PART I

Theoretical Foundations

Methods and Concepts

1 Feedback

At the Heart of – But Definitely Not All of – Formative Assessment

Dylan Wiliam

Introduction

In 1996, Avraham Kluger and Angelo DeNisi, two psychologists at Rutgers, the State University of New Jersey, published a rather remarkable review of the research conducted from 1905 to 1995 on the effects of feedback. In their conclusion, they suggested that most of the research that had been conducted into feedback had shed little light on how to make feedback more effective because more attention had been given to whether, rather than how, feedback worked, and most of the studies had focused on the short-term effects of feedback, taking little account of whether the effects would be sustained over time. In this chapter, I review the history of feedback research and suggest that feedback research will be more effective if this research is located within a wider theoretical framework, taking more account of both micro and macro features of learning. Specifically, feedback research is likely to be more effective if it places greater attention on the cognitive processes that are involved in learning (the micro level) and on the social situations within which feedback is given and received (the macro level).

The Origins of the Concept of Feedback

In 1948, Norbert Wiener, then professor of mathematics at the Massachusetts Institute of Technology, published a book titled *Cybernetics, or control and communication in the animal and the machine* (Wiener, 1948). In it, he pointed out that in many physical systems (whether mechanical or organic), effective action required not only a mechanism for producing the action required but, in addition, some means for monitoring whether the mechanisms had produced the required effect. In railway signaling, when a lever is moved to change signals or points (switches in US English), a light or other confirmatory device shows that the intended change has indeed taken place. In the Navy, on receipt of an order, subordinates repeat the order to show that it has been heard correctly.

Such systems are needed only when the effect of the action taken is uncertain. If the desired effect always occurs, there is no need to check whether the action has been effective. For example, at typical atmospheric temperatures, it takes around 1,200 joules of energy to raise the temperature of 1 cubic meter of air by

1 degree Celsius. So if we know the volume and the temperature of the air in a building, and we know the temperature we would like the air to be, then it appears to be a straightforward task to calculate the heating requirements of the building. Unfortunately, things are not quite that simple. The building loses heat, and the rate at which the building loses heat depends on the quality of the insulation, the exterior temperature, the speed of the wind, and a whole host of other factors. That is why thermostats are used to measure the temperature of the air in rooms, to compare it with the desired temperature, and when the measured temperature goes too far below the desired temperature, some action, such as increasing fuel to a boiler or opening up a damper, is taken to change the temperature of the air. In discussing this "chain of the transmission and return of information" Wiener suggested that it should be called "the chain of feedback" (Wiener, 1948, p. 114).

In the middle part of the nineteenth century, the term "feed back" had been used in the United States to describe the reversion of a system back to some previous state. For example, a patent application in 1865 described what an operator of a machine for making spindles for wagon axles should do when "the carriage is about to feed back" (Jay, 1865). However, the first use of the term in its current sense seems to have been by Karl Ferdinand Braun, winner (jointly with Guglielmo Marconi) of the 1909 Nobel Memorial Prize in physics. In his acceptance address he pointed out that it was possible to separate out two electrical oscillations with a special circuit "*so long as care is taken* that as far as possible the circuit has no feed-back into the system being investigated" (Braun, 1909, p. 233, emphasis in original).

Of course, the idea of self-correcting systems had been around for thousands of years. Some versions of the ancient Greek clepsydra (water clock) incorporated a conical float to regulate the flow of water into the reservoir, and this is probably the earliest known example of a self-regulating device that needed no control from an external agent. James Watt's steam engine employed a "governor" to control the speed of the engine. Two solid metal balls swung on pendulum rods on opposite sides of a rotating shaft, and when the speed of the shaft's rotation increased, the balls swung outward. The pendulum rods were in turn connected to the intake valves of the steam engine, so that when the shaft speeded up, the valves closed down, slowing the speed of the engine, and when the speed of the engine dropped too much, the valves opened up, speeding up the engine. In this way, the system regulated or "governed" the speed of the engine.

Wiener described Watt's governor as an example of *negative* feedback because, in a governor, "the feed-back tends to oppose what the system is doing" (Wiener, 1948, p. 115). The alternative, where the feedback tends to reinforce what the system is already doing, is termed positive feedback. Although positive feedback loops can be useful – for example in amplifying a faint sound or a weak electrical signal – they are unstable. Most people are aware of the problem encountered in sound amplification in which sound from a loudspeaker is picked up by a microphone, which is then further amplified,

which in turn increases the sound received by the microphone, increasing the sound coming out of the loudspeaker further still, leading to a howling sound that is often just called "feedback." Another example of a positive feedback loop is the growth of a population of animals with plentiful food and no predators; the population just keeps on growing exponentially.

A less obvious example of a positive feedback loop is an economic recession, in which people lose their jobs, or have their hours cut back, so they have less money, and so rein in spending, leading to further job losses, leading to further losses of confidence, and so on. This is a positive feedback loop because the information being fed back (there's a recession) has the effect of pushing the system further in the direction it was already going (less economic activity). In engineering, therefore, "positive" feedback is generally not good, since it leads either to explosive increase (as in the case of the amplification system) or collapse (as in the case of a recession). In contrast, negative feedback loops are useful, since they tend to produce stability.

For example, where the supply of food is limited, a population of animals will grow quickly at first, but as competition for food intensifies, the rate of increase will slow. The population will then tend toward a steady state, known as the "carrying capacity" of the environment (Levins, 1966, p. 427). In a similar way, the effect of the room thermostat is an example of negative feedback since the effect of the system is to heat the room when it is too cold, or to cool the room when it is too hot.

Feedback in Psychology and Education

As the term "feedback" was used more and more in engineering, it was picked up in psychology. Early uses of the term focused on motor control or the electrical circuits involved in nerve stimulation (see, for example, Washburne, 1935), but in research on the effects of reinforcement, rewards or other sensory information given to subjects in experiments were sometimes described as "feedback" (W. O. Jenkins & Stanley, 1950; Skinner, 1950). More interestingly, researchers on human performance in organizational psychology were also using the term "feedback" to describe information given to individuals or groups about their own performance (D. H. Jenkins, 1948; Roseborough, 1953; Wilson, High, & Beem, 1954). By 1954, the use of the term was sufficiently established for Robert Gagné – one of America's leading psychologists of education – to suggest that issues about the quality, grain size, and frequency of feedback, and how they impacted performance, were key areas for research on learning (Gagné, 1954).

An important focus for early research in this area examined whether corrective or reinforcement feedback was the most useful. In other words, was feedback more effective when it provided reassurance to learners that they were "on the right track" or was it better if feedback was given only when the actions of the learner deviated from what was desired in some important respect?

Intuitively, there are plausible arguments for each of these ideas. If someone is following a set of driving directions, it can be helpful to be told, "You will pass a BP petrol station on your right" but it is also helpful to know that, "If you pass under the motorway, you've gone too far."

It is tempting to regard the relationship between corrective and reinforcement feedback as being equivalent to the relationship between negative and positive feedback in engineering, but there are important differences. In engineering, whether feedback is positive or negative depends on its effect; if the effect is to push the system further in the direction in which it is already moving, it is positive, while if the feedback serves to oppose the existing trend, it is negative. In psychology, the terms "corrective" and "reinforcement" describe the intent of the feedback, and, as researchers commonly found, the actual effects of the feedback were different from what was intended.

As Kulhavy noted in his review of research on feedback in 1977, much of the early research into the effects of feedback was rooted in the behaviorist paradigm and strongly tied to the "programmed learning" movement (Skinner, 1968). The idea was that telling students that their answers were correct "reinforced" the cognitive processes through which the student had gone in order to arrive at the correct response, and thus would increase the likelihood that the correct response would be given to a similar prompt in the future. B. F. Skinner summed it up as follows:

> the machine, like the private tutor, reinforces the student for every correct response, using this immediate feedback not only to shape his behavior most efficiently but to maintain it in strength in a manner which the layman would describe as "holding the student's interest." (Skinner, 1968, p. 39)

Kulhavy defined feedback as "any of the numerous procedures that are used to tell a learner if an instructional response is right or wrong" (Kulhavy, 1977, p. 211). At its simplest, this would involve simply indicating whether a learner's response to an instructional prompt was correct or not (often called "knowledge of response" or "knowledge of results" or simply "KR"). Other examples included feedback that, for incorrect responses, provided a correct response ("knowledge of correct response" or "KCR") or gave the learner multiple further attempts to produce a correct response, possibly with additional support, such as an explanation provided by the teacher ("repeat until correct"). More complex forms of feedback included feedback that provided the learner with information on what needed improvement ("correctional review"). Ultimately, as it became more and more complex, feedback became indistinguishable from instruction:

> If we are willing to treat feedback as a unitary variable, we can then speak of its form or composition as ranging along a continuum from the simplest "Yes-No" format to the presentation of substantial corrective or remedial information that may extend the response content, or even add new material to it. Hence, as one advances along the continuum, feedback complexity increases until the process itself takes on the form of new instruction, rather than informing the student solely about correctness. (Kulhavy, 1977, p. 212)

However, it is clear that the evidence about the superiority of reinforcement, rather than correction, was not that clear cut:

> [Referring to the statement made by Skinner quoted above] With such confident statements available, it is no surprise that scholars have worked overtime to fit the round peg of feedback into the square hole of reinforcement. Unfortunately, this stoic faith in feedback-as-reinforcement has all too often led researchers to overlook or disregard alternate explanations for their data. One does not have to look far for articles that devote themselves to explaining why their data failed to meet operant expectations rather than to trying to make sense out of what they found. (Kulhavy, 1977, p. 213)

Kulhavy pointed out that many studies (e.g., Anderson, Kulhavy, & Andre, 1971) had found that telling learners that they were on the wrong track (negative reinforcement) was more effective than telling them they were on the right track (positive reinforcement) – something that should not occur if the primary benefit of feedback was as reinforcement. More serious for the "feedback-as-reinforcement" hypothesis was the finding that delaying feedback for a day or more often had no impact on its effectiveness – what is sometimes called the "delay-retention effect" (Kulhavy, 1977). In fact, as a later review by Bangert-Drowns, Kulik, and Kulik (1987) found, delaying feedback until after the learner had been required to attempt a task – called "reducing presearch availability" – consistently enhanced the effects of feedback.

During the 1980s, a number of attempts were made to make sense of the often apparently conflicting research findings. Some used what Robert Slavin (1986) termed a "best evidence" approach (the approach used by Kulhavy), while others, such as Schimmel (1983) and Kulik and Kulik (1988), used meta-analysis. Most of the reviews found that, on average, providing feedback was better than providing no feedback, but beyond that, clear results about how to maximize the benefits of feedback were hard to find.

For example, Kulik and Kulik found that delayed feedback appeared to be more effective than immediate feedback in what they called "applied" studies, where students were tested on material related to, but different from, what they had been taught. In the twenty-eight studies they found where the material to be learned took the form of lists of words or terms, the results were less clear cut. In ten of the studies, immediate feedback was significantly better, while in four of the studies, delayed feedback was significantly better. In the remaining thirteen, the differences were nonsignificant (with six favoring immediate feedback and seven favoring delayed feedback).

A clue as to the mechanisms involved was provided by a review of studies of feedback in "test-like events typical of classroom education and of text-based and technology-based instruction" (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). Their review focused on feedback that was intentional (i.e., intended to improve learning) and mediated (i.e., given by answers in a text or through computer-based instruction, rather than by direct human interaction), and generated four main findings.

Feedback Type: "Correct/incorrect" feedback (what Kulhavy had termed "knowledge of results") produced no benefit; providing correct answers ("knowledge of correct results") provided some effect (Cohen's $d = 0.22$), while "repeating until correct" or providing an explanation provided larger effects ($d = 0.53$ in each case).

Presearch Availability: Where learners were able to find answers before they had attempted the task, feedback did not improve learning, but where this was not possible (where there was, in the jargon, "control for presearch availability"), feedback had a substantial positive impact on learning ($d = 0.46$).

Study Design: Feedback had no effect in studies where students were given a pre-test, but in the studies that relied only on post-tests, there was a significant impact on learning ($d = 0.40$).

Type of Instruction: Feedback in programmed instruction and in computer-assisted instruction had no impact on learning, but in text comprehension exercises, and test performance, feedback had a significant positive effect ($d = 0.48$ and 0.60, respectively).

Bangert-Drowns et al. concluded that the crucial determinant of the effects of feedback was the extent to which the feedback was received in a "mindful" way, and similar conclusions were reached in other reviews of research into feedback conducted at the same time (see, for example, Dempster, 1991, 1992), and a review of research studies published in Dutch also underscored these main findings (Elshout-Mohr, 1994).

In 1996, Kluger and DeNisi published what is perhaps still to this day the most important study on the effects of feedback on performance. They defined a feedback intervention as "actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance" (Kluger & DeNisi, 1996, p. 255). They searched a number of bibliographic databases, including the Social Science Citation Index, PsycInfo, and National Technical Information Services, for any studies – going back as far as 1905 – that included either "knowledge of results" or "feedback" as one key word and "performance" as another, a process that yielded approximately 2,500 papers and 500 technical reports.

To ensure that only high-quality feedback studies were included in their review, Kluger and DeNisi (1996) rejected:

- Studies where feedback was combined with some other intervention, such as target-setting. Even if the intervention was effective, it would be impossible to know whether the increased performance was due to the feedback or the target-setting component of the intervention.
- Studies that lacked a control group. Where performance was measured before and after a feedback intervention, it would be impossible to attribute any increase in performance to the feedback since it could have been achieved simply by maturation.

- Studies where performance was estimated qualitatively rather than measured quantitatively, since the lack of measurement obviously introduces a substantial element of subjectivity into the results.
- Studies where the experimenter participated as a subject (!) or those that involved fewer than ten participants, due to the large errors of sampling involved.
- Studies where the reports of the results did not allow an effect size to be estimated (e.g., studies where the standard deviation of the performance measure was not provided and could not be estimated from the information provided in the report).

Of the original 3,000 studies identified as potentially relevant, only 131 (fewer than 5%!) of the studies met all five of these criteria, and the reports of the selected studies allowed Kluger and DeNisi (1996) to calculate 607 effect sizes based on 12,652 participants (with a total of 23,663 observations). They found reasonably strong evidence that feedback that directed recipients' attention to themselves (e.g., general praise, comparison with others, other general threats to self-esteem) tended to be less effective than feedback that focused attention on the tasks in which they were engaged. They also found that where initial feedback showed a large gap between actual and desired performance, what was crucial was whether subsequent feedback showed a rapid improvement. Where learners saw such improvement, feedback had a substantial positive effect on achievement, but where they did not, the benefits of feedback were much smaller.

However, their most important conclusion was that little could be drawn from a review of the literature on feedback because the quality of the existing studies was relatively poor. A large number of studies were not included in their review due to the lack of a control group, presumably because researchers assumed that the effects of feedback would be positive. Also, in most of the studies included in the review, the impact of feedback was evaluated just once, with the impact being measured very shortly after the feedback intervention. Studies that included measurement of performance on several occasions, with feedback interventions between each, often yielded different results from the "one-off" studies. Moreover, in many of the studies, feedback was evaluated in terms of effects on shallow learning, which Kluger and DeNisi (1996) pointed out might interfere with deeper learning, and thus restrict the ability of individuals to transfer their learning to similar, but different, tasks. Perhaps even more importantly, they pointed out that even if a feedback intervention produced large increase in achievement, it might not be a good idea to implement it widely if the feedback made the recipient more dependent on the feedback for continued progress, especially if that feedback was expensive to provide.

Since Kluger and DeNisi's (1996) review, two further major reviews of the effects of feedback have been published. Shute (2008) sought to provide advice on the best ways to provide feedback to learners in intelligent tutoring

systems. Echoing the conclusions of Kluger and DeNisi, she found that there was no simple answer to the question, "What feedback works?" The effectiveness of feedback varied according to the instructional context, the kinds of tasks, and the characteristics of the students themselves. However, again echoing Kluger and DeNisi, she found that feedback was less effective when it focused on the learner, when it provided information only on the correctness of responses, or when it was so prescriptive that learners did not have to think for themselves. Feedback was more effective when it provided guidance – but not instructions – on how to improve performance – in other words, feedback should encourage mindfulness. She also found that for procedural learning and for tasks well beyond the learner's current capability, immediate feedback tended to be more effective, while delayed feedback was more effective for tasks within the learner's capability, or where the outcome measure involved transfer to other contexts.

Around the same time, Hattie and Timperley (2007) summarized the results of seventy-four meta-analyses of factors affecting student academic achievement that specifically included feedback. Building on the work of Ramaprasad (1983), they defined feedback as "specifically relating to the task or process of learning that fills a gap between what is understood and what is aimed to be understood" (p. 82). They confirmed the earlier findings of Kluger and DeNisi (1996) that feedback about the self as a person tended to be less effective, while feedback about the processing of the task and about self-regulation of learning tended to be more effective. Feedback about the task itself showed positive benefits when the feedback focused on strategy processing, or to enhance self-regulation of learning. Hattie and Timperley also noted that the conditions necessary for feedback to be optimally effective were not commonly found in practice.

Quantifying the Effects of Feedback

In the discussion of the review of research by Bangert-Drowns et al. (1991) above, a number of standardized effect sizes (Cohen, 1988) were quoted, with the most beneficial forms of feedback generating effect sizes of the order of 0.4–0.6. In Kluger and DeNisi's (1996) review, weighting the 231 effect sizes in proportion to the size of the sample generated a mean effect of 0.41 standard deviations, and Shute (2008) suggested that the effect of feedback on student achievement was in the range of 0.4–0.8 standard deviations. An unpublished review of the effects of feedback in college students by Nyquist (2003) found an average effect size, over 185 studies, of 0.4, but the effect varied substantially across different kinds of studies. In the thirty-one studies where feedback took the form of simply indicating the correctness of the results, the average effect size was 0.14, but in the forty-one studies where feedback provided information about guidance for improving performance, the effect size was 0.39, and where the feedback provided the learner with a specific task (ten studies), the average