

The Black Box of Tertiary Assessment: An Impending Revolution

John Hattie

Visible Learning Labs, University of Auckland

Abstract

There has been a formative assessment revolution that has swept our compulsory schooling, and is about to descend on higher education. It has the potential to make major changes as to how assessment is undertaken, with more emphasis on “feedback from assessment” or “assessment for learning” (alongside the traditional “assessment of learning”), using assessment to improve and change what and how we teach. Students familiar with the power of this assessment model in high school (especially since the introduction of NCEA) will expect and demand different forms of feedback from assessment from their lecturers. At the same time, more international comparisons of universities will place pressure on universities to enhance their teaching and quality of assessment. This chapter reviews the multiple outcomes of higher education, the importance of alignment of curricula and assessment, outlines these newer models of assessment, reviews ‘What works best’ relating to teaching and learning in tertiary settings, and outlines the effects of newer assessment models for selecting higher education students.

Keywords: *Assessment for learning, higher education outcomes, power of feedback*

Introduction

A revolution of assessment has swept through many of our primary and secondary schools in New Zealand (and in many other countries). This revolution relates to Assessment *for* Learning and it can be witnessed in innovations such as the National Certificate of Educational Achievement (NCEA) and its standards-based approach, the emphasis on reporting more than on scoring, constructive alignment of learning and outcomes, peer collaborative assessment, learning intentions and success criteria, and the realisation of the power of feedback. These issues have become commonplace discussions in our schools. The higher education community is yet to catch up, but the doors seem to be opening and one hopes not too slowly. My theme is how we need to re-open the black box of tertiary assessment and what a wonderful set of new initiatives we can place into our black box. The revolution of using assessment as an integral part of teaching and learning, and not primarily as a summation of teaching and learning, are about to enter the halls of academia.

There is nothing wrong with black boxes: a ball-point pen is one and it works well – I do not want to know what is inside nor do I need to know, but it works fine. We have, in large part, a black box of assessment in tertiary assessment – it has worked for us for many years. We implicitly trust our academics to know what they value in their subjects, to set examinations and assignments, to mark reliably and validly, and then to record these marks (using more and more sophisticated marking engines), and the students move on, upwards, and/or out. We look at pass marks, we ‘red flag’ courses in which students satisfaction is not high, and we run the occasional course in teaching or assessment (occasional on the part of the lecturer, not staff

development personnel) that touches a few staff (usually the more committed). The newer models for teaching academics how to be teachers and set more appropriate assessment are exciting, as is the trend (mainly out of Europe) about the role of peers in assessment in new and powerful ways. The most exciting, however, is the move to include formative assessments notions in the all-so-often summative black box of tertiary assessment.

The aim of this chapter is to address these new moods. First, the inclusion of multiple outcomes from tertiary assessment begs the introduction of newer forms and foci for assessment. Second, the necessity for tertiary educators to spend more attention to the constructive alignment of what is desired to teach, what is taught, and what is then assessed. Third, there is a review of “What works best?” in the teaching and learning equation which highlights the power of “assessment *for* feedback”. This leads to the fourth and major claim about the revolution in assessment for learning and how it needs to penetrate the corridors of higher education. Finally, the recent changes in New Zealand towards more formative assessment and standards-based assessment in upper secondary school is reviewed to show, among many other benefits, that students may not be satisfied if these methods are not introduced into tertiary assessment. They expect better, they know the power of these methods, and they are more self-regulated than those who came through the black box of normative assessment, where assessment was primarily used to sum up the student’s status in terms of their “mastery of what the instructor thought he/she was asking in the final exam”.

Multiple Outcomes

Both society and students are demanding more from higher education. Chickering’s (1969; Chickering & Reisser, 1993) model is still worth noting in terms of these outcomes:

- i. Achieving competence.
- ii. Managing emotions – from those that interfere with learning (anger, anxiety, hopelessness – see Au, Watkins, Hattie, & Alexander, 2009) to those that assist (optimism, hopefulness).
- iii. Mature interpersonal relations – respecting differences, working with peers (probably the most underestimated power in higher learning).
- iv. Moving from autonomy to independence – moving from needing assurance and approval of others to self-sufficiency, problem solving, and making decisions.
- v. Establishing identity – self-esteem and self-efficacy.
- vi. Developing purpose – from Who am I? and Where am I? to Where am I going?
- vii. Developing integrity.

Chickering’s argument is that universities are in the business of developing all seven outcomes and/or they will be assessed by many on the institutional missions, processes and successes on all seven outcomes. While we spend the most time assessing the first, our funding and public success as well as our continual justification as an attractive place for people to come for higher education is as dependent on the other six. Higher education is as much about identity, reputation enhancement, and growing as it is about becoming critics and problem solvers. The (valuable) by-products are knowing more about a topic, being passionate about content, and being learned about a subject.

My prediction is that there will become more emphasis on the latter six outcomes while of course not forgetting the importance of the first (achieving competence). There have been attempts to assess success on the latter six (e.g., the Australian Course Experience Questionnaire), and occasionally universities consider whether they enhance critical thinking and similar non-subject-dependent attributes. I recall, when I was in an Australian university, the debate about measuring critical thinking and one of the best measures was the “number of complaints” received by students – but it soon lost favour. Probably the most important development hovering on our horizon is the Organisation for Economic Co-operation and Development (OECD)’s “Assessing Higher Education Learning Outcomes (AHELO)”, to explore the possibilities for developing comparative quantitative measures of graduate learning outcomes. This is a project of quite considerable importance as it aims to compare universities across nations, systems and cultures. The AHELO feasibility study has four strands:

The assessment of generic skills: critical thinking, analytic reasoning, problem-solving, written communication skills, generation of knowledge, and the interaction between substantive and methodological expertise;

The assessment of discipline-specific skills: complements the generic skills and will initially focus on selected discipline areas that are most common among universities in the participating countries and less likely to be influenced by unique cultural features. “The aim will be to assess competencies that are fundamental and ‘above content’; that is, with the focus on the capacity of students to extrapolate from what they have learned and apply their competencies in novel contexts unfamiliar to them, an approach that is similar to PISA.” The first subjects to be included are engineering and economics;

The measurement of the ‘value-added’ or contribution of tertiary education institutions to students’ outcomes: This will be based on two measures. The first one relates to the absolute performance or raw scores of students, since “prospective students or employers would want to know the ‘bottom line’ of the performance of HEIs, departments or faculties”; and the second is a measure of incremental learning (or ‘value-added’) “with a view to assess the quality of the teaching services provided by HEIs. It focuses on the scores an institution would attain after accounting for the quality of prior schooling or the degree of selectivity of the programmes and HEIs”; and

The *Contextual strand:* This will capture contextual measures at institutional level as well as appropriate indirect proxies of learning outcomes, such as:

- Academic studies and teaching (contact between students, counselling, courses offered, opportunities for e-learning, study organisation and teaching evaluation);
- Equipment (IT-infrastructure, library, computer workstations, spending for books and journals, rooms);
- International orientation (support for stays abroad);
- Job market and career orientation (employment market-related programmes, practice support);
- Research (number of doctorates, publications and internationally visible publications, extent of third party funding);
- Study location and TEI (amount of sport, level of accommodation rent, size of TEI); and

- Overall opinions (study situation, reputation for academic studies and teaching, research reputation).

These are akin to the PISA, PIRLS, and TIMSS assessments internationally that are becoming more important and now ever-present in our policy discussions in the compulsory sector. It is the case that they have more profound influences in other countries but increasingly so in New Zealand. Given our penchant in the university sector to want to be internationally recognisable, then the potential power of AHELO is high – especially when added to many of the current world rankings which depend mostly on research outcomes only.

Let me give you an example of the implications of adding these kinds of measures into international rankings. The 2008 Academic Ranking of World Universities by the Shanghai Jiao Tong University, based exclusively on research indicators, has American universities dominating the top ten with few changes over the past years (Table 1):

Table 1: 2008 Shanghai Jiao Tong University academic ranking of world universities

2008	(2007)	University
1	(1)	Harvard University
2	(2)	Stanford University
3	(3)	University California – Berkeley
4	(4)	University Cambridge
5	(5)	Massachusetts Institute of Technology (MIT)
6	(6)	California Inst Tech
7	(7)	Columbia University
8	(8)	Princeton University
9	(9)	University of Chicago
10	(10)	University of Oxford

An example of a ranking based more on Chickering's outcomes is the Forbes Rankings (Table 2). It includes a combination of research and teaching (albeit within the US), using measures such as the Listing of Alumni in Who's Who, student evaluations from Ratemyproffessors.com, four-year graduation rates, faculty receiving nationally competitive awards, and average four-year accumulated student debt of those borrowing money. The result is a totally different list (but note Princeton, Harvard, and Columbia are on both lists).

Table 2: Forbes ranking of US universities

2008	University
1	Princeton University
2	California Institute of Technology
3	Harvard University
4	Swarthmore College
5	Williams College
6	United States Military Academy
7	Amherst College
8	Wellesley College
9	Yale University
10	Columbia University

The major differences are that when you account for inputs, the ranking of outputs can change markedly (i.e., more value added). My message is that such assessment in higher education is likely to have profound effects on WHAT we measure in higher education, place more attention on the quality of our assessments – especially of student outcomes, and move the debates away from mainly research to student effects. We also need to show students that assessment can help them in their learning and not just be the end point for instructors to judge them.

The international assessments of tertiary institutions may have the highest probability to move from esteeming research as the major factor to esteeming also the quality of teaching and assessment alongside the important research components of higher education. The Performance-Based Research Fund (PBRF) in New Zealand, for example, is not intended to measure teaching capability – although its effects on postgraduate teaching have been profound and should not be underestimated (see Adams, 2008). For example, the “Research Degree Completions” component of the PBRF has led to increased attention to the quality of supervision and the progress of students *through* their research degrees. That more academics need to be involved in supervision is becoming more a focus of this area, and it is wonderful to see the system move from one where the *students’* woes about completion (they were not full time, they were not on time, they were not as interested after the first years) have turned to *systems’* woes about ensuring that the student is well prepared, well supervised, and well finished in reasonable time.

This should not be interpreted as a plea for a performance-based *teaching* fund. There are many other ways to achieve these outcomes relating to teaching than taking current funding away from tertiary institutions and redistributing it on the basis of teaching qualities (which is what PBRF does based on research qualities). There have been many initiatives to enhance the accountability of teaching in higher education. The PREQ and Course Experience Inventory from Australia, and the national Student Survey from the UK, are but recent examples. The NSS, for example, was designed to measure six factors (Teaching, Assessment and Feedback, Academic Support, Organisation and Management, Learning Resources, Personal Development), plus an overall satisfaction item. Each factor was based on multiple items, a total of 22 items in all. Marsh and Cheung (2007) completed an extensive evaluation of this instrument. First, the differences between universities were trivial – only about 2.5% of the variance (variance component based on multilevel analyses controlling for student characteristics and discipline) could be explained by between university differences. Second, student and institution background characteristics did not explain a lot of the variance in the overall satisfaction ratings. The greatest variance was explained by discipline-within-university groups (i.e., groups of student from the same discipline **and** the same university) and most of these differences were unreliable (due to small sample size). It would be hard to defend the use of this instrument as providing much information.

The Australian Course Experience Questionnaire (CEQ), aimed at monitoring the quality of students’ university experiences, has been available since 1991. The CEQ assesses characteristics of good teaching and effective learning such as enthusiasm, feedback, and clarity of explanations; the establishment of clear goals and standards; the development of generic skills; the appropriateness of the workload and assessment; and an emphasis on student independence (Ainley & Long, 1994; Johnson, 1998; Ramsden, 1991). A similar postgraduate version is the Postgraduate Research Experience Questionnaire (PREQ). There are small differences between disciplines and between universities, and the major source of variance was between supervisors. There were also differences relating to infrastructure (e.g., technical/financial support, computers). Whereas Humanities received the lowest ratings on this scale, this tended to be the case across all universities. Hence these discipline differences appeared to be inherent to the discipline and were

not a function of the supervisors, administration, policies and support in Humanity programs in particular universities.

The message from many of these assessments of teaching (undergraduate, course, and postgraduate) is their notorious unreliability at measuring between-institution or even between-department differences, and thus have little merit other than at the individual lecturer/supervisor level! We need more robust assessments of teaching effects in our tertiary system; the forthcoming international comparisons are most likely to drive the system to attending to these issues as core business, and thence investing in enhancing our teaching and assessments in higher education. The attention, however, will lead to more esteem for the other six Chickering outcomes alongside the academic achievement outcome.

Constructive Alignment

A key precondition for discussing assessment of teaching effects in higher education is the key notion, coined by John Biggs (1999, 2003), of “constructive alignment”. This notion has two premises: students construct meaning from what they do to learn; and the teacher aligns the planned learning activities with the learning outcomes (see Figure 1). Thus, any course needs to be designed so that the learning activities and assessment tasks are aligned with the learning outcomes that are intended in the course. This means that the system is consistent.

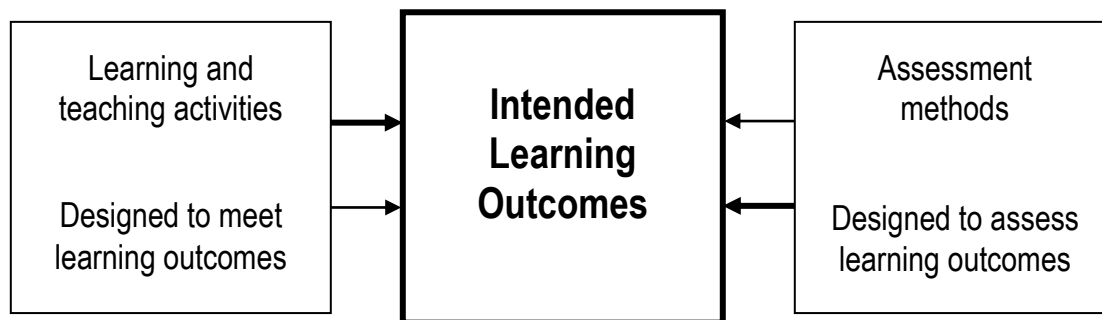


Figure 1: Constructive alignment

In the compulsory sector, this has been termed as “learning intentions” and “success criteria,” and one without the other is less effective. So often, learning in higher education involves students “working out” what is to be learnt and what it means to be successful in that learning – often they only learn this when they get the results of assessment back – and we need instead to make these clear before any assessments to thence maximise student learning. Without such alignment the powerful effects of feedback, reporting from assessment, and self-regulated learning are less likely to occur. I have seen so many studies of these latter notions implemented without attention to the constructive alignment and thus doomed to have little effect, especially little effect on the students. For most students, the information about learning intentions and success criteria are embedded in the assessment tasks themselves, and such wash back is not necessarily powerful.

For example, in preparation for this chapter I reviewed the examination papers in the Social Sciences, Arts, and Education in one university. The majority of assessments are still the “essay”. We know from decades of research on factor analysis of essays that the dominate

factors are: organisation, style, and language factors. Take the study by Page (1985), where 1000 essays were scored by six humans and his computer programme. This programme used variables intrinsic to the language of the written essay such as grammatical accuracy and vocabulary size and rarity, and various approximations of quality, the length of the essay, and the ratio of active to passive voice verbs. The average correlation of agreement between the computer and the six humans was higher than the correlations obtained between the six human judges. Regardless of what the humans thought they were using as the basis for rating each essay, there was a huge language component in their scores. The missing ingredient in human scoring of essays is content and understanding. If I had one wish in the improvement of assessment in higher education, it would be to ban the essay – but you can see my wish is unlikely to be granted.

The more exciting development is the newer automated essay scoring programmes, which indeed do highlight content and understanding far more successfully than humans (Shermis & Burstein, 2003). Also, the quality of the feedback from the essays is much superior; for example, we had teachers and these programmes score writing essays and when we asked teachers and students which they preferred in terms of informing them of what to do next, the overwhelming preference was for the computer feedback, which was also faster, more detailed, and much more reliably, consistently and validly scored. When we introduce these methods into higher education, we can restructure the academic's job from summative evaluator to more a formative role with a greater emphasis on the facility to teach; are they ready?

Similarly, Brown (2008) asked teachers about whether the learning desired in their courses was more surface or deep, and the majority said deep; he asked the students of these teachers, and they said surface. The students primarily used the nature of assessments to make their decision, and teachers need to contemplate more about the nature of wash back as it is much more powerful than their teaching claims – often because there is a lack of constructive alignment between the learning outcomes, success criteria, and the assessment methods and feedback.

To achieve constructive alignment requires a very critically reflective teacher who uses evidence to inform their decisions. It requires an alignment exercise of the learning intentions, success criteria, and assessments; it requires attention to the messages (overt and covert) given students about what is valued and to be learnt; it requires others to help in understanding their unanticipated messages about what is valued; it requires careful selection of teaching materials; and it requires frequent modification of module descriptions (often difficult in New Zealand universities).

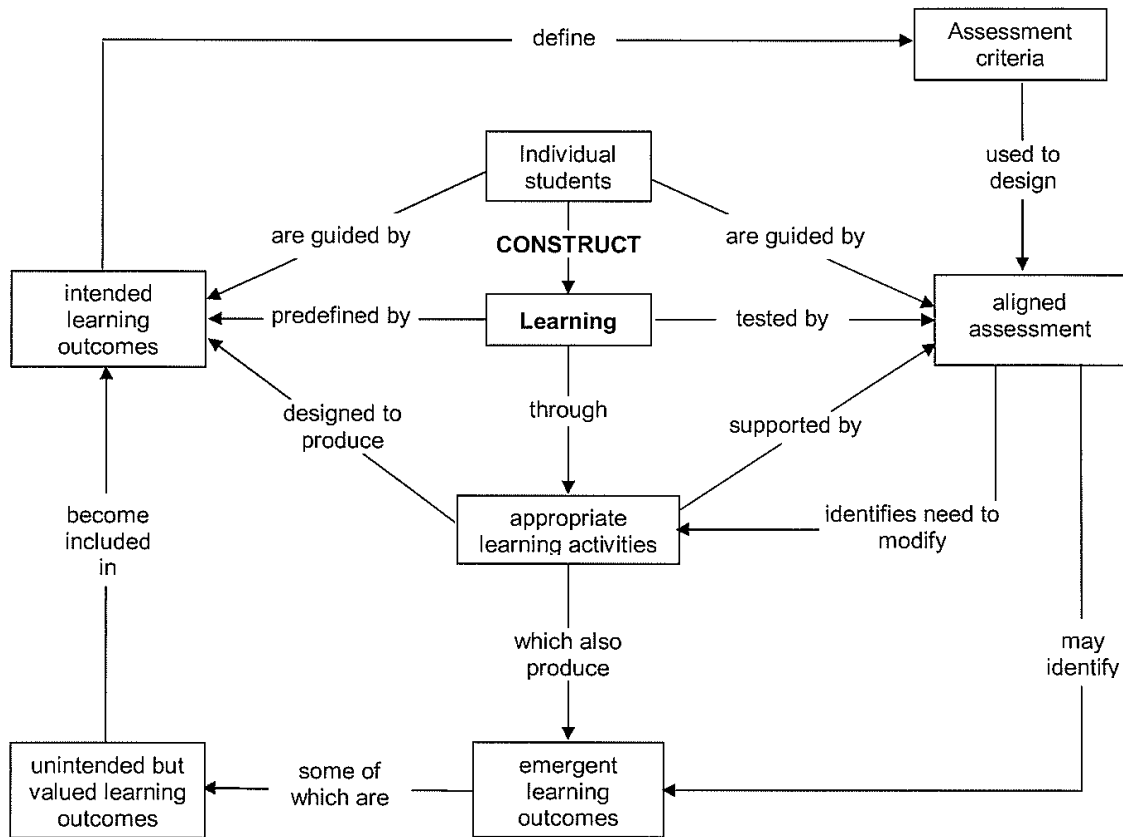


Figure 2: Concept map illustrating the main ideas put forward by Biggs (1996) and the relationships between them in the curriculum design process

Given our penchant towards essays and long scripted answers, then an influential way to increase alignment is NOT to write the essay or open-ended task first. Instead, one would first work out the scoring criteria that illustrate the desired attributes of each grade, and then construct a question or prompt that maximises the chance of the instructors seeing the responses to the scoring rubric. This method increases the reliability and validity of such scoring, makes the learning intentions most clear to the instructor (and to students if shared), and maximises constructive alignment. It is likely, moreover, that when computer scoring becomes more available, then students can trial their essays and can derive considerable learning about how to better write essays (and with obvious implications for summative assessment when such teaching help is available in this manner).

What works best?

One of my pleas to the compulsory sector is to move away from the question: “What works?” and instead ask the question “What works best?” – as I can show that 95%+ of innovations and policies “work”. I have synthesised over 800 meta-analyses, including about 250+ million students, 50,000+ studies, about 150,000 effect-sizes, from early childhood through adult education, in the search for what works best (Hattie, 2009). As can be imagined, these effects cover most subject disciplines, all ages, and a myriad of comparisons. Based on this work, several points can be made:

1. Almost everything works.

Ninety percent of all effect sizes in education are positive. Of the 10 percent that are negative, about half of these are “expected” (e.g., effects of disruptive students); thus about 95% of all things we do have a positive influence on achievement. When teachers claim that they are having a positive effect on achievement or when a policy improves achievement this is almost a trivial claim: virtually everything works. One only needs a pulse and we can improve achievement.

2. Setting the bar at zero is absurd.

If we set the bar at zero and then ask that universities “improve achievement”, we have set a very low bar indeed. No wonder every lecturer can claim that they are making a difference; no wonder we can find many answers as to how to enhance achievement; no wonder every student “improves”. It is easy to find programmes that make a difference. Raising achievement that is enhancing learning beyond an effect size of $d = 0.0$ is so low a bar as to be dangerous and is most certainly misleading.

3. Set the bar at $d = 0.40$.

The average effect size is $d = 0.40$. This average summarises the typical effect of all possible influences in education and should be used as the benchmark to judge effects in education. Effects lower than $d = 0.40$ can be regarded as in need of more consideration, although (as discussed earlier) it is not as simple as saying that all effects below $d = 0.40$ are not worth having (it depends on costs, interaction effects, and so on). Certainly effects above $d = 0.40$ are worth having and a major focus needs to be understanding the common denominators of what makes a difference (i.e., the effect sizes above compared to those below $d = 0.40$). I refer to this $d = 0.40$ effect size as the hinge-point or h-point, as this is the point on the continuum that provides the hinge or fulcrum around which all other effects are interpreted.

4. What works in schools, also works in universities.

While it is the case that most of the studies are compulsory schooling and not universities, there are sufficient studies in universities to make the strong conclusion: that the messages are the same about what works best, but the details may be different. I have sorted out those more pertinent to higher education and there are major messages for assessment at this level (see Table 3).

Table 3: Ranking of effects relevant to higher education

Rank	Domain	Influence	d
1	Student	Self-report grades	1.44
3	Teaching	Providing formative evaluation to lecturers	.90
8	Teacher	Teacher clarity	.75
9	Teaching	Reciprocal teaching	.74
10	Teaching	Feedback	.73
12	Teaching	Spaced vs. Mass Practice	.71
13	Teaching	Meta-cognitive strategies	.69
17	Curricula	Creativity Programs	.65
18	Teaching	Self-verbalisation/Self-questioning	.64
19	Teacher	Professional development	.62
20	Teaching	Problem solving teaching	.61
21	Teacher	Not Labelling students	.61
24	Teaching	Cooperative vs. individualistic learning	.59
25	Teaching	Study skills	.59
29	Teaching	Mastery learning	.58
30	Teaching	Worked examples	.57
34	Teaching	Goals - difficulty	.56
36	Teaching	Peer tutoring	.55
37	Teaching	Cooperative vs. competitive learning	.54
48	School	Small group learning	.49
49	Student	Concentration/Persistence/ Engagement	.48
56	Teacher	Quality of Teaching	.44
63	Teaching	Cooperative learning	.41
70	Teaching	Time on Task	.38
71	Teaching	Computer assisted instruction	.37
79	Teaching	Frequent/ Effects of testing	.34
97	Teaching	Special College Programs	.24
103	Teaching	Teaching test taking	.22
104	Teaching	Visual/Audio-visual methods	.22
106	Structural	Class size	.21
111	Teaching	Co-/ Team teaching	.19
112	Teaching	Web based learning	.18
120	Teaching	Mentoring	.15
122	Student	Gender	.12
126	Structural	Distance Education	.09
130	Structural	College halls of residence	.05

First, note the power of self-reported grades – this means that students are excellent predictors of their own performance. On the one hand, this may mean that we are excellent in giving feedback and cues to students throughout the course about how they are performing, and/or it may mean that students are aware of their prior achievement and how this can impact on subsequent performance. On the other hand, this is a negative as it may set upper bounds on the expectations students have for performance in a course. One of the powerful ways we can overcome and make these expectations false ceilings is to provide students with more confidence, self-efficacy, and set higher goals so that they can excel beyond their own

expectations. For many of us in higher education, we came here because at some point we realised we could contribute, understand, and make a difference in our areas of expertise – probably beyond that which we thought when we first entered higher education. The pursuit, teaching, measurement, and esteem that comes from fostering self-efficacy to excel is a taught notion, a developmental process, and is often the catalyst for enhanced academic achievement.

Second, note the power of feedback. Carless (this book) has written extensively on this subject, and I return to it below. The major point is that the most powerful feedback occurs when feedback is to the instructor: about how well they have taught, who they taught well, what they have or have not taught well. The trickle-down effect from such assessment that informs the instructor down to the student is much greater than the teaching and learning that comes from assessment directly to the student.

Third is the effect of the difficulty of goals. If we set difficult goals we are more likely to attain them than if we set easy “Do your best” goals. The art of making these appropriately difficult goals – and in the tertiary sector an important part of this “appropriately difficult goals” in our larger classes and more distance teaching methods is that they be transparent and achievable – is the learning intentions and the success criteria need to be explicit. If there is one thing we can do to enhance teaching and learning in universities, it is to pay more attention to BOTH these aspects. Over the years we are getting better at providing students with outcomes from the course; but we need more, much more – we need also to articulate the learning intentions of each class and particularly of each assessment. When students know these, and also know what success looks like, then learning is greatly improved. The use of scoring rubrics, worked examples, mastery learning, reciprocal teaching, and model answers provided to the student PRIOR to submission of work makes a major difference to the students’ learning; if these things do not occur, students must best guess what you want! Why not tell them the levels and degrees of what success looks like and see if they can attain this?

To accomplish this aim requires a different kind of professional development – seeing assessment *for* not *of* learning; learning more effective methods for developing criteria of success that students can understand and aspire to; writing different types of examinations and exercises in light of providing the success criteria; and teaching in manners whereby the lecturer seeks to illustrate what they mean by success!

Fourth, most students, bright or slow, need multiple opportunities to engage with the rich ideas underlying what success means on the learning intentions of any course. These multiple opportunities need to be spaced rather than in mass – some time between multiple opportunities, not all at once. Not all of us get the idea the first time; we need several chances to see your world and grasp the nuances, the hierarchies, the details, and the importance of thinking about and assimilating the new ideas. We need others to help us work out the self-questioning we need to develop, to work through what is and what is not important. Cooperative learning, team teaching and learning, group work, and teaching that involves multiple explanations are critical, as is the assessment to provide yet another opportunity to learn the desired outcomes and processes.

Fifth, the quality of teaching is high on the list, but it is quality teaching that provides clear learning intentions, transparent success criteria, listening in to how the students are thinking about and questioning the ideas, and learning how effective our own teaching is through the eyes of the students. I have developed a whole book on this theme of Visible Teaching – Visible Learning (Figure 3; Hattie, 2009).

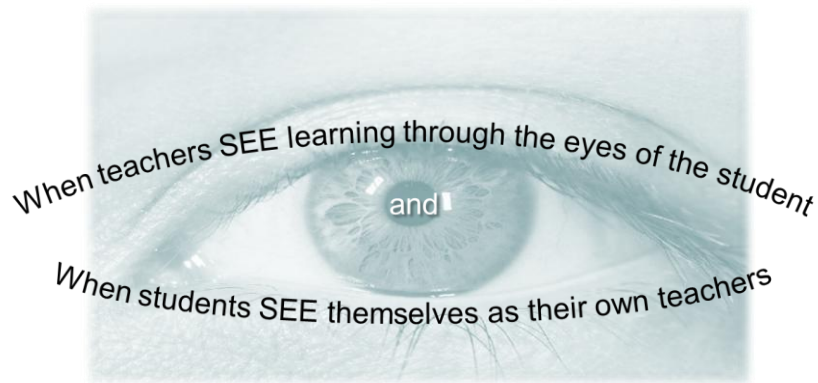


Figure 3: Visible learning

One of the major themes about assessment is that it should be developed, used, and evaluated primarily in terms of what the information from the assessment tells the lecturer about the quality of teaching, and not in terms of how the students advance or not. Lecturers need to see assessment as a way of seeing learning through the eyes of the students. Of course, tutorials are also powerful in this manner, as are labs, workshops, developing and enhancing peer tutoring, peer cooperation in teaching and learning, peer collaboration in so many forms, and many other methods of teaching that allow lecturers to hear who and what their students are thinking and questioning. So also for assessments.

Visible teaching and learning occurs when learning is the explicit goal, when it is appropriately challenging, when the teacher and the student both (in their various ways) seek to ascertain whether and to what degree the challenging goal is attained, when there is deliberate practice aimed at attaining mastery of the goal, when there is feedback given and sought, and when there are active, passionate, and engaging people (teacher, student, peers, and so on) participating in the act of learning. It is teachers seeing learning through the eyes of students, and students seeing teaching as the key to their ongoing learning. The remarkable feature of the evidence is that the biggest effects on student learning occur when teachers become learners of their own teaching, and when students become their own teachers. When students become their own teachers they exhibit the self-regulatory attributes that seem most desirable for learners (self-monitoring, self-evaluation, self-assessment, self-teaching). Thus, it is visible teaching and learning by teachers and students that makes the difference. (Hattie, 2009, p.271)

An important methodological consideration relates to the power of effect-sizes (which I have used above). These easy-to-calculate measures should become more known in the university sector, as they allow us to make the comparisons between different influences on our teaching and student learning. They allow us to evaluate which courses have more “value added” than others, and they can be used to begin the discussion on what we need to change, improve, and drop. Effect-sizes are standardised measures of the difference between two lecturers, and/or between students learning over time. Note the work of Pascarella and Terenzini (2005) who set some benchmarks by using effect-sizes. For example, over the years of undergraduate education, how much change occurs in student development?

Table 4: University effect sizes

Effects of Universities	ES
Quantitative skills	.24
Verbal skills	.56
Piagetian (formal) reasoning	.33
Written communication	.50
Speaking skills	.60
Mathematics courses	.62
Science skills and knowledge	.62
Social sciences	.73
Reading and literature courses	.77
Specific subject matter knowledge	.84
Critical thinking skills	1.00
Reflective judgment-thinking (use of reasons, addressing ill-structured problems)	1.00
Conceptual complexity	1.20

Students not only make major gains in subject matter knowledge and knowledge during their undergraduate years but also become more critical, reflective, and sophisticated thinkers. When compared to the .40 hinge-point, however, we should have reason to pause and reflect on the knowledge and skills compared to the sophisticated thinking, and ask to what degree these latter appear in our learning intentions and success criteria. We can also take courage in promoting these latter more directly in our courses and assessments.

Assessment for learning/ Feedback from assessment

Of all the factors that make a difference to student outcomes, the power of feedback is paramount in any list. The overall effect-sizes of feedback from over 1000 studies based on 50,000+ students reveals that feedback is among the highest of any single factor, and it underpins the causal mechanisms of most of the factors in the top 10-20 factors that enhance achievement.

When I completed the first synthesis of 134 meta-analyses of all possible influences on achievement (Hattie, 1992), it soon became clear that feedback was among the most powerful influences on achievement. Most programs and methods that worked best were based on heavy dollops of feedback. When I was presenting these early results in Hong Kong, a questioner asked what was meant by feedback, and I have struggled to understand the concept of feedback ever since. I have spent many hours in classrooms (noting its absence, despite the claims of the best of teachers that they are constantly engaged in providing feedback), worked with students to increase self-helping (with little success), and have tried different methods of providing feedback. The mistake I was making was seeing feedback as something *teachers provided to students*. They typically did not provide feedback, although they made claims that they did it all the time, and most of the feedback they did provide was social and behavioural. It was only when I discovered that feedback was most powerful when it is from the *student to the teacher* that I started to understand it better. When teachers seek, or at least are open to, feedback from students as to what students know, what they understand, where they make errors, when they have misconceptions, when they are not engaged, then teaching and learning can be synchronised and powerful. Feedback to teachers helps make learning visible.

Recently a colleague and I published a paper devoted to the power of feedback which provides a deeper explanation than can be presented in this chapter (Hattie & Timperley, 2007). But, in summary, feedback is information provided by an agent (e.g., teacher, peer, book, parent, or one's own experience) about aspects of one's performance or understanding. For example, 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 correctness of a response. *Feedback is a "consequence" of performance.*

To assist in understanding the purpose, effects, and types of feedback, it is useful to consider a continuum of instruction and feedback. At one end of the continuum is a clear distinction between providing instruction and providing feedback. However, when feedback is combined with a correctional review, feedback and instruction become intertwined until "the process itself takes on the forms of new instruction, rather than informing the student solely about correctness" (Kulhavy, 1977, p. 212). To take on this instructional purpose, feedback needs to provide information specifically relating to the task or process of learning that fills a gap between what is understood and what is aimed to be understood (Sadler, 1989), and it can do this in a number of different ways. For example, this may be through affective processes, such as increased effort, motivation, or engagement. Alternatively, the gap may be reduced through a number of different cognitive processes, including helping students to come to a different viewpoint, confirming to students that they are correct or incorrect, indicating that more information is available or needed, pointing to directions that the students could pursue, and indicating alternative strategies to understand particular information. Winne and Butler (1994) provided an excellent summary in their claim that "feedback is information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (p. 5740).

The effect sizes reported in the feedback meta-analyses show considerable variability, which indicates that some types of feedback are more powerful than others. The most effective forms of feedback provide cues or reinforcement to the learner, are in the form of video, audio or computer-assisted instruction feedback, or relate feedback to learning goals. It is also worth noting that the key is feedback that is received and acted upon by students; many teachers claim they provide ample amounts of feedback but the issue is whether students receive and interpret the information in the feedback. At best, each student receives moments of feedback in a single day (Nuthall, 2005; Sirotnik, 1983). Carless (2006) asked students and teachers whether teachers provided detailed feedback that helped students improve their next assignments. About 70% of the teachers claimed they provided such detailed feedback often or always, but only 45% of students agreed with their teachers' claims. Further, Nuthall (2005) found that most feedback that students obtained in any day in classrooms was from other students, and most of this feedback was incorrect.

The most systematic study addressing the effects of various types of feedback was published by Kluger and DeNisi (1996). Their meta-analysis included studies of feedback interventions that were not confounded with other manipulations, included at least a control group, measured performance, and involving at least 10 participants. Although many of their studies were not classroom or achievement based, their messages are of much interest. From the 131 studies, they estimated 470 effect sizes, based on 12,652 participants, and the average effect size was $d = 0.38$; 32% of the effects were negative. Specifically, feedback is more effective when it provides information on correct rather than incorrect responses and when it builds on changes from previous trials. The impact of feedback was also influenced by the difficulty of goals and tasks. There is highest impact when goals are specific and challenging but when task complexity is low. Giving praise for completing a task appears to be ineffective, which is hardly

surprising because it contains such little learning-related information. Feedback is more effective when there are perceived low rather than high levels of threat to self-esteem, presumably because low threat conditions allow attention to be paid to the feedback.

A main purpose of feedback is to reduce discrepancies between current understandings and performance and a learning intention or goal. The strategies that students and teachers use to reduce this discrepancy depends partly on the level at which the feedback operates. These levels include the level of task performance, the level of process of understanding how to do a task, the regulatory or meta-cognitive process level, and the self or person (unrelated to the specifics of the task). Feedback has differing effects across these levels.

The major feedback questions are **“Where am I going?”** (learning intentions/goals/success criteria), **“How am I going?”** (self-assessment and self-evaluation), and **“Where to next?”** (progression, new goals). An ideal learning environment or experience is when both teachers and students seek answers to each of these questions. These three questions do not work in isolation at each of the four levels, but typically work together. Feedback relating to “How am I going?” has the power to lead to doing further tasks or “Where to next?” and “Where am I going?” As Sadler (1989) has convincingly argued, it is closing the gap between where the student is and where they are aiming to be which leads to the power of feedback.

So far so good, but the difficulty arises from the way in which feedback works at four levels noted above. First, feedback can be about the task or product, such as the work is correct or incorrect. This level of feedback may include directions to acquire more, different, or correct information, such as “You need to include more about the Treaty of Versailles”. Second, feedback can be aimed at the process used to create the product or complete the task. This kind of feedback is more directly aimed at the processing of information, or learning processes required for understanding or completing the task. For example, a teacher or peer may say to a learner, “You need to edit this piece of writing by attending to the descriptors you have used so the reader is able to understand the nuances of your meaning”, or “This page may make more sense if you use the comprehension strategies we talked about earlier”. Third, feedback to the student can be focused at the self-regulation level, including greater skill in self-evaluation, or confidence to engage further on the task. For example, “You already know the key features of the opening of an argument. Check to see whether you have incorporated them in your first paragraph.” Such feedback can have major influences on self-efficacy, self-regulatory proficiencies, and self-beliefs about the student as a learner, such that the student is encouraged or informed how to better and more effortlessly continue on the task. Fourth, feedback can be personal in the sense that it is directed to the “self” which, it will be argued below, is too often unrelated to performance on the task. Examples of such feedback include, “You are a great student”, “Well done!”

The art is to provide the right form of feedback at, or just above, the level where the student is working – with one exception. Feedback at the self or personal level (usually praise) is rarely effective. Praise is rarely directed at addressing the three feedback questions and so is ineffective in enhancing learning. When feedback draws attention to the self, students try to avoid the risks involved in tackling a challenging assignment, to minimise effort, and have a high fear of failure (Black & Wiliam, 1998) in order to minimise the risk to the self. Thus, ideally, teaching and learning moves from the task, to the processes or understandings necessary to learn the task, to regulation about continuing beyond the task to more challenging tasks and goals. This process results in higher confidence and greater investment of effort. This flow typically occurs as the student gains greater fluency and mastery.

We need to be somewhat cautious, however. Feedback is not “the answer” to effective teaching and learning; rather it is but one powerful answer. With inefficient learners or for learners at the acquisition (not proficiency) phase, it is better for a teacher to provide elaborations through instruction than to provide feedback on poorly understood concepts. If feedback is directed at the right level it can assist students to comprehend, engage, or develop effective strategies to process the information intended to be learnt. To be effective, feedback needs to be clear, purposeful, meaningful and compatible with students’ prior knowledge and to provide logical connections. It also needs to prompt active information processing on the part of the learner, have low task complexity, relate to specific and clear goals, and provide little threat to the person at the self level. The major discriminator is whether feedback is clearly directed to the various levels of task, processes, or regulation, and not directed to the level of “self”. These conditions highlight the importance of classroom climates that foster peer and self assessment, and allow for learning from mistakes. We need classes which develop the courage to err.

Thus, when feedback is combined with effective instruction in university lectures and other forums, it can be very powerful in enhancing learning. As Kluger and DeNisi (1996) noted, a feedback intervention provided for a familiar task that contains cues that support learning, attracts attention to feedback-standard discrepancies at the task level, and is void of cues that direct attention to the self, is likely to yield impressive gains in students’ performance. It is important to note, however, that under particular circumstances, instruction is more effective than feedback. Feedback can only build on something; it is of little use when there is no initial learning or surface information. In summary, feedback is what happens second, is one of the most powerful influences on learning, too rarely occurs, and needs to be more fully researched by qualitatively and quantitatively investigating how feedback works in the classroom and learning process.

Assessment to get into university

There has been much debate about using assessment to make choices about how gets into University and into specific courses. This debate has a long history, although we should be reminded that it is only in the last 40 or so years that ability or achievement has been the foremost consideration. The selection into the more prestigious universities used to be on the basis of choosing “future leaders” and this criterion used measures such as old-boy recommendations, attestations of citizenship, and involvement in cultural and political activities of high interest to universities that were much more important. Indeed, Yale, Harvard and Princeton had quotas (no more than 10%) for those who could enter on the basis of academic performance (Karabel, 2005). The move to meritocracy began with the protest movements in the 1960s. Since, then we have hustled to find the best predictors of success.

The typical finding is that 10% of the variance of the end of first year performance can be accounted for by school related achievement, or ability test scores - and these school measures are among the highest of all predictors we can find! For example, Goldberg and Alliger (1992) completed a meta-analysis of 10 studies investigating the predictability of secondary school achievement against first year University grade point averages (GPA) in Psychology and found an average correlation of .15 between these two scores. In their meta analysis of 22 studies, Morrison and Morrison (1995) found correlations of .22 and .28 between secondary school results and various university GPAs. In a more recent meta analysis Kuncel, Hezlett, and Ones (2001) used 1753 studies and 6589 effect-sizes and found a sample-size-weighted average correlation of .13 to .38 between secondary school results and undergraduate grade point average. These average correlations of .20 and .35 between various measures of high school performance and GPAs from first year university results have been relatively stable over many

years. There is a large literature aiming to understand why these relations are so low (e.g., unreliability of first year examinations, the need to include study skills and personality measures such as effort and perseverance, restriction of range).

One of the more exciting developments in New Zealand has been the introduction of the NCEA. This method, despite its many hiccups in implementation, is fundamentally sound and a model for many other countries. Notwithstanding all the details, the major reform related to identifying many standards within a subject, asking students and teachers to make or choose a mix of internal and external assessments, grading on a criterion or standards-based scoring rubric, and accumulating successes over a variety of experiences during the last three years of high school. This is contrary to the former method which was studying for the last years, and then having a one-shot on one-day at one-examination and accumulating results across subjects (with or without “scaling”). The question is: which method is more like the modern university experience?

Long ago, the single final university examination (with or without “terms”) disappeared. It is perhaps not surprising, therefore, that the NCEA rather than the single final high school exam is immeasurably superior at predicting the results of first year. We found predictive correlations between .5 and .6 from NCEA (Shulruf, Hattie, & Tumen, 2008), indicating to us that the NCEA more appropriately mimics what is valued in undergraduate courses: students learning effort and self-regulated learning makes the difference; students and teachers battling to make more sense of what the standards are prior and during the teaching and learning; an appropriate mix of internal and external assessments; and attention to what is meant by a subject (just look at the comparability of what is offered in the name of a subject across universities and viva la difference); and more attention on the meaning of discrimination between score names (Not Achieved, Achieved, Merit, Excellence in NCEA, A,B,C,D in Universities) with the associate considerations of moderation and constructive alignment.

Conclusions

The major message in this chapter revolves around the power of feedback and the important role feedback plays in the assessment issues in higher education. Particularly the feedback from assessments *to the* instructor about what they taught well, gaps, strengths, to whom in the class, and how their learning intentions were realised by them or not. It is the feedback from assessment that is most paramount and critical. This requires a change of mindset from thinking assessments are *for* and *about* the students; as the feedback from this model has been low level and too time intensive. Instead, we need to move past the “red biro ticks” to optimising the feedback in comments relative to constructively aligning the intentions and the success.

The revolution of computerised essay scoring, the computerised peer critique, the use of in-class clickers (or audience response systems, Bruff, 2009; Draper & Brown, 2004), and the computerised interactive tasks will allow lecturers to spend more time and energies to provide feedback and devoting more time to the arts of teaching and learning. The black box of tertiary assessment is about to be re-opened, re-engineered, and put back together in a totally different manner. We can be on the forefront of a revolution to induce punctuated change in tertiary assessment

Alexander Graeme Bell invented the telephone from a tool he made for teaching the deaf. If he saw the telephone in 1990 he would still recognise it. If he saw it now in 2008 he would not immediately see the connection between what he invented and Skype, internet, faxing, texting, virtual marts, twitter, i-tunes etc. There will be little familiar to Bell. A similar punctuated

equilibrium is close in tertiary assessment. Nineteenth century professors would see many similarities between their assessments and today's university assessment. Over the next decade we will witness the greatest revolution in the role of assessment in tertiary education – it will move from a device to sum up what we think students need to know, to providing feedback into the teaching and learning cycle; it will involve more than surface and greater emphases on deeper knowledge and understanding; it will involve peer assessment and computerised scoring; it will involve aspects of Second Life and interactivity; it will see more use of computerised adaptive testing; and the quality of these assessments will be set higher, the qualities will be more public, and students will be the major beneficiaries of this revolution. The revolution will encompass “*feedback from assessment*” and *the development of visible learning and visible teaching*.

References

- Adams, J. (2008). *Assessment process for the Performance-Based Research Fund*. Wellington, NZ. Tertiary Education Commission.
- Au, R.C.P., Watkins, D., Hattie, J., & Alexander, P. (2009). Reformulating the depression model of learned hopelessness for academic outcomes. *Educational Research Review*. <http://dx.doi.org/10.1016/j.edurev.2009.04.001>
- Ainley, J., & Long, M. (1994). *The course experience survey: The 1992 graduates*. Canberra, ACT.: Australian Government Printing Service.
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 1-18.
- Biggs, J. (1999). *Teaching for quality learning at university*. Buckingham: SRHE and Open University Press.
- Biggs, J. (2003). *Aligning teaching and assessment to curriculum objectives*. Imaginative Curriculum Project, LTSN Generic Centre.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7-75.
- Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. New York: Nova Science Publishers.
- Bruff, D. (2009) *Teaching with classroom response systems: Creating active learning environments*. San Francisco: Jossey-Bass.
- Carless, D. (2006). *How assessment supports learning: Learning-oriented assessment in action*. Hong Kong: Hong Kong University Press.
- Chickering, A. (1969). *Education and identity*. San Francisco: Jossey-Bass.
- Chickering, A., & Reisser, L. (1993). *Education and identity* (2nd ed.). San Francisco: Jossey-Bass.
- Draper, S., & M. Brown (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning*, 20(2), 81-94.
- Goldberg, E. L., & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement*, 52(4), 1019-1027.
- Hattie, J.A.C. (1992). Measuring the effects of schooling. *Australian Journal of Education* 36(1), 5-13.
- Hattie, J.A.C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. London: Routledge.
- Hattie, J.A.C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112.
- Johnson, T. (1998). *The 1997 course experience survey*. Melbourne, Australia: McPherson & Morison.
- Karabel, J. (2005). *The Chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton*. Boston: Houghton Mifflin.
- Kluger, A.V., & 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.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, 47(1), 211-232.

- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127(1), 162-181.
- Marsh, H.W., & Cheung, J. (2007). *National student survey of teaching in UK universities: Dimensionality, multilevel structure, and differentiation at the level of University and discipline*. Interim Report. Oxford, UK: Oxford University Press.
- Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the graduate record examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement*, 55(2), 309-316.
- Nuthall, G. (2005). The cultural myths and realities of classroom teaching and learning: A personal journey. *Teachers College Record*, 107(5), 895-934.
- Page, E.B. (1985). Computer scoring of essays. In T. Husen & T.N. Postlethwaite (Eds.), *The International Encyclopaedia of Education*, Vol. 2. Oxford: Pergamon Press.
- Pascarella, E.T., & Terenzini, P.T. (2005). *How college affects students: Vol. 2: A third decade of research*. San Francisco: Jossey-Bass.
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The course experience questionnaire. *Studies in Higher Education*, 16, 129-150.
- Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science* 18(2), 119-144.
- Shermis, M.D., & Burstein, J.C., (Eds.). (2003). *Automated essay scoring: A cross- disciplinary perspective*. Hillsdale, NJ: Erlbaum.
- Shulruf, B., Hattie, J., & Tumen, S. (2008). The predictability of enrolment and first year university results from secondary school performance studies in higher education. *Higher Education*, 56(5), 613-632.
- Sirotnik, K.A. (1983). What you see is what you get: Consistency, persistency, and mediocrity in classrooms. *Harvard Educational Review*, 53(1), 16-31.
- Winne, P.H., & Butler, D.L. (1994). Student cognition in learning from teaching. In T. Husen and T. Postlethwaite (Eds.), *International encyclopedia of education* (2nd ed., pp. 5738-5745). Oxford: Pergamon.

Email: j.hattie@auckland.ac.nz