blog/the-high-cost-of-effective-feedback-the-triple-marking-fiasco

[5] The TeacherToolkit blog (27oct2014) suggests that this marking frenzy occurred to fill the gap from lesson grading, i.e. because Ofsted inspectors are now seeking evidence in exercise books instead of observation.

[6] Many of the limitations are acknowledge in the Technical Appendices, and some of the differences between research studies are shown among the references, but in practice the league table format and average effects in terms of 'months progress' will lead to simplistic and erroneous responses.

[7] See for example M Russo (2007) *How to review a meta-analysis*. Robert Coe (*It's the effect size, stupid*, 2002) makes exactly the same point: "Given two (or more) numbers, one can always calculate an average. However, if they are effect sizes from experiments that differ significantly in terms of the outcome measures used, then the result may be totally meaningless." Unfortunately his Toolkit team seems to have missed his sound advice.

[8] Ironically, this quotation comes from meta-analysis enthusiast John Hattie (Visible Learning, p174, 2009).

[9] P Black and D Wiliam (1998) Assessment and classroom learning, p40 Unfortunately, the statement they made elsewhere about other people's studies "typically reporting" effect sizes between .4 and .7 is taken out of context and been read as an authoritative statement of fact.