Assessment & Evaluation in Higher Education

A study of sustainable assessment theory in higher education tutorials

Robert J. Beck [a] , William F. Skinner [a] & Lynsey A. Schwabrow [a]

[a] Lawrence University, Appleton, WI, USA

fiFirst

PLEASE SCROLL DOWN FOR ARTICLE

A study of sustainable assessment theory in higher education tutorials

Robert J. Beck, William F. Skinner and Lynsey A. Schwabrow*

Lawrence University, Appleton, WI, USA

A study of sustainable assessment theory in nine tutorial courses at four colleges demonstrated that three long-term learning outcomes improved: Independence, Intellectual Maturity and Creativity. Eight of 10 traits associated with these outcomes were validated through internal reliability, faculty and student rubrics, and faculty case studies reporting pedagogic innovations and improvements of student abilities in self-assessment. The findings suggest that sustainable assessment theory should be applied using methods encompassing a strong commitment to equity, including shared criteria for long-term learning outcomes and faculty and student monitoring of student progress towards outcomes through periodic rubrics and reflective sessions.

Keywords: sustainable assessment; tutorial; tutorial learning outcomes; liberal arts; case studies

Introduction

This study responded to recent criticism that questioned whether undergraduates made adequate gains in intellectual development during their college careers (Arum and Roksa 2010; Spellings 2006). Gains in intellectual development refer to long-term learning skills that 'equip students to learn beyond the academy once the infrastructure of teachers, courses, and formal assessment is no longer available' (Boud and Falchikov 2006, 399). Long-term learning abilities do not refer exclusively to content knowledge but rather concern 'habits of mind' and metacognitive skills that embody cognitive and social cognitive abilities that are useful in improving students' learning skills. We selected for study long-term learning skills that enable students to learn on their own, approach problems from multiple perspectives, and work with complex issues. However, higher education lacks assessment methods for determining whether students have made gains on these traits over their undergraduate careers. This problem dovetails with recent calls by prominent educators (Association of American Colleges and Universities [AAC&U] 2008; Bok 2006; Connor 2006; Katz 2008) for new methods of assessment and that colleges develop quality standards themselves and not rely on external agencies. These quality standards refer to learning goals and assessments *anchored in the tutorial and liberal arts course curricula*. Sustainable assessment theory (Boud 2000; Boud and Falchikov 2006) holds that educators should use explicit long-term learning quality standards as criteria for making judgements about progress in learning and involve

*Corresponding author. Email: lynsey.schwabrow@lawrence.edu

students in self-assessment for the purpose of becoming assessors of their own learning.

Many colleges have answered this criticism by undertaking research projects to determine whether and how undergraduates developed socially/intellectually from their entry to graduation (see http://www.teaglefoundation.org for an extensive review of assessment research projects). Katz (2008) discussed outcomes relevant to taking the true measure of a liberal education:

> Most advocates of liberal education deny that a primarily content-based evaluation of learning assesses the totality of a senior's educational experience. Most of us are committed to the notion that liberal learning has more to do with cultivating qualities of mind and the capacity to recognize and analyze significance than with the mastery of any quantum of information. (32)

In a speech in Washington DC, Connor (2006), former president of the Teagle Foundation, also supported increased focus on the assessment of student learning in liberal arts colleges. Of primary interest to the present problem is the trend he cited of moving from reliance alone on assessments of *content* or *subject matter*, called achievement goals by educators, to *cognitive goals*, in which enduring cognitive capacities are measured as outcomes. Connor concluded that unless we understand cognitive capacities in the context of relevant content these new goals might prove to be unduly abstract. Thus, a consensus has developed that new assessment methods are needed to measure students' long-term learning in higher education.

Literature review

Sustainable assessment theory

Sustainable assessment theory is an emerging approach to assessment complementing summative and formative assessment methods (Boud 2000; Boud and Falchikov 2006). The premise is that assessment needs to be brought into alignment with teaching and learning for the purpose of equipping students to assess their abilities to learn in a variety of non-academic, relatively unpredictable contexts after graduation. Sustainable assessment is part of a 'constructive alignment' between the teaching system and assessment tasks in which the latter are part of teaching and learning. In this approach, students need to become more active participants in assessing their own learning (Boyd and Cowan 1986). By contrast, summative assessment is a traditional end of course or other instructional activity grade or other quantitative measure of accomplishment made unilaterally by instructors that certifies that some level of learning has been achieved. Formative assessment refers to qualitative feedback by instructors that is offered to students during learning activities, such as discussion, that are not intended to provide conclusive measures but rather as part of teaching that helps students grasp particular aspects of subject matter or approaches to learning.

Boud and Falchikov (2006) criticise both summative and formative assessment because they place students in the position of 'always attending to the judgements of others' and preventing students from having the 'opportunity to see how the process of assessment actually works' (403). Traditional assessments undermine the independence of students in making their own judgements. In comparison, sustainable assessment theory proposes to move beyond summative and formative assessment by positing that students should be more actively involved in their own

assessment by increasing their participation both in the process of identifying assessment criteria and in making judgements themselves. Black and Wiliam's (2009) approach to formative assessment borders on sustainable assessment in that it does include increased involvement by students in 'clarifying and sharing learning intentions and criteria for success', 'activating students as instructional resources for each other' and 'activating students as the owners of their own learning' (8). Perhaps, the most significant new features in sustainable assessment theory that distinguish it from formative assessment would be, in principle, to develop in students the ability to be *sustainable assessors* of their own *long-term learning skills* and to develop *assessment devices* for student self-monitoring. The findings of the study will suggest, however, that formative assessment of the kinds described by Black and Wiliam (2009) will emerge during educational activities that employ sustainable assessments.

Social cognitive and metacognitive learning outcomes from tutorials

While the purposes of sustainable assessment are evident, the nature of long-term learning outcomes requires clarification. One pertinent question concerns the type of academic context in which one might be able to observe and measure long-term learning outcomes. There is evidence to suggest that long-term learning might be supported and more visible in individualised instruction settings in which both the social and cognitive skills of students are more in evidence. In a study of tutorial education at the University of Oxford, Beck (2007) supported Moore's earlier Oxford study (1968) that the principal learning outcomes of tutorials were to develop students' abilities to think for themselves, work independently, develop a sceptical orientation, acquire mental flexibility, demonstrate creativity and imagination, learn to argue, engage in continuous self-assessment, and produce a documented example of original work. The kinds of student outcomes represented social cognitive and metacognitive alternatives to the way student learning has often been assessed in education through examinations of content knowledge.

Oxbridge and American liberal arts tutorials typically involve 2–4 students and a tutor in a variety of subjects. Tutorials feature close and repeated interactions and intensive oral and written communications between faculty and students, enabling faculty to assess changes in student cognitive and metacognitive capabilities. Because of the presence of nearly continuous assessment in the dialogic method of tutorials, in which questioning and the assessment of responses serve both to offer students feedback and guidance, we may readily explore the relationships between teaching and assessment. Tutorials draw on several comparable records pertinent to student learning, such as a series of essays and related discussions in a disciplinary area. Because they take place in intimate, easily observable and recordable settings, tutorials could, therefore, provide data-rich environments to research the interaction of cognitive and content goals in liberal education. Simple assignment of grades and correction of papers do not do justice to the richness of student development in tutorials. Assessment of tutorials requires greater attention to enduring learning abilities.

Marzano (1992) defined *habits of mind* as mental dispositions or traits individuals can develop to render their thinking and learning more self-regulated. Several of Marzano's habits of mind are *metacognitive* and employed regularly during intellectual work in tutorials, such as being aware of one's own thinking, self-assessment

of one's actions and being sensitive to feedback. A student's metacognitive abilities refer to his or her active control in thinking or reasoning about thinking, and thinking about how one learns (Flavell 1979). Metacognition entails strategies for planning, monitoring and evaluating progress towards learning goals. Such techniques as self-questioning and self-assessment are considered vital to the development of students' ability to engage in higher-order thinking and self-regulated learning (Brown 1987). In tutorials, both faculty and students are often engaged in clarifying what students are thinking and in making continuous formative assessments of the effectiveness of their work in writing, oral presentation and discussion. And, Costa and Kallick's (2000) taxonomy of habits of mind describes the kinds of problems – dilemmas and uncertainties – that are part and parcel of typical student assignments in writing and discussing essays in tutorials.

Tutorials and sustainable assessment theory

There is evidence to suggest that tutorials may represent a superior form of education at any stage of education. In a recent review of the role of dialogue and tutoring in education, Hacker and Graesser (2007) wrote, 'The superiority of one-on-one tutoring over traditional classroom instruction has been well documented' (259). In an experimental study of K-12 students, Bloom (1984) demonstrated that tutorial instruction had two-standard deviation superiority over all other forms of instruction. Cohen, Kulik and Kulik (1982) conducted an excellent meta-analysis of several studies that evaluated tutorial education and came to the same conclusion, although tutorial superiority did not reach the astounding levels found by Bloom. Yet, Hacker and Graesser (2007) cautioned that 'the identification of exactly what aspects of one-to-one tutoring contribute to this superiority has received little attention' (259). Nevertheless, most colleges in the Annapolis Group – an organisation of 130 of the leading liberal arts colleges in the USA – have some form of one-on-one instruction to teach disciplinary subjects and to conduct independent studies and research, and individualised instruction is often cited as a distinctive feature of liberal education.

If as the literature suggests, tutorials are, in fact, superior in supporting student learning, and we can determine which aspects of one-on-one tutoring contribute to this superiority, then tutorials should also prove to be a highly advantageous research environment for observing and measuring social cognitive and metacognitive long-term learning outcomes and, hence to research and develop sustainable assessment. Setting goals of social cognitive and metacognitive learning outcomes for students and providing them with comprehensive assessments could improve the pedagogy of tutorials as well as contribute to the development of sustainable assessments.

Long-term learning outcomes from tutorials

The long-term learning outcomes in this study were selected through a variety of methods. Independent Thinker has been identified as an outcome of tutorials and one of its traits, self-assessment, is integrally related to the goals of sustainable assessment. Creativity had been studied in a consortium of colleges that sought to identify long-term learning outcomes in liberal arts colleges. While the overall findings of The Five Colleges of Ohio Creative and Critical Thinking Project (2009), including one college in this study, indicated that creativity was not closely

associated with the classroom, and first year students tended to see more possibilities for creative thinking campus-wide than seniors, 'as students progressed through the curriculum, their belief that creativity can be taught [and hence could be assessed] increases'. Moreover, the investigators concluded that 'participating faculty believe that teaching practices are most likely to stimulate creative thinking if they include "active learning" techniques that facilitate student engagement with course material . . .' (3–5). Tutorial courses qualify as active learning because students assume teacher roles in presenting and discussing course material. In an online survey in this project, faculty indicated that using creativity rubrics helped them learn more about creativity, influenced subsequent teaching and were shared with colleagues (21). In exit interviews with the faculty they expressed the positive values associated with developing and working with rubrics to assess creativity:

> The rubric made me much more aware of students' weaknesses in certain areas [of creativity]: I realized that if I genuinely valued creativity, I should construct assignments that facilitated its expression'; 'I paid closer attention to the criteria used to evaluate the assignments'; 'Don't laugh . . . but it helped me to see some of what may be the motivation for the movement toward assessment. (22)

Based on these results and because two of our participating professors were explicitly concerned with creativity in their courses (Studio Art and Art History), the participating faculty decided to develop and test rubrics assessing creativity in the present study.

During a two-day conference on student learning outcomes in tutorials, a group of nine professors from four liberal arts colleges also selected Intellectual Maturity, the ability to deal with complexity and taking risks, as among the most important goals of college education.

We propose that Independent Thinker, Intellectual Maturity, and Creativity as articulated below meet several of Boud's (2000) criteria for long-term learning skills so that students: (a) can learn on their own; (b) deal with super complexity and ambiguity in work and life situations; (c) act effectively in new situations; (d) be self-initiating and (e) become assessors of their own learning in different knowledge domains and in different circumstances.

Independent Thinker

We theorise that student dispositions to thinking for oneself are expressed as Independent Thinker in their approaches to learning. A student can learn on his/her own and with others dialogically by finding sources of information, employing individual strategies for acquiring ideas from sources and combining this knowledge with his or her own prior knowledge in creating new knowledge. We examined four traits of the Independent Thinker as they are commonly expressed in tutorial settings:

Independence

The ability to take teacher roles by setting topics, asking questions to originate new topics, summarising discussion and assessing ideas

Tutorials support the development of the student as an Independent Thinker because the tutorial relationship, in principle, has the *ideal of teacher-student equality.*

Students become Independent Thinkers by assuming the teacher role during tutorials and this undercuts the negative influence of hierarchic power in higher education. As Oxford tutor Mirfield (2001, 38) stated: 'My own tutor never taught otherwise than as his equal'. When students present their essays to the tutor they are acting/performing their argument, as if they are lecturing from a prepared text just as teachers do. However, their presentations are ones in which the audience may constantly interrupt and engage them in discussion about the very line or previous argument they have uttered. In this role, the students learn that in their presentations teachers must be prepared to justify and defend their propositions and their supporting evidence. If there are two or more students per faculty, then each one has the opportunity also to teach his/her peer. Our position, therefore, is that faculty, with various levels of awareness, seek to balance their structurally superior roles by placing students in the teacher or co-teacher role. One way they do this is by creating an environment in which teachers and students work together to help students think for themselves.

Developing an inquiring mind

Questioning ideas and self-questioning; asks high-level questions, such as why-questions

Teacher questioning serves both a formative assessment function and a scaffold that points students in directions they have hitherto not considered. In this latter sense the teacher uses questions to explore the various subtopics of the primary question or problem of the student essay. But the teacher tends not to disclose these subtopics didactically, so much as through indirect teaching by probing or asking. This encourages the student to do the work relatively independently. When a teacher asks a student a question, such as a why-type question, that has no pre-existing, informational answer, the student will think for him(her)self in formulating a reply. If as often occurs the reply is incomplete or displays error in some way, then the teacher's follow-up questions ask students to think further to refine their propositions.

Acquiring self-assessment skills

Self-assesses one's claims and arguments; the ability to recognise biases in thinking and predispositions to make certain judgements; shares or communicates critique of one's biases and prejudgements

Student responses to faculty questions provide material for assessment and other forms of feedback. We assume that students' long-term self-assessment traits derive from the continuous formative assessment they receive in tutorials. Richard Mash, a Fellow in Economics at Oxford, proposed that tutorials are exceptional settings for providing extensive feedback:

> Tutorials should offer excellent opportunities for feedback that is positive (while always being honest), and the frequency of feedback should help the process whereby students settle in mentally and feel that, subject to the required effort, they can be successful. (Mash 2001, 91)

Emma Smith, a Fellow in English at Oxford, described her role in the tutorial as less of a teacher than as a critic, who engaged in assessment more or less

continuously (Smith 2001). Mayr-Harting (2006), an Emeritus Regius Professor of Ecclesiastical History at Oxford, offered some excellent advice about criticism and feedback on students' essays:

> A pupil needs to hear why the tutor *thinks* it is a good essay ... to understand what *makes* a good piece of work a good piece of work; even for weak essays, the tutor should build up the credit column all he or she can before going into the debit column. (5)

Thus, the extreme density and fine calibration of assessments offered to students lead not only to a propensity for both teachers and students to use assessments constructively but for students to acquire the trait of self-assessment and the practice of seeking assessments from others.

Learning to argue

Argues effectively by making conceptual claims backed with supporting theory, reasoning and evidence; does not imply emotional and contentious argument

During tutorial discussion, argument serves as a means for the production of constructive activities such as collaborative learning. Argumentation ensues when there is communication about 'an issue that has two sides and which provides for two opposing communicator roles: a protagonist who puts forward a claim and an antagonist who doubts that claim, contradicts it, or otherwise withholds assent' (van Eemeren et al. 1997, 209). The process by which argument is exposed and shaped is necessarily dialogic and dialectical (oppositional) because it requires an extended series of questions and feedback to help students understand how intricate arguments may be parsed and constructed. During tutorials, the cultural rules of argumentation, mediated by tutors, serve to assess student arguments.

Intellectual Maturity

Through extensive discussion and debate, the nine faculty participants in this study generated a selection of traits based on what they expected students to learn from the tutorial experience. Different instructors variously defined tutorial learning outcomes and their associated traits and provided different behavioural criteria for the same trait. As the discussion proceeded degrees of consensus were reached on the nature of learning outcomes, and how students might be classified in acquiring such traits in the tutorial setting. As a result, the faculty proposed Intellectual Maturity as a new contribution to the assessment rubrics. One trait, *takes intellectual chances*, was adapted from the trait of risk-taking from The Five Colleges of Ohio Creative and Critical Thinking Project (2009). Another trait, *complexity*, was selected because students often dealt with complex problems in their tutorial essays. The faculty consensus was that two important traits of Intellectual Maturity should be tested in the study:

(1) *Complexity*: the ability to work with complex problems, issues, and information.
(2) *Takes intellectual chances*: willingness to state positions and arguments without worry of saying something wrong, making mistakes, or risking failure.

Creativity

The Five Colleges of Ohio Creative and Critical Thinking Project (2009) used surveys, focus groups and interviews with faculty to research characteristics associated with long-term creative processes and products. The participants produced well-designed and validated rubrics for use in a variety of educational settings to assess creative and critical thinking and to foster more effective pedagogies. The basis for skills or abilities associated with the creativity traits assessed in this study was derived from their rubric to measure creative and critical thinking. The four traits assessed in this project include:

(1) *Idea generation*: generates new ideas, variations of or alternatives to solving problems, a novel way of analyzing or re-conceptualizing a topic or idea in the context of what the student knows and understands, interesting and creative restatements of others' ideas, unusual ideas, interesting theories.
(2) *Curiosity*: the desire to learn or know more, ability to become absorbed in the topic, discovers a new line of inquiry or question of a topic and wishes to persist and sustain in exploring the topic.
(3) *Multiple perspectives*: sees a problem from multiple perspectives, compares and contrasts approaches, uses multiple disciplines.
(4) *Connectivity*: ability to bring together or synthesize disparate bits of information, makes connections between already established ideas or theories, connects disciplines.

Integrating sustainable assessment theory and method

According to Boud (2000) and Boud and Falchikov (2006, 407–10) sustainable assessment theory requires the following: focus on long-term learning outcomes that are applicable not only to course activities but also to the workplace; explicit criteria defining student outcomes; co-participation by students and teachers in making judgements in assessment activities; and, the 'development of devices for self-monitoring and judging progression toward goals' (Boud 2000, 161). We selected specific long-term learning trait outcomes as introduced above and tutorial settings for studying such outcomes. But, a method for integrating sustainable assessment theory into tutorial practice remained to be developed. We were in need, therefore, of a 'standards-based framework to enable students to view their own work in the light of acceptable practice' (Boud and Falchikov 2006, 407). The standards in the present study were comprised of the definitions of the particular traits of Independent Thinker, Intellectual Maturity, and Creativity and the acceptable practice would have to be judged according to prevailing practices in tutorials. The tutorial settings provided what Boud called 'active engagement with learning tasks with a view to testing understanding and application of criteria and standards' (161).

Moreover, effective sustainable assessment would have to 'provide practice in discernment [for faculty and students] to identify critical aspects of problems and issues and knowledge required to address them' (Boud and Falchikov 2006, 408). The approach would have to provide training and practice in identifying, developing and engaging with criteria and standards. In other words, students as well as teachers would have to analyse and judge their own behaviour traits of Independent Thinker, Intellectual Maturity, and Creativity. We would have to communicate and

establish up front the 'goal of training and practicing self-assessment abilities of these traits'. It was particularly apt in this regard that self-assessment was both a trait held to be relevant to outcomes of tutorials (an Independent Thinker trait) and required in sustainable assessment. Because tutorials usually eschewed letter grading, at least until the end of the course, the method could incorporate another criterion: 'grades and marks [should be] subordinated to qualitative feedback' (409). As such, the method we developed fostered in depth consideration of the epistemology of sustainable assessment as aligned to teaching and learning.

Sustainable assessment theory also emphasises that assessment takes place in association with others. The high presence of interaction in tutorial courses, in effect, was identified as a determining factor that indicated that the target traits could be developed in that context. Because tutorial courses often involve both teachers and peers, the new method could involve giving and receiving feedback from people with whom students worked collaboratively. While some criteria and training could be established in advance of tutorial courses, other features suggested by Boud and Falchikov might have to emerge in practice. The emergent properties could involve the use of greater student participation through practice in assessment, a more intensive role for collaboration, the addition of reflection as an element of the process as would occur, for example, for students in filling out rubrics or discussing criteria in class during reflective sessions, and specific reference to the role of interaction in observing whether criteria had been met. While we could encourage reflective assessment with peers, much would depend on actual circumstances that provided opportunities for fostering such peer critique. From the teachers' perspective another emergent property might well be 'finding appropriate assistance to scaffold understandings' (408), strategies that teachers could innovate during the course, which might encourage students to demonstrate independence, or creativity, and intellectual maturity.

Finally, Boud (2000) called upon educators to develop 'strategies' and 'devices' for judging whether progress was being made towards outcomes.

> This involves the use and development of a range of strategies and devices deployed in the process of learning. These include everything from the setting of intermediate goals and checking progress at regular intervals, to the keeping of learning journals, to more sophisticated meta-cognitive devices. Not only is it necessary to know what are the appropriate standards and criteria, it is necessary to be able to detect the extent to which the work one has produced meets them. (161)

We regard the 'Shared Assessment Method' (SAM) of our project as integrating such learning techniques.

Method

The purpose of this study was to research and develop sustainable assessment for tutorial courses in liberal arts education. The study was designed to answer three research questions:

Q1. Were the traits validated during the tutorials conducted by faculty?

    Q1a. Did student traits improve during tutorials? If traits improved this would demonstrate that rubrics were tapping coherent learning outcomes.

> Q1b. How internally reliable were the traits for each learning outcome? If groups of traits had reliability then they were, in fact, addressing related aspects of long- term learning goals.

Q2. Did the use of traits as outcomes and rubrics enhance the pedagogy in tutorial courses from the perspective of faculty and students? If traits and rubrics were effective they should become integrated into course design, instructional roles and peer interaction strategies.

Q3. Did the use of traits inform the design of assessments of other non-tutorial courses? If traits could only be measured in tutorials, then it would suggest that only in such contexts was there sufficient interaction to allow traits to be expressed and visible.

Procedure and data collection

This research was conducted over a two-year period during which an iterative process was used to develop the long-term learning outcomes and associated traits, assessment rubrics and case study guidelines that were tested by faculty and student participants.

Proposing and assessing traits for inclusion in the study

During two conferences the nine faculty participants discussed the learning outcomes and associated traits to determine whether they were serviceable in terms of their own experiences of tutorials. To further validate the constructs, they were assessed for importance and scalability by the faculty. Those that were most applicable to the faculty and disciplines involved were determined. As previously outlined, the learning outcomes and associated traits that emerged from extensive discussions and two cycles of testing the rubrics through student and faculty assessment in tutorial courses were as follows:

Independent Thinker: (1) Independence; (2) Developing an inquiring mind; (3) Acquiring self-assessment skills and (4) Learning to argue.

Intellectual Maturity: (1) Complexity/Uncertainty and (2) Taking intellectual chances.

Creativity: (1) Idea generation; (2) Curiosity; (3) Multiple perspectives in problem solving and (4) Connectivity.

Learning outcome rubrics

Rubrics measured observations of the traits for each learning outcome through ratings for which faculty and students used a 5-point scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Frequently and 5 = Very frequently) to record the frequency in which each trait was observed across written work and oral presentations. Each trait included zero as a rating possibility, signifying that the trait was not applicable to observations of the work. Faculty were also given the opportunity to record detailed qualitative observations of the traits exhibited by their tutorial student(s), and were encouraged to record additional traits observed or fostered during the course.

Assessment rubrics were completed at three intervals throughout each tutorial course: a baseline assessment during the second or third week of class, a midpoint assessment during the fifth to eighth week of class, and a final assessment during

the last weeks of class. During each assessment period, students were instructed to self-assess their progress by completing learning outcome rubrics, and faculty completed rubrics for each student enrolled in the tutorial.

Case studies

The 10 observable traits contained in the rubrics for the three student learning outcomes were validated through action research in which faculty developed case studies of their tutorials. The data for the case studies included student work (papers, presentations, exhibits), learning outcome rubrics, and transcripts from reflective sessions between faculty and students that were recorded (a session early in the course; midpoint; and, towards the end). Some tutorials collected a fourth transcript. Drawing on portfolios of faculty and student self-reported ratings, and observations and analyses of student performance and work during tutorials, validation of the traits was sought through the correlation and triangulation of both subjective survey evidence and objective evidence, such as textual passages found in student essays and transcripts of tutorial discussions.

Forms of assessment in the study

While we used sustainable assessment theory to guide our methods, including outcome criteria specifications, joint faculty-student rubrics, and course and teaching strategies, some methods might better be classified as formative and summative.

Formative methods. The early rubrics at baseline and midpoint and the orientation and midpoint reflective sessions were formative. Teaching strategies that appeared spontaneously during the course were formative for the faculty and students. Student self-assessments during the course also served a formative role for students. Transcripts made early in tutorial courses were used to monitor progress formatively.

Summative methods. The final rubrics, exit reflective sessions, transcripts and faculty case studies provided summative information on the outcomes of the programme and were used to validate the theory of sustainable development.

Participants

Nine faculty and 20 third and fourth year tutorial students participated in the study. In consultation with the Dean and Provosts, the faculty were recruited from four private liberal arts colleges in the USA and represented departments in four-year undergraduate programmes in the fine arts, humanities, natural sciences and social sciences.

Results

Rubric analysis

Q1a. Did student traits improve during tutorials? If traits improved this would demonstrate that rubrics were tapping coherent learning outcomes

There are two analyses that demonstrate student traits improved during tutorials. Figures 1 and 2 show the faculty mean ratings for each trait across three assessment periods – baseline, midpoint and final. Table 1 contains the student mean ratings
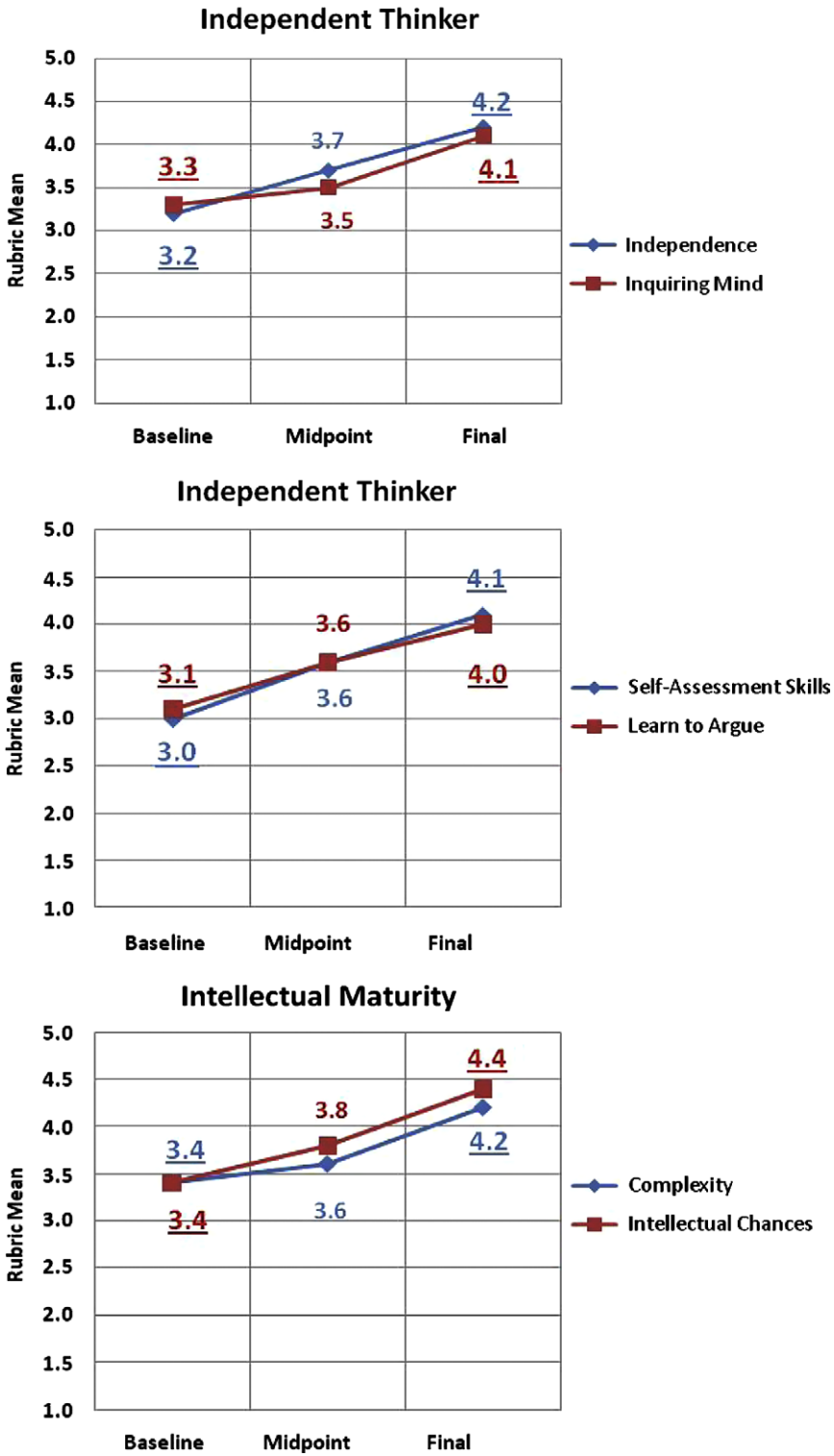
Figure 1.   Faculty rubrics: Independent Thinker and Intellectual Maturity.
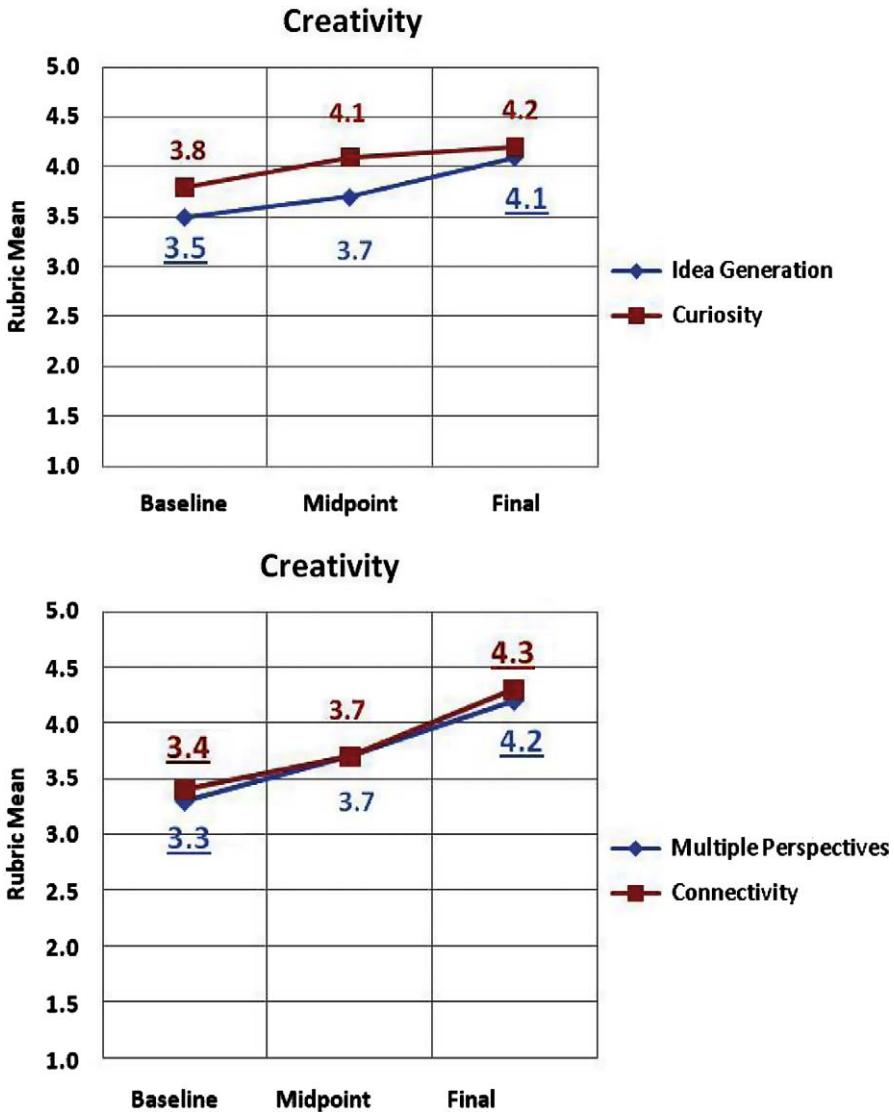
Figure 2.   Faculty rubrics: Creativity.

and a comparison of the faculty and student mean ratings for the traits of the three learning outcomes.

Based upon a paired *t*-test analysis, all of the traits associated with Independent Thinker (Figure 1) showed statistically significant improvement from baseline to final as indicated by the underlined means. It is interesting to note that all the traits for Independent Thinker started at baseline between 3.0 and 3.3 (indicating observed 'Sometimes') and increased at the final measurement between 4.0 and 4.2 (indicating observed 'Frequently'). The mean rating for each trait of Intellectual Maturity also significantly improved from baseline to final. Figure 2 shows that the only trait associated with Creativity that did not significantly improve between baseline and final was *curiosity*. The baseline mean of 3.8 for *curiosity* is the highest

Table 1.   Faculty and student baseline, midpoint and final rubric means.

Trait	Faculty means ($N = 20$)			Student means ($N = 20$)		
	Baseline	Midpoint	Final	Baseline	Midpoint	Final
Independent Thinker						
Independence	3.2	3.7	4.2	3.3[a]	3.9	3.9[a]
Inquiring mind	3.3	3.5	4.1	3.5[a]	4.0	4.3[a]
Self-assessment skills	3.0[b]	3.6	4.1	3.8[ab]	4.2	4.3[a]
Learn to argue	3.1[b]	3.6	4.0	3.6[ab]	3.8	4.1[a]
Intellectual Maturity						
Complexity	3.4	3.6	4.2	3.7[a]	4.1	4.3[a]
Intellectual Chances	3.4	3.8	4.4	3.7[a]	4.0	4.3[a]
Creativity						
Idea generation	3.5	3.7	4.1	3.3[a]	3.9	4.1[a]
Curiosity	3.8	4.1	4.2	4.3	4.6	4.7
Multiple perspectives	3.3	3.7	4.2	3.7	3.9	4.1
Connectivity	3.4	3.7	4.3	3.6[a]	4.2	4.4[a]

Notes: [a]Baseline faculty and student means significantly different at $p < 0.05$. [b]Baseline faculty and student means significantly different at $p < 0.05$. Baseline and final student means significantly different at $p < 0.05$.

starting point of all other traits. Thus, students enter the tutorial with a relatively high level of 'desire to learn or know more'.

The analysis in Table 1 is based on the total number of students who completed the baseline, midpoint and final rubrics and the corresponding faculty ratings of those students. According to the student self-ratings, all traits showed improvement from baseline to final. Based upon a paired *t*-test analysis, statistically significant improvement (indicated by superscript 'a') in the traits from baseline to final occur for all the traits of Independent Thinker and Intellectual Maturity and the Creativity traits of *idea generation* and *connectivity*. Based on an independent sample *t*-test analysis, statistically significant differences (indicated by superscript 'b') in faculty means compared to student means only occur at baseline for *self-assessment skills* and *learn to argue*. And, it is interesting to note that as with the faculty assessment, the highest baseline assessment coming from students occurs for *curiosity* and the rating does not significantly increase by the end of the tutorial.

Q1b. How internally reliable were the traits for each learning outcome? If traits had reliability then they were, in fact, addressing related aspects of long-term learning goals

The reliability analysis indicated that the traits associated with the learning outcomes were highly interrelated for both the faculty and student self-assessments at the baseline and final time periods. For faculty, Cronbach's alpha coefficients for the faculty baseline and final assessments were 0.71 and 0.95 for Independent Thinker, 0.87 and 0.90 for Intellectual Maturity, and 0.93 and 0.95 for Creativity, respectively. For students, the alpha coefficients for baseline and final assessments were 0.69 and 0.65 for Independent Thinker and 0.85 and 0.74 for Creativity. For Intellectual Maturity, the alpha coefficients for the baseline assessment were 0.68 but only 0.10 for the final assessment. Further analysis indicated there were three students who rated themselves either high on complexity and low on taking intellectual chances or visa-versa. When those cases were removed from the analysis, the alpha coefficient was 0.66.

Case studies analysis

Q2. Did the use of traits as outcomes and rubrics enhance the pedagogy in tutorial courses from the perspective of faculty and students? If traits and rubrics were effective they should become integrated into course design, instructional roles and peer interaction strategies

In their case studies, the 10 traits were generally described by faculty as 'crucial', 'extremely important' and 'foundational' to outcomes in liberal arts education. There were a few recommendations to consolidate some Intellectual Maturity and Creativity traits but no faculty suggested eliminating any of the three groups. Some individual faculty members judged that a few of the traits were less relevant to their particular courses.

Some faculty used traits to re-envision the goals of the course at several levels: how courses were conceived, course design and teaching strategies. They offered that the rubrics expanded their ideas concerning the potential for student development. Most faculty found that peer interactions of 2–4 students, a characteristic of most of the tutorials, brought the traits into greater visibility and enabled faculty to

adapt instruction for the traits. Other faculty found that use of the traits 'freed them up' from constantly evaluating through conventional letter grading.

At a *tutorial conception level* the following quotes from case studies reveal how the traits and rubrics changed faculty thinking:

> Working with SAM has absolutely realigned this perspective of mine. I am able to envision the goals of this tutorial (and others) both more broadly and more specifically. In other words, I understand the potential for student development to be well beyond the parameters I first envisioned, yet I have a much more sophisticated understanding of the varied levels of experience that make for that development.

> Perhaps the most beneficial aspect of this study was having the tutorial students go through the rubric at the beginning of the year and discuss it with the instructors. The rubric led to wonderful discussions between instructors and students, where all parties were able to articulate some clear goals for the upcoming senior independent study project ... I believe that this is an educationally sound practice – to give a student the rubric by which he/she will be assessed at the onset of the project, and to re-evaluate at mid-tutorial and at the end of the tutorial. If this process would take place with all tutorials, and honest feedback were to be given, then the grading process at the end of the year would not seem so mysterious.

At a *course design level*, the faculty were able to more or less seamlessly integrate the traits into their course plans. Faculty embedded specific traits into their assignments such as requiring students to assume teacher roles by making presentations (Independent Thinker) or ensuring that students had required readings that had a high level of *complexity*, a trait of Intellectual Maturity:

> I found SAM to be useful in assessing how the students progressed through the course of the tutorial and how well the tutorial succeeded as a whole. Carefully articulated learning outcomes can be a useful tool, but there needs to be a more regular way of keeping students focused on them. Filling out the rubrics after each tutorial session would be one way to do this. I feel confident that, at the very least, I will consult them (rubrics) about how to assess students in future courses, both tutorial and non-tutorial.

From the perspective of *teaching*, the faculty found that they were able to employ particular strategies to encourage the development of traits in students. These strategies were implemented in every aspect of the instruction including the selection of works to be read and analysed, conversations, written comments on papers, private meetings and peer-to-peer advice. Questioning was the most ubiquitous strategy, but also teaching interventions occurred in student evaluations and in class feedback to encourage student development of the learning outcomes. Here are a few examples of tutor teaching strategies.

> Independence (taking teacher role) was among the most influential traits, if not the most important. I often invoke the aphorism, 'to teach is to be twice taught', with my students, and so they have been primed to think how leadership in the classroom both evidences and solidifies one's learning. Part of the influence of this trait is in setting the role of the tutor. Though I remain poised to enter a discussion, and do enter as needed, the trait reminds me to let there be silence. Sometimes it requires a little theatrical touch, such as looking at my notes or something as if to emphasize that I am

not going to jump in on the conversation any time soon. In the vacuum, I am more likely to get students to initiate new directions.

While strategies for both the discussion and writing were offered, most of the teaching innovations occurred during the discussion phase. One faculty's innovation concerning writing, however, strikes us as valuable. When students review their own papers and come up with two problematic aspects, they are taking both the teacher role and self-assessment role, just as they do in presenting their papers to the group and in leading the discussion.

As reported in the case studies, the validity of the relevance of the traits in tutorials was born out in *students' responses to trait-based pedagogic strategies from faculty and peers*. The faculty reported that most students improved their levels of traits. These data were corroborated by positive changes in the levels of traits as analysed in the rubric scores. Students offered additional evidence for trait validity by integrating the rubrics into their learning goals. Some students checked with faculty that their behaviour indicated one trait or another. Other students reflected on their progress on certain traits. Others specifically praised rubrics for helping them with their papers or increased abilities to self-assess and make complex arguments. Students contributed to peers' growth in selected traits by bringing in information from their own background as double majors. Peer evaluations of traits provided additional feedback.

> RE: ability to take teacher roles. MA: 'I like inviting audience participation through questions, but you should have a plan to lead your audience to the desired answer if nobody gets it. You can do this by restating the question, providing hints, or asking progressively simpler questions to get your audience started on the right path.' PK: 'You also have a natural technique for physically pointing out the things on the slide as you are describing it and drawing on the board when you need to make the point. I think that shows you are acting as a "teacher" to explain something to the audience.' LY: 'I could tell you made the effort to slow down and clearly describe complex concepts ... you also oriented the audience to your slides and almost always provided a clear "take-home" message for each slide'.

In faculty judgements, all tutorials contained stronger and weaker students. The rubrics discriminated between students who varied in performance. In many student assessments, faculty found clear patterns of high levels of traits across all Independent Thinker, Intellectual Maturity, and Creativity constructs in superior students and inversely so. In some cases, faculty noted that students of equal grade point averages performed unequally on the trait assessments, suggesting that the rubrics captured qualities above and beyond content mastery.

> At the onset of the project, it appeared that the rubric was motivating the weaker of my two students ... it would seem natural to ask whether the rubric provides additional motivation to a more 'average' student. After a few initial meetings, R. was able to lead the discussion and initiate topics she wanted to discuss. K. on the other hand, needs a lot of guidance, and we had to prod her to keep moving.

Self-assessment may be a trait that is deeply engrained in students prior to their entry in the tutorial and therefore might have had little potential for improvement. Unlike another trait, *curiosity* (Creativity), however, the tutorial used in this study offered continuous opportunities to engage in *self-assessment*, e.g. in writing and

revising papers and in filling out rubrics. Therefore, a tutorial employing sustainable assessment may well have contributed to the rise in student ratings for this trait.

Q3. Did the use of traits inform the design of assessments of other non-tutorial courses? If traits could only be measured in tutorials, then it would suggest that only in such contexts was there sufficient interaction to allow traits to be expressed and visible

The traits and rubrics were not found to be applicable to large non-tutorial courses, because there was insufficient interaction between instructors and students, and peers, to observe and rate student trait performances. In some cases faculty did not even attempt to observe students' traits judging that their prior experience in large courses informed them that it would be fruitless.

Summary

Based on faculty and student quantitative rubric scores and faculty case studies, we found evidence for validity and pedagogic application of the three learning outcomes and associated traits:

- Both faculty and students were able to observe and measure the traits during two cycles of tutorial courses.
- Mean levels of student self-estimated and faculty-estimated rubrics statistically improved during the tutorials for all traits of the Independent Thinker and Intellectual Maturity learning outcomes, and for all the Creativity traits except *curiosity*.
- Evaluations from faculty and students indicated that the traits were integrated into the learning goals for courses.
- Faculty found methods for encouraging student development of the traits through course plans, assignments and teaching strategies.
- In some cases, faculty noted that students of equal grade point averages performed unequally on the trait assessments, suggesting that the rubrics captured qualities above and beyond content mastery, that is, the assessments discriminated different levels of traits in stronger and weaker students.
- Trait validity findings were consistent across tutorials offered by a wide range of faculty representing disciplines in the Arts, Humanities and Sciences.
- A reliability analysis found internal consistency for each group of traits.

Tutorial settings, because of their high levels of interactive feedback, and social science assessment methods such as standardised rubrics, make positive contributions to the research of sustainable assessment theory.

Discussion, limitations and conclusions

To our knowledge this was the first study to test sustainable assessment theory. To meet the requirements of sustainable assessment theory, the outcome traits needed to refer to long-term learning abilities, embrace a shared, teacher-student methodology and help students to become assessors of their own learning. All three out-

comes, Independent Thinker, Intellectual Maturity, and Creativity, refer to long-term learning outcomes that were used as effective organizers of instruction. The criteria and standards required by sustainable assessment theory were, in fact, the definitions of the outcomes. While no formal learning contract was used, both faculty and students used these criteria in working towards these outcomes. We assume that to test sustainable assessment theory, we needed to validate the constructs underlying the outcomes (Cronbach and Meehl 1955). Validity is tested by the 'pattern of results consistent with the pattern specified by the conceptualisation of the construct' (Fiske 2002, 171). The results were consistent with our claim that the constructs were valid.

First, while a small sample size precluded a factor analysis to examine the measurement structure of the learning outcomes, we did determine that each learning outcome (construct) had a high degree of internal reliability among the associated traits. The results of the study as analysed from rubrics and case studies suggest that 8 of the 10 traits selected to comprise the three learning outcomes were valid. In addition, several faculty concluded that one student trait, *curiosity* (Creativity), was already at high levels at the start of courses and no important strategies emerged for supporting this trait. Other faculty found too much overlap between the two Creativity traits of *connectivity* and *multiple perspectives*. We propose, therefore, to reduce the final number of traits to eight: Independent Thinker (4); Intellectual Maturity (2); Creativity (2).

Second, another way to test validity was to ask whether the rubrics, which can be considered as tests, predicted course behaviours (Fiske 2002). Both the faculty and student ratings of almost all traits rose during the courses and we learned from the course transcripts and faculty case studies that both faculty and students integrated the construct outcomes into their behaviours. For example, faculty adapted their course plans and assignments as well as assessment strategies and student evaluations during discussions of the criteria, indicating that they integrated the traits into their learning goals. Some students periodically asked faculty whether they were successfully employing behaviours associated with the constructs.

Third, Messick's (1975, 1995) 'generalizability' requirement for validating constructs is concerned with the extent to which the scores of the test rubrics generalised to and across groups, settings, tasks and time periods. As we reported the results were consistent across four colleges, professors and students engaged in courses representing different disciplines, and a variety of tutorials and independent studies involving varying class sizes and time periods from 10 to 15 weeks. Several faculties proposed that using the rubrics could be generalised to the assessment of other courses, such as college-wide tutorial and seminar programmes. In some cases, faculty noted that students of equal grade point averages performed unequally on the trait assessments, suggesting that the rubrics captured qualities above and beyond content mastery.

We propose that the methodology in this study can be used to enhance sustainable assessment theory and practice. Including the study investigators, 12 researchers developed the constructs and criteria over two cycles of courses for the development of a sustainable assessment programme. We think such a collaborative and extended approach is required to formulate and test constructs. While apparently violating one of the tenets of sustainable assessment theory, we also claim that it is not feasible to engage students in the preliminary definitions of each construct. The concepts of independence and creativity, for example, have presented educators

with considerable conceptual challenges over their long history and did so during the four days of face to face meetings before the present group arrived at the specifications for these outcomes. Nevertheless, student collaboration can be obtained through the use of baseline, midpoint and exit reflective sessions during the courses. These discussions reinforced the utility of the faculty defined target outcomes. While students did not change the language of the outcome criteria, their self-reports, particularly concerning self-assessments, improved the faculty's observation of the relevant behaviours. We think the use of rubrics is indispensable not only in researching quantifiable measures of sustainable assessment theory, but also in improving students' performance on the selected outcomes. By requiring students to make multiple judgements of their target behaviours, we induced them to reflect and assess how well they were doing. We argue that these judgements contributed to their improvements of self-assessment, which is the most important aspect of the theory – for students to become assessors of their own learning. Asking for multiple rubric judgements also requires students and faculty to revisit and improve their understanding of the criteria as connected to memories of relevant behaviours over multiple test periods. This suggests that sustainable assessment may need periodic reflective discussions to consolidate examples implicated in progress and perhaps, also, to discuss criteria themselves. In the present study this did not lead to revisions in criteria, but this is a possibility.

The study showed that students' papers, discussion skills and presentation abilities are potentially valuable sources of information about the traits under investigation. In the present research context we recorded samples of oral behaviours (discussions/presentations) that might otherwise be difficult to use as verification data in ordinary, non-experimental courses. We are equivocal on whether the reflection sessions ought to be recorded and transcribed. On the one hand, this comprises an unaccustomed time and expense to instruction that may already be stressed. Nevertheless, the transcripts, which are now easily obtained at low cost, proved especially valuable to faculty in assessing the effectiveness of the programme. One professor used her cell phone to record these sessions and there are now available apps for obtaining transcriptions. The focus on student work also conjures up a body of literature on the effect of feedback that students receive from faculty in a tutorial setting. Weaver (2006) found that students valued tutor feedback, but thought that it could be much more helpful by tying it to assessment criteria and learning outcomes, as well as by being mindful of giving timely feedback to allow maximum progress to be made. 'Learning depends upon feedback to the learner' and 'the language in which feedback is delivered is an important variable' in students' concept formation (Falchikov 1995, 159). In this study faculty were specifically requested to use the language of traits in annotating papers as well as during discussions. An implication of this practice is that faculty and student e-portfolios may be a useful method for storing and retrieving evidence concerning trait ratings and for application to final evaluations whether in experimental or ordinary course programmes.

The study of independence traits in tutorials may have provided a unique context for developing theory and method in sustainable assessment. Sustainable assessment theory requires a unique commitment to *student-teacher equity*. For example, the theory assumes shared criteria in which all parties are on the same page using a common language. Tutorials also feature a role reversal in which students assume the teacher role thereby contributing to greater equity in the relation-

ship. The focus on independence traits, particularly in the inclusion of self-assessment, meant that sustainable assessment in the sense of equipping students to become life-long assessors of their learning became a specific objective of the programme. These several factors may have contributed to the success of the programme. Nevertheless, it should again be noted that students in this study had no part in developing the assessment criteria, although they voluntarily accepted the standards and appeared to work diligently with them. While all these factors could be viewed as limiting the application of SAM, it has also indicated that sustainable assessment theory should be developed by expanding the use of equity criteria in structuring assessments and making self-assessment a more intentional goal.

Another of our goals was to develop assessment methods for measuring outcomes in university education. Colleges that already use their independent study courses as culminating projects essentially allow these courses to serve as assessments for the full undergraduate education. Other institutions are now introducing 'senior experiences' as culminating projects for students. Both of these examples refer to tutorial-like programmes. SAM, therefore, would be useful for assessing students' four-year liberal arts education as it prepares them for further study, work or life. Could SAM be adapted for application to non-experimental tutorials and other courses that feature small class sizes and coordinated writing/discussion literacy intensive activities? We think SAM devices, such as rubrics, and pedagogic strategies reported could be used in such contexts. We recommend the introduction of assessment criteria in syllabi and instructors would discuss the rubrics at the onset, review progress at the midpoint, and evaluate their worth at the end of courses. Students and faculty should fill out rating scales at least three times per course. Faculty should develop pedagogic strategies that target groups of traits and use the language of traits in annotating and correcting papers. There is some evidence to the suggestion that sustainable assessment theory complements both formative and summative assessment. We found that data produced by formative and summative assessments were useful in elaborating sustainable assessment theory.

These are still early days in the state of the art for assessment researchers of higher education outcomes, but we should be working towards a common classification system of outcomes and criteria, while recognising the intrinsic problems. The development of rating scales and rubrics for the assessment of undergraduate outcomes needs to proceed in conjunction with the collection of behavioural criteria for each trait. Thus, it would not be sufficient for an instructor to rate a student's trait without documentation of the observed behaviours that served to inform the rating. Further, given the emphasis on metacognitive traits it seems sensible to ask students to rate their own possession of the quality, to document what they think constitutes demonstration of the trait, and to what degree they estimate they have developed the trait during the selected period of instruction. We question to what extent it is productive for students to be co-constructors of long-term learning outcomes and rubrics, but their participation in sustainable assessment is essential, as we have argued throughout the paper. A formal assessment process should encourage self-assessment, according to Boud, Cohen, and Sampson (1999), as assessment in higher education 'has a dual function of judging for the purpose of providing credentials and for the purpose of improving learning' (419). Above all, researchers and instructors alike should provide end of course opportunities for discussing with students the degree to which they have become assessors of their own learning.

Acknowledgements

Notes on contributors

Robert J. Beck, PhD, is visiting professor of Education at Lawrence University and professor emeritus of Education at University of California, Irvine. His research interests include educational discourse processes in higher education, teacher education, and educational technology.

William F. Skinner, PhD, is director of Research Administration at Lawrence University. His research interests include assessment and retention in higher education.

Lynsey A. Schwabrow, PhD, is associate director of Research Administration at Lawrence University. His research interests include retention issues in higher education and environments for children.

References

Arum, R., and J. Roksa. 2010. *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago.

Association of American Colleges and Universities, and Council for Higher Education Accreditation. 2008. *New leadership for student learning and accountability: A statement of principles, commitments to action*. Washington, DC: AAC&U and CHEA.

Beck, R.J. 2007. *The Oxford tutorial: A learning theory*. Oxford: Oxford Magazine.

Black, P., and D. Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability* 21: 5–31.

Bloom, B.S. 1984. The 2 Sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher* 13: 4–16.

Bok, D. 2006. *Our underachieving colleges*. Princeton, NJ: Princeton University Press.

Boud, D. 2000. Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education* 22, no. 2: 151–67.

Boud, D., R. Cohen, and J. Sampson. 1999. Peer learning and assessment. *Assessment & Evaluation in Higher Education* 24, no. 4: 413–26.

Boud, D., and N.F. Falchikov. 2006. Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education* 31, no. 4: 399–413.

Boyd, H.R., and J. Cowan. 1986. A case for self-assessment based on recent studies of student learning. *Assessment & Evaluation in Higher Education* 10, no. 3: 225–35.

Brown, A.L. 1987. Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In *Metacognition, motivation, and understanding*, ed. F.E. Weinert and R.H. Kluwe, 65–116. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, P.A., J.A. Kulik, and C.C. Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal* 19: 237–48.

Connor, W.R. 2006. 'From Foxes to Hedgehogs'. Center of Inquiry in the Liberal Arts at Wabash College. Podcast address given in Washington, DC, January 27, 2006.

Costa, A.L., and B. Kallick. 2000. *Habits of mind: A developmental series*. Alexandria, VA: Association for Supervision and Curriculum.

Cronbach, L.J., and P.E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302.

Falchikov, N. 1995. Improving feedback to and from students. In *Assessment for learning in higher education*, ed. P. Knight. London: RoutledgeFalmer.

Fiske, D.W. 2002. Validity for what? In *The role of constructs in psychological and educational measurement*, ed. H.I. Braun, D.N. Jackson, and D.E. Wiley, 169–78. Mahwah, NJ: Lawrence Erlbaum.

Flavell, J.H. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34: 906–11.

Hacker, D.J., and A.C. Graesser. 2007. The role of dialogue in reciprocal teaching and naturalistic tutoring. In *Talking texts: How speech and writing interact in school learning*, ed. R. Horowitz, 253–70. Mahwah, NJ: Lawrence Erlbaum.

Katz, S.N. 2008. Taking the true measure of liberal education. *Chronicle of Higher Education* 54, no. 37: 32.

Marzano, R. 1992. *A different kind of classroom*. Alexandria, VA: Association for Supervision and Curriculum.

Mash, R. 2001. Tutorial teaching in economics. In *The Oxford tutorial*, ed. D. Palfreyman, 87–92. Oxford: Blackwell's.

Mayr-Harting, H. 2006. *The Oxford tutorial. Convocation curriculum delivered. November, 2006*. Appleton, WI: Lawrence University.

Messick, S. 1975. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist* 30: 955–66.

Messick, S. 1995. Validation of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meanings. *American Psychologist* 50: 741–9.

Mirfield, P. 2001. Teaching law, learning law: Growing up intellectually. In *The Oxford tutorial*, ed. D. Palfreyman, 37–41. Oxford: Blackwell's.

Moore, W.G. 1968. *The tutorial system and its future*. Oxford: Pergamon.

Smith, E. 2001. English: A shared enterprise. In *The Oxford tutorial*, ed. D. Palfreyman, 105–9. Oxford: Blackwell's.

Spellings, M. 2006. *A test of leadership: Charting the function of US higher education*. Washington, DC: US Department of Education.

The Five Colleges of Ohio Creative and Critical Thinking Project. 2009. Creative and critical thinking: Assessing the foundations of a liberal arts education. http://www3.wooster.edu/teagle/docs/Y3 Teagle Final Report.pdf.

van Eemeren, F.H., R. Grootendorst, S. Jackson, and S. Jacobs. 1997. Argumentation. In *Discourse as structure and process*, ed. T.A. van Dijk, 208–29. London: Sage.

Weaver, M.R. 2006. Do students value feedback? Student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education* 31, no. 3: 379–94.