

# FEEDBACK AND STUDENT LEARNING? – A CRITICAL REVIEW OF RESEARCH

*Stefan Ekecrantz*

## SAMMANFATTNING

Formativ bedömning i allmänhet och feedback i synnerhet har sin givna plats i rådande utbildningsvetenskapliga paradigmen. Genom stora översikter, meta-studier och synteser ses feedback på och för lärande som ett empiriskt synnerligen väl grundat fenomen. Den typen av närmast konsensusliknande försanthållanden riskerar att undslippa kritisk granskning över tid. I den här studien görs en uppföljande närläsning av ett särskilt inflytelserikt segment inom feedbackforskningen. Resultatet visar att den underliggande primärforskningen i det fallet inte alls bygger på forskning om elevers och studenter lärande, tvärtom hur denna forskning har refererats och använts vidare i efterföljande meta-analyser av till exempel Hattie (2009) och Hattie och Timperley (2007). Konsekvenser därav för forskning och evidensbaserad praktik diskuteras.

*Keywords: Feedback, formative assessment, learning, meta-analysis*

## INTRODUCTION

The accumulation of scientific knowledge requires courageous, bold conjectures that, to paraphrase Popper, can and need to be contested by the research community (Popper, 1959/2005, p. 278). In this era of educational research, when numerous authoritative meta-studies make highly generalizable (i.e. bold and courageous) claims, this needs to be done in many different ways to do this research justice. Through the influential work by e.g. Black and Wiliam (1998) and Hattie (2009), formative assessment in general and feedback in particular have been established as highly effective with regard to student learning. Their reviews, meta-studies and syntheses, along with those of others, have created what might be described as a consensus that feedback is "one of the most powerful influences on learning" (Hattie, 2009, p. 178). This notion continues to have a strong influence on research, policy and evidence-based recommendations for practice and large-scale implementations thereof (Hopfenbeck, Flórez Petour & Tolo, 2015; Jonsson, Lundahl & Holmgren, 2015; Ratnam-Lim & Tan, 2015).

## STEFAN EKECRANTZ

*Fil. Dr. i Historia*

*Verksam vid institutionen för pedagogik och didaktik  
Stockholms universitet, 106 91 Stockholm.*

*E-post: stefan.ekecrantz@su.se*

All such conceptions need to be constantly scrutinized and the impressive amount of knowledge that these meta-studies and the like represent is not to be misinterpreted as claims of finality. Or, as Hattie continues his quote above: “[Feedback] needs to be more fully researched by qualitatively and quantitatively investigating how feedback works in the classroom and learning process” (2009, p. 178). In addition, I argue, we do not only need to better understand how and under what circumstances teacher feedback on student performances promotes learning, but also continue to question the generalized claim itself: Does it? How do we know this? Could there be alternative explanations to these results? What is its empirical basis, and what are the main limitations therein? How has this research been used and what can it tell us about the need for future research?

To date, some critical voices have been heard from researchers that are sceptical of meta-analyses in educational research on principle. Such criticisms include questioning the assumption that meaningful knowledge of such complex and highly contextualized phenomena can be gained by quantitative analyses of alleged apples and oranges (e.g. Skourdoumbis & Gale, 2013). A different form of criticism comes from quantitative researchers who have questioned some claims on more methodological grounds (Bennett, 2011; Dunn & Mulvenon, 2009; Kingston & Nash, 2011). This study relates to both of these perspectives in varying ways.

The initial aim was to unveil contexts that had been decontextualized in a meta-analytic process. By focusing on different types of learning outcomes, varying academic disciplines, local assessment cultures and age groups in a selection of original research, the idea was to unveil possible white patches on a feedback and learning map that is often assumed to be more or less complete. However, as the work progressed it had to be renamed “a critical review” after the fact, zeroing in on mainly methodological issues concerning validity and relevance. In the selected studies it became clear that this research rarely focused on feedback leading to students learning something of academic relevance – contrary to how this particular research is presented in the formative assessment and feedback literature.

## FROM HATTIE & TIMPERLEY TO KLUGER & DENISI – A GENEALOGICAL CASE STUDY

The term *critical review* here refers to what Petticrew and Roberts (2008) describe as a: “term sometimes used to describe a literature review that assesses a theory or hypothesis by critically examining the methods and results of the primary studies [...], though not using the formalized approach of a systematic review.” (p. 41). The aim is not to cover the field as a whole, as is often the case in state-of-the-art reviews and the like, but rather to problematize some of its empirical foundations. It could perhaps best be described as a *genealogical case study* of sorts:

- A cornerstone of formative assessment and assessment for learning is various forms of teacher feedback on student learning (e.g. Wiliam, 2011).

- One of the most widely cited sources in support of the effectiveness of feedback for student learning is Hattie and Timperley's (2007) meta-study "The power of feedback", which is built on thirteen existing meta-studies.
- Hattie and Timperley's main reference, in turn, is Kluger and DeNisi's (1996) meta-analysis "The effects of feedback interventions on performance", building on 131 individual empirical studies and a total of 12,652 participants.

Kluger and DeNisi's work will be the main focus of this review. Their study is described by Hattie and Timperley as the "most systematic" (p. 84) of the thirteen meta-studies, and was also the most recent one. (Incidentally, Hattie and Timperley discuss twelve such studies, but thirteen are in fact listed). It is also by far the largest, building on 131 out of a total of 196 empirical studies in the thirteen meta-studies combined. In conclusion, these 131 studies make up a substantial portion of the empirical foundation of Hattie and Timperley's synthesis. Moreover, Kluger and DeNisi continue to be cited in a plethora of authoritative and emerging literature in the field of formative assessment, explicitly as empirical support of large effect sizes regarding feedback and student learning (i.e. Andrade & Cizek, 2010, pp. 20, 91; Hattie, 2009, p. 178; Jonsson, et al, 2015, p. 107; Van der Kleij, Feskens & Eggen, 2015, p. 2; Vlachou, 2015, p. 3; Voerman, Meijer, Korthagen & Simons, 2012, p. 1008). In other cases, Kluger and DeNisi's meta-study is used in a way that would lead most readers to conclude that they build their analysis on empirical data about student learning, even if this is not stated explicitly (e.g. Hattie & Timperley, 2007; Wiliam, 2011).

As in all research, a collection of accumulating support branch out into a breadth of previous research. Each step in these collective, accumulative arguments rely on multiple primary and secondary sources, which stands on the shoulders of other collective giants in the same manner. The other side of this is that each step in any such sequence leaves the original empirical research in an ever more distant past. Because of this, much of the original researchers' explicit reservations, insecurities and stated limitations risk being blurred and eventually forgotten. For this reason, the 131 studies in Kluger and DeNisi were analysed with regard to student age, subject, methodology and the type of outcome measured. The intention was to create a descriptive overview of the empirical basis for this influential strain of evidence: What were the main methodological caveats? Are some age groups more represented in this particular research segment than others? Does it rely more heavily on some subjects, such as writing, math, languages, social sciences or other? Were some contextual factors concerning feedback and student learning less researched?

Before proceeding to the examination of Kluger and DeNisi's research, the other twelve meta-studies in Hattie and Timperley's synthesis need to be described. L'Hommedieu, Menges and Brinko's (1990) meta-analysis deals with student evaluation of teachers, i.e. student feedback and how this affected subsequent evalua-

tions of teachers. This was not related to student learning or performance. Two were unpublished or inaccessible in full-text (Wahlberg, 1982; Moin, 1986). Skiba, Casey and Center's (1985-1986) study is a compilation of research on classroom management and behaviour in special education. Tenenbaum and Goldring (1989) reported on motor skills in combination with instruction only, and not in combination with learning in the cognitive domain. Three meta-studies dealt with extrinsic rewards, praise and punishment (Getsie, Langer & Glass, 1985; Rummel & Feinberg, 1988; Wilkinson, 1981). This is a feedback category identified by Hattie and Timperley as the least effective – and sometimes even detrimental – and is, therefore, of limited relevance for the type of feedback on learning most often associated with formative assessment.

Four other meta-studies deal specifically with feedback and student learning in the cognitive domain. All four focus on certain aspects of feedback and learning, rather than general perspectives. Bangert-Drowns, Kulik, Kulik and Morgan (1991) analyse the effects of testing frequency rather than feedback per se. Kulik and Kulik (1988) analyse research on timing aspects of feedback, and comparisons were not made between interventions with and without feedback, but rather between immediate and delayed feedback. Yeany and Miller's (1983) meta-study only covers feedback in science education, and its effects on attitudes and performance. Lastly, Lysakowski and Walberg's (1982) analysis is about instructional cues, student participation, reinforcement and corrective feedback. Their meta-analysis could have been a viable alternative to Kluger and DeNisi in this study, but is substantially smaller, builds mainly on studies from the late 1960s and early 1970s, and has had less impact in the field.

## STUDIES NOT DIRECTLY RELATED TO STUDENT LEARNING

In the analysis of Kluger and DeNisi's work, it soon became evident that a large number of the original studies covered areas that were either vaguely or only indirectly related to students and learning in formal education. Hattie (2009) makes note of this: "The most systematic study addressing the effects of various types of feedback was published by Kluger and DeNisi. [...] Although many of their studies were not classroom or achievement based, their message are of much interest" (p. 175). So do Hattie and Timperley: "[Kluger and DeNisi's] meta-analysis included studies of feedback interventions that were not confounded with other manipulations, included at least a control group, measured performance, and included at least 10 participants. Many of their studies were not classroom based." (2007, pp. 84-85). So what does classroom versus non-classroom-based research mean in this context? What was the nature of these studies?

Kluger and DeNisi covered research on feedback and performance in the most general sense possible. This led to the inclusion of research built on a range of so called performance outcomes – outcomes that often had little or nothing to do with student learning and had to be disregarded in this descriptive analysis. The main principle used for this exclusion was that only individual studies that could

be expected to be used in relation to student learning, without further evidence, were to be included in this overview. Since all of these studies are still part of the combined body of evidence in present day literature about formative assessment and feedback, there is reason to describe these excluded studies in some detail.

A number of the 131 original studies on feedback effectiveness dealt with workplace productivity and behaviour. This included ways to improve mental health centre staff productivity (Calpin, Edelstein, & Redmon, 1988), effect on productivity and satisfaction in organizations (Kim & Hamner, 1976), and the promotion of ear protection use in high noise workplaces (Zohar, Cohen & Azar, 1980). Another area covered by several studies was how feedback could increase worker vigilance in performing repetitive and monotonous tasks (e.g. Chung & Dean, 1976). Such workplace-related feedback research may or may not be relevant to some aspects of student learning, but would arguably not be used in isolation without further evidence of relevance. For this reason, 32 studies on feedback and productivity, safety and satisfaction in the workplace were excluded.

For the same reason, studies that dealt with the cognitive functions of the elderly (e.g. Rebok, & Balcerak, 1989), recognition memory of obese adults (Gardner, Sandoval & Reyes, 1986), judgements during driving (Lucas, Heimstra & Spiegel, 1973) and airplane pilot selection processes (Fowler, 1981) were excluded. As the overview was intended to describe research on learning in the cognitive domain, eight studies on feedback and motor skills were excluded. Such studies included research on how positive and negative information influenced elated and depressed subjects' motor skills (Anshel, 1987) and a study on stabilimeter precision (Wade & Newell, 1972). A single study about the performance of a hockey team, a study that reported on an increase in the team's number of legal body checks after a feedback intervention, was also excluded (Anderson, Crowell, Doman & Howard, 1988).

Another field not included was research on so-called *helplessness*, where subjects were asked to perform tasks that, unbeknown to them, were impossible to complete. The outcome often used was the time it took before the subject gave up, as a measure of personality and apathy rather than of learning. As many as ten out of the 131 studies were about such induced helplessness (e.g. Mikulincer, 1989). Other studies excluded were those that dealt with outcomes that were simply deemed too distant from student learning to be included. These included studies on mood manipulation in marketing research (Hill & Ward, 1989), IQ-test methodology (Kratochwill & Brody, 1976), post-stress performance (Foushee, Davis, Stephan & Bernstein, 1980) and strenuous exercise results on ergometers (Bandura & Cervone, 1983). Furthermore, one article that was a meta-study itself rather than primary empirical research was not included (Hulin, Henry & Noon, 1990).

A rather peculiar study on parapsychology and ESP (*sic!*) was also excluded (Vitulli, 1982). The psi-ability among the 26 test subjects in that study presumably improved when they were given feedback. When they received positive, correct feedback they got 18.43 out of 75 responses correct, while no feedback or incorrect feedback

rendered only 13.67 and 13.50 correct responses respectively. Unfortunately, the difference was not statistically significant, but the author argued that a  $p < 0.05$  threshold might not be optimal for this type of research. The reasons for excluding this particular study can be seen as self-evident. In all, 66 studies had to be excluded in the descriptive analysis for reasons explained above.

One criticism against Black and Williams' (1998) early reviews on formative assessment has been that they – and subsequently many that build on their work – did not distinguish between studies about students with special needs and other students (Dunn & Mulvenon, 2009). Black and Williams' main reference is a meta-study by Fuchs and Fuchs (1986) where 83 per cent of the original research was about students with cognitive disabilities. Since this group is known to be more positively affected by formative assessment and feedback compared to the general population (Cf Skiba, et al, 1985-1986), Black and Williams' claims regarding effect sizes were inflated according to Dunn and Mulvenon. To address this potential problem, ten studies about students with disabilities are not included in this overview, even though several of them reported on performance outcomes that could possibly be linked to student learning.

## FEEDBACK AND LEARNING? STUDENTS IN THE K-12 RANGE

Among the remaining 55 studies, as many as 36 used students in higher education as test subjects. This should not be interpreted as a particular interest in higher education learning, but rather as a consequence of where most of this predominantly psychometric research was being conducted. As a large meta-study from the University of British Columbia shows, first-year students in western Psychology and Education departments are highly over-represented in research about human cognition across age groups and cultural divides (Henrich, Heine & Norenzayan, 2010).

In a majority of the remaining 19 studies with students in the K-12 age range, the primary outcome was something other than learning. As a consequence, many articles lacked in transparency regarding the precise subject matter and what kind of learning that might have taken place due to feedback-based interventions. In some cases, an unspoken assumption of non-learning was in fact used as an independent variable, so that changes in post-feedback performance could be interpreted as evidence of something else, such as text-anxiety, motivation, self-efficacy, vigilance, concentration, or other. This is perhaps seen most clearly in studies that used aspects of IQ-tests and rather limited feedback interventions, where significant changes in the test subjects' spatial visualization ability or similar would not be expected.

In this analysis, a broad characterization was made based on whether the potential learning measured could reasonably be at least akin to some intended learning outcomes in education, such as creative thinking, communication skills or knowledge of science – which is labelled *academic*. The other category was *non-academic* subjects that would not likely be seen as intended learning outcomes in

themselves, such as memorization of pictures, non-verbal IQ-test and similar. One example of the latter is a study about two fifth grade creative arts classes, where the feedback intervention was aimed at classroom management and discipline (Winett, 1974). One class received group feedback on appropriate and inappropriate classroom behaviour and the other did not. The outcome measure was the degree of talking out of turn, ignoring teacher directions and the like, and not changes in creative arts achievements. This feedback improved discipline to a degree that was described as not dramatic but at least statistically significant.

The only study with K-12 students above ninth grade was an experiment with 45 high school juniors and seniors attending a university class (Glover, 1989). The students were selected from a group of particularly gifted science students with an IQ average of 131. The main objective was to investigate if inserted questions and feedback on correctness could improve the students' ability to estimate their own performance after having read a ten-page essay about the solar system. The students were divided into three groups. One group just read the text, one group read inserted multiple-choice factual questions – recall of isolated facts – and one group read the same questions but answered them and received feedback on the correctness of the replies. This was followed up by a similar test measuring post-test performance. The main results presented were that the students that received feedback estimated their own performance more correctly than the other two groups.

However, regarding actual performance on the test itself, the more significant difference was between the control group and the two groups that had read inserted questions with and without feedback. The control group got 11.12 correct answers on average, while the inserted questions groups scored 13.18 and 13.93 respectively. Thus, as for feedback and student learning in school ages above ninth grade, the only empirical result among all of Kluger and DeNisi's studies consists of a single experiment with highly gifted students attending a college course. In this, the average difference between 15 students receiving feedback and 15 students receiving no feedback was a meagre 0.75 out of 20 possible correct answers on an MCQ test about factual recall.

## FEEDBACK AND LEARNING? STUDENTS IN HIGHER EDUCATION

In a second experiment in the same study by Glover (1989), 60 freshman college students were tested in a similar fashion but with a control group, a group that received feedback on inserted factual knowledge questions and a group that received feedback on questions designed to be analytical as defined in the Bloom taxonomy. In this case, the group that received feedback on higher order thinking performed significantly better than the other groups on the same post-test factual knowledge MCQ: 11.45, 14.74 and 18.74, respectively. This makes it one of only a handful of studies that clearly reports a plausible relationship between a particular feedback intervention and significant improvement of student learning of an authentic academic topic.

Table 1. Subject and age groups. Students and feedback in studies included in Kluger & DeNisi (1996). Number of individual studies.

	1-3rd grade	4-6th grade	7-9th grade	High School	Higher ed.	Total
Academic						
Advertisement					1	1
Biology					1	1
Communication skills	2					2
Education					2	2
Math					3	3
Medicine					1	1
Psychology					5	5
Science			1	1		2
Science, arithmetic, social studies		1				1
Vocabulary					4	4
Non-academic						
Behaviour, discipline		1				1
IQ tests or similar			1			1
Memorization, pictures/letters	1	2			2	5
Multiple cues					2	2
Non-verbal IQ-test or similar	2	4			2	8
Numbers matching					3	3
Puzzle solving					1	1
Reaction time					7	7
SAT non-verbal			1			1
Shapes and forms	1	1				2
Visual monitoring					2	2
Total	6	9	3	1	36	55

To make sense of what is actually measured in these studies, it is often necessary to follow up on a substantial part of the literature that the authors relate their work to. At first glance, a multiple choice test before and after a feedback intervention might seem to be related to some kind of learning, but would, at closer inspection most often turn out to be a measure of test takers' motivation, test anxiety or other emotions. Other possible causes to achievement differences, such as learning, would then be seen as a methodological problem. Most authors dealt with this problem in the design and choice of outcome measure but did not address it explicitly. One exception can be found in Tinderley and colleagues (1991) who describe their efforts not to accidentally measure student learning:

Because subjects were randomly assigned to feedback conditions, false feedback ensured that the attributions evoked by the feedback would be independent of the subjects' past histories or abilities. Second, *in order to isolate the motivational consequences of feedback, the feedback could not contribute to performance improvement through learning*. Consequently, it was necessary that the feedback not be associated with any real performance differences among the subjects. (Tindale, Kulik & Scott, 1991, p. 47 [emphasis added]).

## FEEDBACK AND (PLAUSIBLE) LEARNING

Out of the 55 studies presented in table 1, only eleven were identified that seemingly reported on student learning that was of academic relevance. Again, this was not always the main objective of this research, but some kind of student learning could at least be deemed plausible. As evidenced in table 2, the limited number of studies represents a diverse body of research that covers only a minute fraction of the field. Perhaps most striking is that the only study other than Glover (1986) that focused on ages below higher education and above 6-year olds was a single study by Hanna (1976) with 1,391 5th and 6th grade students. These students completed two 18-item tests designed to mirror upper elementary, standardized testing of data interpretation in science, arithmetic and social studies. The students were divided into three groups where one got no feedback, one got partial feedback on correctness and the other group got total feedback, meaning that they got to continue trying if an item was answered incorrectly.

The students that received so called total feedback on the first test scored, on average, 10.26 on the second test, while the no feedback control group scored, on average, 9.63. With the presented data translated into effect size, the difference represents a Cohen's *d* of 0.23, which would usually be seen as a rather modest level of improvement. A methodological problem that is not addressed by the author is that the total feedback group got to spend 22 minutes on the test, while the control group only got to spend 15 minutes. The rationale for this was simply that the multiple attempts needed in the total feedback format took longer to complete. It seems plausible to assume that the 0.63 higher score average – out of 18.00 – in the feedback group, can be partly attributed to having been allowed to be actively engaged with the material for a 47 per cent longer time period. This type of shortcoming is quite common in a majority of the examined studies. This is most likely due to the genre and style they were written in, where a variety of alternative explanations and devil's advocate type discussions would not be expected.

More significant results were reported by Clark and colleagues (Clark, Guskey & Benninga, 1983), in a study about so-called 'mastery learning'. Out of 197 undergraduate education majors, 55 were selected to partake in a mastery learning group. Two instructors volunteered to join the mastery learning part of the project. Among other tasks, the 55 students completed formative tests with accompanying feedback and corrective activities throughout the length of a semester. In a final, authentic test, these students performed significantly better, with an average score of 26.39, compared to 23.69 in the control group. From the data provided, this

translates to a rather impressive effect size of 0.73. Final grades improved at a similar rate. Another important observation was that prior knowledge correlated significantly with the test results in the control group ( $r=0.356$ ) while this relationship was near zero in the experimental group ( $r=0.099$ ), suggesting that the intervention had been successful on an individual level.

This research illustrates a methodological trade-off that is quite common in educational research. On one hand, authentic and highly relevant learning outcomes could be seen following a real-life intervention implemented during a full semester, as opposed to the limited seven minute intervention in Hanna's study. On the other hand, the conditions for the control group and experimental group were not identical. Among other things, the selection of teachers was not randomized and teacher effects cannot be ruled out. Furthermore, the intervention was highly complex, and the results can therefore not be attributed to the administered feedback alone.

Table 2. Students, feedback and learning outcomes in studies included in Kluger & DeNisi (1996).

Author	Age	Subject	Outcome	Format	Caveats	Strengths
Reid et al (1988)	1-3rd grade	Communication skills	Explain, understand, critique	Verbal	6 year old children, at the far end of the age spectrum.	Long term effects. Authentic, relevant learning outcomes.
Sonnenschein (1986)	1-3rd grade	Communication skills	Explain, understand, critique	Verbal	6 year old children, at the far end of the age spectrum.	Novel task. Relevant learning outcomes. Substantial effects.
Hanna (1976)	4-6th grade	Science, arithmetic, social studies	Interpret	MCQ/ completion	Moderate results.	Very large sample. Unanticipated gender differences. Relevant learning outcomes.
Glover (1989) Experiment 1	High School	Science	Recall facts	MCQ	Highly gifted students. Main focus self-assessment ability, not learning. Moderate results.	Academically relevant outcomes.
Clark et al (1983)	Higher Ed	Education	"Range of course objectives"	"Final examination"	Complex intervention (mastery learning). Teacher effects cannot be ruled out.	Highly positive results. Authentic, classroom-based intervention over a full semester. Independent measurements of outcome.
Fulmer & Rollings (1976)	Higher Ed	Psychology	N/A	MCQ	Unlikely that the results reflect authentic learning.	Highly positive results. Authentic content.
Glover (1989) Experiment 2	Higher Ed	Science	Recall facts	MCQ	Main focus self-assessment ability, not learning.	Highly positive results. Academically relevant outcomes.
Lytle & Kulhavy (1987)	Higher Ed	Biology	Recall facts	MCQ	Small scale.	Highly positive results. Explicitly about learning.
Morgan & Morgan (1935)	Higher Ed	Psychology	N/A	True/false	Only knowledge of results. No explanatory feedback.	Highly positive results. Explicitly about learning.
Newman et al (1974)	Higher Ed	Psychology	Recall facts	MCQ	No significant difference between feedback and no feedback.	Explicitly about learning and retention. Authentic material.
Rees (1986)	Higher Ed	Medicine	N/A	True/false	Positive results only on repeated questions.	Explicitly about learning and retention. Authentic material. Long term retention.
Schloss et al (1988)	Higher Ed	Education	Explain, understand, recall	MCQ	Control and feedback groups were not controlled for ability or knowledge pre-test.	Highly positive results. Mostly higher order knowledge. Authentic academic topic.

Although the eleven studies presented in table 2 are quite few and dated, it is important to note that they met Kluger and DeNisi's quality criteria regarding size, control groups and experimental or quasi-experimental research designs. As a lot of classroom-based research cannot satisfy those demands, and may thus become ineligible for future meta-analyses, they could possibly serve as exemplars of how this can be achieved in feedback research.

## DISCUSSION – META ANALYSES, TIME AND REITERATION

One aspect of meta-studies rarely discussed is the relatively old age of the underlying primary research, something that is even more accentuated in multiple layers of syntheses. As mentioned, Kluger and DeNisi (1996) was the most recent meta-analysis in Hattie and Timperley's synthesis. Still, by the time Hattie and Timperley (2007) was published, the average age of the 131 studies was 27.7 years (SD = 11.3). This begs the question of its relevance for our understanding of feedback for learning today. If empirical results from earlier eras are accepted as sound, the main difference might be different foci – what might or might not be deemed important, or differing causal explanations – what best might explain these outcomes. Results from a particularly successful feedback intervention may be attributed to stimuli-response and positive reinforcement of behaviour by the original researchers, while someone else, perhaps decades later, might use the same results in support of a completely different theoretical perspective.

There are, however, instances when such differences could be seen as more incommensurable. One such instance might be differing views on context-dependency and human learning. For someone who is convinced that learning can only be understood as socially, historically and culturally situated, research with experimental designs that explicitly aims to nullify such factors may be of limited relevance. From a viewpoint that academic learning needs to be understood in its specific disciplinary context, generic research that lumps together the learning of algebra, history and languages across all age groups might be of little use. Lumping together workplace satisfaction and student learning should present even more of a problem. Thus, it seems evident that the quantitative analyses of Kluger, DeNisi, Hattie and Timperley are sometimes used by researchers who would refrain from using much of the underlying primary research if they had detailed knowledge of it.

The most common objectives were to establish effects on *behaviour modification* and test subjects' *motivation to perform* – in the workplace, in a laboratory or in some cases a classroom. Kluger and DeNisi themselves summarize their conclusions in a feedback intervention theory, where the main mechanism is feedback's potential to *affect locus of attention*. When asked to count slides of intact and faulty cups, (Bustamante, Moreno, Rehbein & Vizueta, 1980) or monitor horizontal dots at a distance of 10.8 versus 13.3 cm apart on a screen (Wiener, 1975), the test subjects were seemingly motivated to stay concentrated when they were kept informed about results. Feedback – in the widest possible use of the term – also motivated people to improve performance in studies such as “Knowledge of

performance as an incentive in repetitive industrial work" (Hundal, 1969), "Effects of knowledge of results and differential monetary reward on six uninterrupted hours of monitoring" (Montague & Webber, 1965) and "Improving oral hygiene with videotape modeling" (Murray & Epstein, 1981).

Through what seems akin to a game of Chinese Whispers, all these test subjects have all been transformed into *students* over time. And, perhaps more troubling, a wide variety of emotive, motivational, cognitive and other outcomes were translated into *performance*, which then became *achievement*, which then became *learning* in the formative assessment literature:

Formative feedback has been widely studied due to its *enormous potential to support learning* and a large number of meta-analyses and reviews have been published on this topic, especially in relation to *classroom settings* (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Narciss, 2008; Shute, 2008)." (Coll, Rochera & de Gispert, 2014, p. 53 [emphasis added]).

Research has demonstrated *powerful effects of feedback* for student achievement in *individual learning settings* (see Hattie & Timperley, 2007; Kluger & DeNisi, 1996, for meta-analyses and overviews). (Asterhan, Schwarz & Cohen-Eliyahu, 2014, p. 34) [emphasis added].

So, does that mean it is time to place the work of, Kluger, DeNisi, Hattie and Timperley on the history shelves of feedback and learning research? Quite the contrary, I argue. Not only did Kluger and DeNisi contribute significantly to the field at the time, they also made several points that are still highly relevant. One such point is that not all feedback is beneficial, and that all types of feedback do not work in the same way. Their conclusion that extrinsic rewards and punishment might produce negative outcomes and that goal oriented, elaborated feedback can be expected to lead to positive outcomes, have been widely cited. Much would be gained if research and recommendations for practice maintained such a nuanced perspective regarding elaborated feedback for learning in the same manner. They also stressed that nearly a century of diverse empirical feedback research had resulted in a need for more developed theoretical frameworks and models. Much has been done in that respect (i.e. Hattie & Timperley, 2007; Shute, 2008; Nicol & Macfarlane Dick, 2006; Black & Wiliam, 2009), but there is still much to be done, perhaps especially with regards to less generic, discipline-, age- and context-specific models and theories.

Hattie and Timperley (2007) have been less scrutinized in this study, but considering their reliance on Kluger and DeNisi, the partly critical review presented here does, to some degree, apply to their meta-analysis as well. A reading of the other twelve meta-studies in Hattie and Timperley does not refute this description. That being said, their contribution to the field can hardly be overstated, and this contribution supersedes the mere notion that feedback is supposedly an effective educational tool. They also stress that different types of feedback work differently, and that there is more to the story than just 'more is better'. In addition, some of their reservations would arguably deserve more attention in the general formative assessment literature:

[Feedback about the task] is more powerful when it is about faulty interpretations, not lack of information. If students lack necessary knowledge, *further instruction is more powerful than feedback information*. (Hattie & Timperley, 2007, p. 91 [emphasis added]).

The impact of feedback was also influenced by the difficulty of goals and tasks. It appears to have the most impact when goals are specific and challenging *but task complexity is low*. (Hattie & Timperley, 2007, p. 85-86 [emphasis added]).

These assertions might lend some relief to the teacher who struggles with a perceived notion that the most effective remedies for lack of knowledge and understanding are formative assessment and individualized feedback in any and all cases.

## CONCLUSIONS AND CALL FOR FURTHER RESEARCH

This study has shown that, contrary to popular belief, the meta-analyses of Kluger and DeNisi (1996) and Hattie and Timperley (2007) are not based exclusively on empirical research about students learning from feedback. In the former, it is an exception rather than a rule. Their combined reliance on research that is only vaguely or indirectly related to learning puts their quantitative analyses, and how these effect sizes have often been used in the literature, in a different light. There exists the notion that the effectiveness of feedback for student learning is more or less conclusively researched on a meta-level with these analyses as references. Rather than upholding this, I would argue that it is time for new, large meta-studies on feedback and learning – studies that could possibly build on the theoretical framework and categories presented by Hattie and Timperley.

There are several reasons why this is called for. For one, there has been a surge in primary research on formative assessment, feedback and student learning in the last decade, allowing for new syntheses that do not have to rely on research areas of lesser relevance. Parallel to these developments, assessment for learning and formative assessment have been increasingly emphasized in policy and educational discourse in many parts of the world. Teacher feedback on student learning remains an integral part of these multifaceted concepts, which means that more research is needed. Such research would need to consider two methodological problems that have not always been adequately addressed in much of the present formative assessment literature: assessment validity and the concept of effectiveness.

In summative assessment literature, both the validity and reliability of tests and other instruments are very much at the forefront. To what degree do we really measure intended learning outcomes, and how reliable are these measurements? Concepts like the *hidden curriculum*, *cue seeking*, *consequential validity* and *teaching to the test* problematize and highlight these difficulties in general assessment literature and primary research alike. This is particularly evident in scientific and ideological debates on high stakes testing and quality assurance (Klein, Hamilton, McCaffrey & Stecher, 2000; Haney, 2000, 2001; Jones, 2007; Nichols, 2007; Ullucci & Spencer, 2009). In the feedback and formative assessment literature, these critical perspectives are often absent, equating post-intervention summative

assessment performances with intended learning outcomes (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; Briggs, Ruiz Primo, Furtak, Shepard & Yin 2012). A possibly controversial question would be if formative assessment regimes that yield positive results, to some degree may function through teaching-to-the-test processes, i.e. independent of, or even detrimental to, student learning?

As for educational effectiveness, the introduction of meta-analyses and effect sizes in educational research was a necessary development at a time when measures of correlation and statistical significance were the dominating quantitative metrics. Larger syntheses most often had to rely on narrative reviews or a crude vote count methodology. Not only were analyses of different outcomes made possible, but it was also possible to make substantiated claims about whether a change in outcomes was large or small – i.e. if it was *significant* in the non-statistical meaning of the word. Albeit arbitrary by necessity, standards for what constitutes large and small effects proposed by e.g. Cohen (1992) or Hattie (2009) contributed with much needed reference points. It could also be argued that Hattie's stipulation that an effect size of 0.40 constitutes a particularly important threshold in education is not any more arbitrary than the notion that the threshold between chance and statistical significance is to be placed at the  $p < 0.05$  level.

However, such effect sizes – usually mean differences divided by pooled standard deviations – leave researchers, policy makers and individual practitioners with only part of the story. The most urgent information would reasonably be evidence-based input on whether one alternative can be expected to yield better results than another – with a given set of resources. On a micro level, a teacher has a finite number of hours at his or her disposal. Such a real life educator might have to make a choice between spending the last couple of work hours preparing next day's class, designing a summative assessment task or writing individualized feedback on students' drafts. Evidence regarding what alternative can be expected to lead to larger effects may be bordering on useless if time and resources are not part of the equation.

There could be at least two ways to address this problem in future research. One would be to create research designs where the time and resources spent in experimental and control groups are kept at a constant. This is something that is not the case in most feedback versus no feedback comparisons. Another way would be to use methods to quantify effectiveness rather than merely effects sizes in absolute numbers. To be relevant for policy makers and educational leaders, this would most likely need to involve a multitude of fiscal aspects, whereas student learning per time spent might be the most relevant metric from a teacher perspective. Individualized, qualitative feedback may or may not fare as well in that perspective compared to present day research on its absolute effects.

## REFERENCES

- Anderson, D., Crowell, C., Doman, M. & Howard, G. (1988) Performance posting, goal setting, and activity-contingent praise as applied to a university hockey team. *Journal of Applied Psychology*, 73(1), 87-95.
- Andrade, H. & Cizek, G. (Eds.) (2010) *Handbook of formative assessment*. New York: Routledge.
- Anshel, M. (1987) The effect of mood and pleasant versus unpleasant information feedback on performing a motor skill. *Journal of General Psychology*, 115(2), 117-129.
- Asterhan, C., Schwarz, B. & Cohen-Eliyahu, N. (2014) Outcome feedback during collaborative learning: Contingencies between feedback and dyad composition. *Learning and Instruction*, 34, 1-10.
- Bandura, A. & Cervone, D. (1983) Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45(5), 1017-1028.
- Bangert-Drowns, R.L., Kulik, C.L., Kulik, J.A. & Morgan, M.T. (1991) The instructional effect of feedback in test-like events. *Review of Educational Research*, 61, 213-237.
- Barley, P. (1986) Trust, perceived importance of praise and criticism, and work performance: An examination of feedback in the United States and England. *Journal of Management*, 12(4), 457-473.
- Bennett, E. (2011) Formative assessment: a critical review. *Assessment in Education: Principles, Policy and Practice*, 18(1), 5-25.
- Betz, N.E. & Weiss, D.J. (1976) Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Rep. No. 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Black, P. (2015) Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, 22(1), 161-177.
- Black, P. & Wiliam, D. (1998) Assessment and classroom learning. *Assessment in Education*, 5(1), 7-74.
- Black, P. & Wiliam, D. (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-31.
- Briggs, D., Ruiz Primo, M., Furtak, E., Shepard, L. & Yin, Y. (2012) Meta Analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice*, 31(4), 13-17.
- Bustamante, J., Moreno, P., Rehbein, L. & Vizueta, A. (1980) Effects of feedback and reinforcement in tachistoscopic training on a fault-detection task. *Perceptual and Motor Skills*, 51, 987-993.
- Butler, R. (1987) Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology*, 79, 474-482.
- Butler, R. & Nisan, M. (1986) Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology*, 78, 210-216.
- Calpin, J., Edelstein, B. & Redmon, W. (1988) Performance feedback and goal setting to improve mental health center staff productivity. *Journal of Organizational Behavior Management*, 9(2), 35-58.
- Church, R. & Camp, D. (1965) Change in reaction-time as a function of knowledge of results. *American Journal of Psychology*, 78, 102-106.
- Clark, C.R., Guskey, T.R. & Benninga, J.S. (1983) The effectiveness of mastery learning strategies in undergraduate education courses. *Journal of Educational Research*, 76(4), 210-214.

- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1(3) 98-101.
- Coll, C., Rochera, M.J. & de Gispert, I. (2014) Supporting online collaborative learning in small groups: Teacher feedback on learning content, academic task and social participation. *Computers & Education*, 75, 53-64.
- Davies, A., Busick, K., Herbst, S. & Sherman, A. (2014) System leaders using assessment for learning as both the change and the change process: developing theory from practice. *Curriculum Journal*, 25(4), 567-592.
- DeLuca, C., Klinger, D., Pyper, J. & Woods, J. (2015) Instructional Rounds as a professional learning model for systemic implementation of Assessment for Learning. *Assessment in Education: Principles, Policy & Practice*, 22(1), 122-139.
- Dunn, K. & Mulvenon, S. (2009) A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, 14(7), 1-11.
- Flórez Petour, M. (2015) Systems, ideologies and history: A three-dimensional absence in the study of assessment reform processes. *Assessment in Education: Principles, Policy & Practice*, 22(1), 3-26.
- Foushee, H., Davis, M., Stephan, W. & Bernstein, W. (1980) The effects of cognitive and behavioral control on post-stress performance. *Journal of Human Stress*, 6(2), 41-48.
- Fowler, B. (1981) The aircraft landing test: An information processing approach to pilot selection. *Human Factors*, 23(2), 129-137.
- Fuch, L. & Fuch, D. (1986) Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53(3), 199-208.
- Fulmer, R.S. & Rollings, H.E. (1976) Item-by-item feedback and multiple choice test performance. *Journal of Experimental Education*, 44(4), 30-32.
- Gardner, R., Sandoval, Y. & Reyes, B. (1986) Signal-detection analysis of recognition memory of obese subjects. *Perceptual and motor skills*, 63(1), 227-234.
- Getsie, R.L., Langer, P. & Glass, G.V. (1985) Meta-analysis of the effects of type and combination of feedback on children's discrimination learning. *Review of Educational Research*, 55(1), 9-22.
- Glover, J.A. (1989) Improving readers' estimates of learning from text: The role of inserted questions. *Reading Research and Instruction*, 28(3), 68-75.
- Hanna, G. (1976) Effects of total and partial feedback in multiple-choice testing upon learning. *Journal of Educational Research*, 69, 202-205.
- Hattie, J. (2009) *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Abingdon: Routledge.
- Hattie, J., & Timperley, H. (2007) The power of feedback. *Review of educational research*, 77(1), 81-112.
- Hayward, L. (2015) Assessment is learning: the preposition vanishes. *Assessment in Education: Principles, Policy & Practice*, 22(1), 27-43.
- Hayward, L., Higgins, S., Livingston, K., Wyse, D. & Spencer, E. (2014) Special issue on assessment for learning. *Curriculum Journal*, 25(4), 465-469.
- Henrich, J. Heine, S. & Norenzayan, A. (2010) The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.
- Hermansen, H. (2014) Recontextualising assessment resources for use in local settings: opening up the black box of teachers' knowledge work. *Curriculum Journal*, 25(4), 470-494.
- Hill, R. & Ward, J. (1989) Mood manipulation in marketing research: An examination of potential confounding effects. *Journal of Marketing Research*, 26(1), 97-104.

- Hopfenbeck, T., Flórez Petour, M.T. & Tolo, A. (2015) Balancing tensions in educational policy reforms: large-scale implementation of Assessment for Learning in Norway. *Assessment in Education: Principles, Policy & Practice*, 22(1), 44-60.
- Hulin, C., Henry, R., & Noon, S. (1990) Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin*, 107(3), 328-340.
- Hundal, P.S. (1969) Knowledge of performance as an incentive in repetitive industrial work. *Journal of Applied Psychology*, 53(3), 224-226.
- Jones B. (2007) The unintended outcomes of High-Stakes Testing. *Journal of Applied School psychology*, 23(2), 65-86.
- Jonsson, A., Lundahl, C. & Holmgren, A. (2015) Evaluating a large-scale implementation of Assessment for Learning in Sweden. *Assessment in Education: Principles, Policy & Practice*, 22(1), 104-121.
- Kingston, N. & Nash, B. (2011) Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28-37.
- Klein, S., Hamilton, L., McCaffrey, D. & Stecher, B. (2000) What do test scores in Texas tell us? *Education Policy Analysis Archives*, 8(49) 1-16.
- Kluger, A. & DeNisi, A. (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254-284.
- Kratochwill, T. & Brody, G. (1976) Effects of verbal and self-monitoring feedback on Wechsler Adult Intelligence Scale performance in normal adults. *Journal of Consulting and Clinical Psychology*, 44(5), 879-880.
- Kulik, J. & Kulik, C. (1988) Timing of feedback and verbal learning. *Review of Educational Research*, 58(1), 79-97.
- L'Hommedieu, R., Menges, R.J. & Brinko, K. T. (1990) Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology*, 82(2), 232-241.
- Leong, W.S., & Tan, K. (2014) What (more) can, and should, assessment do for learning? Observations from 'successful learning context' in Singapore. *Curriculum Journal*, 25(4), 593-619.
- Lucas, R., Heimstra, N. & Spiegel, D. (1973) Part-task simulation training of drivers' passing judgments. *Human Factors*, 15(3), 269- 274.
- Lyhle, K. & Kulhavy, R. (1987) Feedback processing and error correction. *Journal of Educational Psychology*, 79, 320-322.
- Lysakowski, R. & Walberg, H. (1982) Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal*, 19(4), 559-578.
- Mikulincer, M. (1989) Coping and learned helplessness: Effects of coping strategies on performance following unsolvable problems. *European Journal of Personality*, 3(3), 181-194.
- Moin, A. (1986) Relative effectiveness of various techniques of calculus instruction: A meta-analysis. Unpublished doctoral dissertation, Department of Mathematics, University of Syracuse, Syracuse, New York.
- Montague, W.E. & Webber, C.E. (1965) Effects of knowledge of results and differential monetary reward on six uninterrupted hours of monitoring. *Human Factors*, 173-180.
- Morgan, C. & Morgan, L. (1935) Effects of immediate awareness of success and failure upon objective examination scores. *Journal of Experimental Education*, 4(2), 63-66.
- Murray, J. & Epstein, L.H. (1981) Improving oral hygiene with videotape modeling. *Behavior Modification*, 5(3), 360-371.
- Murtagh, L. (2014) The motivational paradox of feedback: teacher and student perceptions. *Curriculum Journal*, 25(4), 516-541.

- Newman, M., Williams, R. & Killer, J. (1974) Delay of information feedback in an applied setting: Effects on initially learned and unlearned items. *Journal of Experimental Education*, 42(4), 55-59.
- Nichols, S. (2007) High-Stakes Testing: Does It Increase Achievement? *Journal of Applied School Psychology*, 23(2), 47-64.
- Nicol, D. & Macfarlane Dick, D. (2006) Formative assessment and self regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2), 199-218.
- Petticrew, M. & Roberts, H. (2008) *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell Publishing.
- Popper, K. (1959/2005) *The logic of scientific discovery*. London: Routledge.
- Ratnam-Lim, C. & Tan, K. (2015) Large-scale implementation of formative assessment practices in an examination-oriented culture. *Assessment in Education: Principles, Policy & Practice*, 22(1), 61-78.
- Ratnam-Lim, C.T.L. & Tan, K.H.K. (2015) Large-scale implementation of formative assessment practices in an examination-oriented culture. *Assessment in Education: Principles, Policy & Practice*, 22(1), 61-78.
- Rebok, G. & Balcerak, L. (1989) Memory self-efficacy and performance differences in young and old adults: The effect of mnemonic training. *Developmental Psychology*, 25(5), 714-721.
- Rees, P.J. (1986) Do medical students learn from multiple choice examinations? *Medical Education*, 20, 123-125.
- Reid, L., Lefebvre-Pinard, M. & Pinard, A. (1988) Generalization of training speaking skills: The role of overt activity, feedback, and child's initial level of competence. *Perceptual and Motor Skills*, 66, 963-978.
- Reid, L., Lefebvre-Pinard, M. & Pinard, A. (1988) Generalization of training speaking skills: The role of overt activity, feedback, and child's initial level of competence. *Perceptual and Motor Skills*, 66, 963-978.
- Rummel, A. & Feinberg, R. (1988) Cognitive evaluation theory: A meta-analytic review of the literature. *Social Behavior and Personality*, 16(2), 147-164.
- Rust, J.Q., Strang, H.R. & Bridgeman, B. (1977) How knowledge of results and goal setting function during academic tests. *Journal of Experimental Education*, 45, 52-55.
- Schloss, P.J., Wisniewski, L.A. & Cartwright, G.P. (1988) The differential effect of learner control and feedback in college students' performance on CAI modules. *Journal of Educational Computing Research*, 4(2), 141-150.
- Shute, V. (2008) Focus on formative feedback. *Review of educational research*, 78(1), 153-189.
- Skiba, R., Casey, A. & Center, B.A. (1985-1986) Nonaversive procedures in the treatment of classroom behavior problems. *Journal of Special Education*, 19, 459-481.
- Skourdoumbis, A. & Gale, T. (2013). Classroom teacher effectiveness research: a conceptual critique. *British Educational Research Journal*, 39(5), 892-906.
- Sonnenschein, S. (1986) Developing referential communication: Transfer across novel tasks. *Bulletin of the Psychonomics Society*, 24(2), 127-130.
- Strang, H.R. (1983) The effects of knowledge of results upon reaction time performance. *Journal of General Psychology*, 108, 11-17.
- Strang, H.R., Lawrence, E.C. & Fowler, P.C. (1978) Effects of assigned goal level and knowledge of results on arithmetic computation: A laboratory study. *Journal of Applied Psychology*, 63(4), 446-450.
- Tenenbaum, G. & Goldring, E. (1989) A meta-analysis of the effect of enhanced instruction: Cues, participation, reinforcement and feedback and correctives on motor skill learning. *Journal of Research and Development in Education*, 22, 53-64.

- Tindale, R.S., Kulik, C.T. & Scott, L.A. (1991) Individual and group feedback and performance: An attributional perspective. *Basic and Applied Social Psychology*, 12(1), 41-62.
- Trembath, R. & White, T. (1979) Mastery achievement of intellectual skills. *Journal of Experimental Education*, 47(3), 247-252.
- Ullucci, K. & Spencer, J. (2009) Unraveling the myths of accountability: A case study of the California high school exit exam. *The Urban Review*, 41(2), 161-173.
- Van der Kleij, F., Feskens, R. & Eggen, T. (2015) Effects of feedback in a computer-based learning environment on students' learning outcomes – A meta-analysis. *Review of Educational Research*, (ahead of print), 1-37.
- Vitulli, W. (1982) Effects of immediate feedback on computer-assisted testing of ESP performance. *Psychological Reports*, 51(2), 403-408.
- Vlachou, M. (2015) Does assessment for learning work to promote student learning? The England paradigm. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 88(3), 101-107.
- Voerman, L., Meijer, P., Korthagen, F. & Simons, R. (2012) Types and frequencies of feedback interventions in classroom interaction in secondary education. *Teaching and Teacher Education*, 28(8), 1107-1115.
- Wade, M. & Newell, K. (1972) Performance criteria for stabilimeter learning. *Journal of Motor Behavior*, 4(4), 231-239.
- Walberg, H.J. (1982) What makes schooling effective? *Contemporary Education Review*, 1, 1-34.
- Wiener, E.L. (1975) On simultaneous monitoring and tracking. *Journal of Applied Psychology*, 60(1), 100-105.
- William, D. (2011) What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3-14.
- Wilkinson, S.S. (1981) The relationship of teacher praise and student achievement: A meta-analysis of selected research. *Dissertation Abstracts International*, 41(9-A), 3998.
- Willis, J. & Adie, L. (2014) Teachers using annotations to engage students in assessment conversations: recontextualising knowledge. *Curriculum Journal*, 25(4), 495-515.
- Winett, R. & Vachon, E. (1974) Group feedback and group contingencies in modifying behavior of fifth graders. *Psychological Reports*, 34(3), 1283-1292.
- Yeany, R.H. & Miller, P.A. (1983) Effects of diagnostic/remedial instruction on science learning: A meta-analysis. *Journal of Research in Science Teaching*, 20, 19-26.
- Zohar, D., Cohen, A. & Azar, N. (1980) Promoting increased use of ear protectors in noise through information feedback. *Human Factors*, 22(1), 69-79.