Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'

Calvin D. Smith, Kate A. Worsfold, Lynda Davies,
Ron Fisher and Ruth McPhail

Griffith University, Australia

Corresponding Author:

Calvin Smith, Griffith Institute for Higher Education, Griffith University, Brisbane, Australia.

email: calvin.smith@griffith.edu.au

Abstract

In this paper we report on a study to quantify the impact on student learning and on student assessment literacy of a brief assessment literacy intervention. We first define 'assessment literacy' then report on the development and validation of an assessment literacy measurement instrument. Using a pseudo-experimental design we quantified the impact of an assessment literacy-building intervention on students' assessment literacy levels and on their subsequent performance on an assessment task. The intervention involved students in the experimental condition analysing, discussing and applying an assessment rubric to actual examples of student work that exemplified extremes of standards of performance on the task (e.g. poor, excellent). Results showed that such a procedure could be expected to impact positively on assessment literacy levels and on student performance (on a similar or related task). Regression analyses indicated that the greatest predictor of enhanced student marks (on the assessment task that was the subject of the experiment), was the development of their ability to judge standards of performance on student work created in response to a similar task. The intervention took just 50 minutes indicating a good educational return on the pedagogical investment.

Keywords: assessment; assessment literacy; student learning; first-year experience; first-year students.

Assessment literacy and its role in student learning

Introduction

A key issue for academics, especially those who teach first-year students, is how to enable students to feel part of their programs' academic culture while encouraging them to take responsibility for their own learning. (Nicol, 2009). To become self-regulated learners, students need to be able to judge their work, identify its merits, locate its weaknesses and determine ways to improve it (Nicol, 2009; Sadler, 2009). Part of that judgement includes evaluating the appropriateness of their responses to assessment tasks and whether they have done what they were asked to do (Sadler, in press). It also requires them to judge how good their response is in relation to the relevant academic achievement standards (Sadler, 2009). Students' understanding of the purposes of assessment and the processes surrounding assessment is part of the context within which they learn to make those judgements and become effectively self-regulating.

Francis (2008) argues, however, that first-year students in particular are likely to over-rate their understanding of the assessment process and that there is a disjuncture between what they think they are being assessed on and what the marking criteria and achievement standards require of them. Using formative assessment, Nicol (2009) addressed this issue as part of the REAP Project supported by the Scottish Funding Council. He tested Yorke's premise that "in order to be successful, students in the first year must have a clear understanding of what is required by academic study" (2009:336). By helping to clarify the meaning of learning goals and criteria, and through the provision of feedback, formative assessment encourages students to keep realigning their work to what is required. Nicol consequently applied a framework based on task structure, learner-regulation and an associated set of assessment principles to inform the redesign of formative assessment in two first-year courses. The framework uses some of Gibbs and Simpson's (2004) eleven assessment *conditions* (relating to the assessment tasks), and Nicol and Macfarlane-Dick's (2006) seven *principles* of good feedback practice. This framework, and work by others (Boud, 2000; Braxton, 2000; Rust,

O'Donovan, and Price, 2005; Yorke, 2005a, 2005b))) suggests that, for students to have a sense of control over their own learning, formative assessment practices must help them develop the skills needed to monitor, judge and manage their learning (Nicol, 2009:338).

One way the sector has moved to help students understand the expectations of assessment has been the development of criteria and standards descriptors that ostensibly set out what elements students' work will be judged against, in responding to a given assessment task. O'Donovan, Price and Rust (2004) have researched the outcomes of criterion-referenced assessment and argue that the current approach in Higher Education of single-mindedly relying on the explicit expression of assessment standards and criteria cannot, on its own, adequately help students to understand assessors' perceptions and expectations of assessment.

Following the outcomes of their five-year research project, and in the context of the quality assurance environment in the UK, these authors have detected that the sector is slowly acknowledging the difficulty of trying to define threshold standards with language that is meaningful to all stakeholders. The authors have been developing a growing body of work on criterion-based assessment methods including the development and use of assessment rubrics (grids), grade descriptors, and benchmark statements. While part of their work identifies the problems involving the effective construction of the rubrics and descriptors, how those rubrics and descriptors are received and used by students also raises concern. They therefore exhort colleagues to be consistent in their choice of language right through the assessment enterprise so students can more accurately grasp the meaning of particular terms, tasks, and expectations expressed by teachers.

The literature here tends to suggest that students' capacity to become successful self-regulated learners can be affected by various aspects of the assessment process. We argue that first, students need to *understand the purpose* of assessment and how it connects with their learning trajectory. Second, they need to be *aware of the processes* of assessment and how they might affect students' capacity to submit responses that are on-task, on-time and completed with appropriate

academic integrity. Third, opportunities for them to practice *judging* their own responses to assessment tasks need to be provided so students can learn to identify what is good about their work and what could be improved. We therefore conceptualised students' capacity to develop these aspects of assessment as *assessment literacy*, and defined this as students' understanding of the rules surrounding assessment in their course context, their use of assessment tasks to monitor, or further, their learning, and their ability to work with the guidelines on standards in their context to produce work of a predictable standard.

While there exists in the literature a small number of questionnaires designed to focus on students' assessment *experience*, there is no research that has specifically looked at the concept of student *assessment literacy*. However, reviewing those questionnaire-based studies (Gibbs & Dunbar-Goddet, 2009) and other research on students' assessment experience (Biggs, 1987, 2003; Francis, 2008; Nicol, 2009) provided a basis for identifying the gaps in the measurement literature relating to assessment and to students' assessment literacy.

Similar to the notion of assessment literacy, O'Donovan, Price and Rust (2004; Rust, Price, & O'Donovan, 2003) conducted a study on facilitating students' understanding about assessment. They argued that knowledge has both explicit and tacit dimensions, and that learners need to construct that knowledge from experience for themselves for it to have meaning for them. This is true for both discipline knowledge and "meaningful knowledge [of] assessment requirements and criteria" (2004:331). Their approach was to aim to develop, through structured activities, students' knowledge of how assessment responses would be marked, and in turn their understanding of how their own responses would be judged. O'Donovan and colleagues point out that it is the shared experience of marking and moderation that helps teachers to build their tacit knowledge base and students are not often offered such opportunities. Their intervention, conducted with students in their first term of university, therefore mirrored aspects of the marking process by requiring students to 'mark' assessment responses and discuss their judgements.

The intervention was based on the notion that that once the students started making judgements about the quality of the work in front of them they could apply that evaluative way of thinking to their own work to help them self-monitor it during its production and identify ways to improve its quality. Students were invited to a 90 minute marking workshop. Prior to the workshop they were given two exemplar pieces of work that they had to mark and provide feedback for. The assignments were similar in nature and format to the next piece of assessment that the participating students were about to commence for their own coursework, but covered different topics with different instructions. During the workshop students discussed their marking and rationales in small groups before reporting to the whole class the marks they awarded and their justifications. At this point the lecturer lead a discussion of the students' rationales and related them to the application of the marking criteria. The small student groups then had a chance to reconsider their marks and rationale, and finally, the lecturer finally provided the whole class his/her annotated assignment exemplars showing the feedback, mark, and rationale.

This activity was replicated over a three-year period and the authors found that students who participated in the workshop showed significant improvement in subsequent assessment pieces compared with students who did not participate. Outcomes of their research show that relying only on the explicit expression of assessment criteria, standards, and processes as a method of transferring knowledge about assessment does not work (see also Sadler, 2009). O'Donovan, Price and Rust's (2004; Rust et al., 2003) results suggest that the provision of explicit criteria and summarised standards descriptors needs to be complemented by opportunities for students and staff to share the experience of judging the quality of responses in order to build tacit knowledge into the students' repertoire, improve their assessment literacy and hence their assessment outcomes.

In the present study we aimed to test this assertion by quantifying the impact of developing students' assessment literacy on their assessment literacy levels and on their learning outcomes.

Further, we set a stringent high-risk testing scenario in which the assessment literacy-developing intervention was much briefer (50 minutes) than those used in the O'Donovan, Price and Rust (2004; Rust et al., 2003) studies. To do this we first developed a questionnaire to operationalise some key concepts in the assessment literacy arena; we then implemented these measures in a pre- and post-test framework, such that, between the measurement episodes, the students in the experimental cohort were exposed to an assessment literacy-building intervention. A control group in the same program of study, but at a different campus location, received only the pre-test instrument and no intervention. This paper reports on the results of this intervention.

Method

Participants

A convenience sample of 369 undergraduate students was obtained from a public university in Queensland, Australia across two different campus locations (Campus A and Campus B). Participants comprised first year business students who took part voluntarily in the study (56% females, 44% males, with a mean age of 19.1 years). The study was quasi-experimental in design, which means that, in order to assess the impact of an intervention, two groups were included; one which received the intervention (Campus A: intervention group), one which did not (Campus B: control group). It is 'quasi'-experimental because allocation of students to the two groups is not random; rather it is a matter of convenience to the students which campus group they attend. Campus A (intervention group) students completed both a pre-test and post-test assessment literacy survey while Campus B (no intervention group) students completed a pre-test equivalent survey only. Of the surveys collected, 20 cases were deleted due to missing student numbers or a failure to complete both pre and post-test surveys (Campus A), leaving a total of 349 useable cases for further analysis.

Because there was a chance that the intervention would benefit those in the intervention or experimental group, relative to the students who did not receive the intervention, it was agreed that were there any discrepancy in the two groups' final grades, in favour of the experimental group, then

a normalising factor would be applied to the results of the students in the experimental group to bring the two groups' final grades into line with each other. Approval from the University's Human Research Ethics Committee was sought prior to commencement of the study.

Materials

Assessment Literacy. This was measured using the Assessment Literacy Survey developed by the authors, which included 30 items designed to test a range of related constructs, including:

- Students' understanding of the local protocols and performance standards (6 items) (e.g., "I understand the criteria against which my work will be assessed");
- Students' use of assessment tasks for enhancing or monitoring their learning, including *assessment for learning* (6 items) (e.g., "I use assessment to figure out what is important to learn") and *assessment for grading* (4 items) (e.g., "I think the University makes me do assessment to: produce work that can be judged for the University's marking and grading purposes").
- Students' orientation to putting into the production of assessable work the minimum amount of effort necessary merely to pass the course requirements (6 items) (e.g., "My aim is to pass the course with as little work as possible");
- Students' ability to judge their own and others' responses to assessment task (8 items) (e.g., "I feel confident that I could judge my peer's work accurately using my knowledge of the criteria and achievement standards provided").

Responses to all items were rated on a five-point Likert scale, ranging from 1 = "Strongly Disagree" to 5 = "Strongly Agree".

Assessment performance. As part of their general course assessment, participants were evaluated across a number of different tasks: quizzes, a report (related to the intervention), and a final exam. Prior to the intervention, participants completed two multiple-choice quizzes (with a combined potential mark ranging from two to 10) as part of their course assessment. This covariate

of “pre-intervention quiz” was used in subsequent analyses as a proxy control for participants’ pre-intervention academic ability.

The dependent variable of primary interest was student achievement (grades) on the intervention-related assessment task, which took the form of a report (scored from one to 30). Instructions for the report were given via a face-to-face transmission of information (relating to timing, due date, lodgement procedures and assessment standards). A rubric consisting of a matrix of criteria gridded against standards provided descriptions of performance against each criterion for each performance standard.

The final exam was designed to test participant’s knowledge of material covered in lectures, tutorial classes and within the course textbook. The final exam contained multiple-choice, written and short-answer case-scenario questions.

Procedure

The study was quasi-experimental in design with Campus A receiving an intervention aimed at improving assessment literacy levels, while Campus B received no intervention. The procedure, described below, was completed as part of a one hour lecture in week six of the semester.

Assessment rubric phase*.* Participants at *both* locations were presented with the assessment rubric. This phase comprised several steps as follows:

(a) Participants were told how they could, and should, use the rubric when completing the task;

(b) Criteria were explained, including what each “component” might look like in the task;

(c) Participants were informed that assessor’s judgments would be based on how well their responses addressed the criteria; and

(d) That those judgments would be aligned with the academic standards described briefly in the rubric.

Pre-test assessment literacy survey phase. At *both* locations, the Assessment Literacy Survey was then administered. As such, the survey functioned as a pre-test measure for Campus A participants (intervention group), or a "pre-test equivalent" for Campus B (control group). No post-test survey was administered in the control group because the amount of time spent in the 'usual' approach to informing students about the assessment rubric was minimal. The pre-test measure allowed comparison between the two campuses to detect any initial group differences in assessment literacy and establish baseline levels. Having the pre-test data in both cohorts also gave a greater pool of cases for the initial factor analysis. Completion of the questionnaire took approximately five minutes.

Intervention phase (Campus A only). The intervention phase lasted approximately 45 minutes and comprised:

(a) Facilitation of a "think, pair, and share" exercise where participants considered, judged and practiced marking two exemplars (actual examples of student work) – to determine their quality: "excellent", "good", "satisfactory", or "bad". The aims of the activity focused on helping participants learn how to: judge a piece of work; identify the criteria they use to make that judgment; and use criteria and recognize different academic achievement standards.

I. Participants made practice judgments on the exemplars ("think"), then explained and justified their judgments to the person next to them ("pair and share").

II. Randomly selected pairs shared their decisions with the whole class.

III. Out of this conversation emerged a list of criteria expressed by the participants in their own language.

(b) Differentiation of exemplars then took place. The Course Convenor asked (via a show of hands), which exemplar was identified as the weaker response, and which one the better response. He then asked the participants to indicate the mark they gave each of the exemplars (a number out of 30). This was done through show of hands indicating what range of marks their judgement fell within. The Convenor then divulged his marking and the reasons for it. Most of the class were within a mark or two of the Convenor.

(c) Participants then referred to the assessment rubric and compared the basis for their judgements against the academic standards expressed in it.

Post-test assessment literacy survey phase. Following the intervention, the Assessment Literacy Survey was administered to Campus A (intervention group) participants again to assess changes in assessment literacy levels.

Assessment outcome phase. Three weeks after the initial "rubric" lecture, (i.e. in teaching week nine), participants at both locations completed a literature review and 1500 word report which formed part of the course-work assessment, using the same rubric previously introduced within the week six lecture. Subject tutors across the two locations were responsible for grading the reports using criterion-referencing. The participant's report mark (one to 30) is used as a dependent or outcome variable in later statistical analyses.

Results

Preliminary Analyses

Following the initial scale development phase, a number decision making or review steps were undertaken to finalise scale development, including data screening, corrected item-total correlation analysis, exploratory factor analysis (using Principle Axis Factoring, or PAF) and reliability analysis. Following these steps and the determination of the final scale, scale scores were calculated for each factor in preparation for later statistical analysis.

Normality was examined via inspection of: (a) Kolmogorov-Smirnov tests of normality, (b) plots (including histograms and boxplots), and (c) skewness and kurtosis statistics. Across both test times, some variables were not normally distributed, with extreme scores detected across three pre-test variables and ten post-test variables. Factor analysis is however robust against departures from normality (Allen & Bennett, 2008) and may still be conducted (Tabachnick & Fidell, 2001). After removal of cases with missing data via listwise deletion, the final pre-test sample size included 317 (32 cases removed) students, and the post-test sample size was 158 (12 cases removed) students. As a general rule of thumb for factor analysis, sample sizes of 300 and over are preferable (Tabachnick & Fidell, 2001), or alternatively samples should have at least a ratio of five participants for each item (Allen & Bennett, 2008). Both of these conditions were satisfied in the present study.

An item total correlation refers to the correlation between the item score and the overall scale score to which that item belongs. Items with low item-total correlations (< .5) were identified and noted for possible later deletion.

Factor analysis was undertaken to confirm the underlying structure of the survey items. As an underlying factor structure was predicted prior to analysis, a PAF method of extraction was chosen to examine both pre- and post-test responses. As the factors were expected to be correlated, an oblique rotation was selected. A variety of factor analysis criteria were utilised, including: (a) Kaiser-Meyer-Olkin Measure of Sampling Adequacy (requirement of $\geq$ .6), (b) Bartlett's Test of Sphericity (requirement of $p < .05$), (c) item correlations to at least one other item by .3, (d) anti-image correlation matrix diagonals to be at least .5 or higher, (e) communalities to be at least .3 or higher, and (f) factor loadings to be at least .3, with any cross-loadings to be less than .3. If particular items were found to have poor factor loadings, communalities, correlations or anti-diagonals, they were noted for possible exclusion.

In computing the first PAF using the pre-test data, all items were included. Nine factors (based on eigenvalues greater than one) were identified, accounting for 63.15% of variance, but this

solution was theoretically incoherent. A priori, a five-factor structure made theoretical sense, but after inspection of the scree plot which was indicative of either a two- or four-factor solution, PAF was again computed with two-, four- and five-factor solutions examined. The two factor solution only explained 31% of the variance and was not theoretically coherent so this model was not interpreted. Of the four- and five-factor solutions, the four-factor solution was considered more theoretically and statistically viable. Following removal of particular items based on the factor analysis criteria, the final four-factor solution (17 Items) accounted for 58% of the variance: factor one (30%), factor two (10%), factor three (9%), and factor four (8%). All factors had eigenvalues greater than one. All other factor analysis criteria were upheld, including: KMO sampling adequacy was acceptable (.83), Bartlett's test of sphericity was significant (χ^2 (136) = 1850.11 ($p < .05$), anti-image diagonals were all .65 or greater, items were correlated to at least one other item by .3, with communalities ranging from .30 to .76.

To confirm this pre-test factor structure, an additional PAF was conducted using the post-test data. The same structure was obtained with all factor loadings greater than .5. The post-test four-factor solution accounted for 70% of the variance. The factors accounted for the following amounts of variance: factor one (41%), factor two (11%), factor three (9%), and factor four (8%). All factors had eigenvalues greater than one. All other factor analysis criteria were upheld, including: KMO sampling adequacy was acceptable (.86), Bartlett's test of sphericity was significant (χ^2 (136) = 1555.62 ($p < .05$), anti-image diagonals were all .64 or greater, items were correlated to at least one other item by .3, with communalities ranging from .32 to .78. As illustrated in Table 1, internal reliabilities for both the pre- and post-test PAF were acceptable ($\alpha > .65$). The four factors represented the following constructs, respectively: *Assessment Literacy (Understanding)* (AU); *Assessment for Learning* (AL); *Minimum Effort Orientation* (MEO); and *Assessment Literacy (Judgment)* (AJ). The factor loading matrix for this final solution is presented in Table 1 (pre-test and post-test responses). As evidenced by the following table, the four-factor model was found to be reliable across both the pre and post-tests.

TABLE 1: ***PRINCIPLE AXIS FACTORING ANALYSIS WITH OBLIQUE ROTATION. FACTOR STRUCTURE AND ITEM LOADINGS ACROSS PRE- AND POST-TEST MEASURES (N = 158)***

Items (abbreviated)	Factor 1 (AU)		Factor 2 (AL)		Factor 3 (MEO)		Factor 4 (AJ)	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	α.82	α.90	α.73	α.87	α.69	α.75	α.77	α.85
The Department's assessment procedures are clear to me	.74	.83						
I understand the rules applying to assessment	.67	.81						
I understand the *criteria*	.66	.78						
I understand the *achievement standards*	.63	.80						
What I need to do to get the mark or grade I want	.51	.63						
Advance my learning to achieve the standard I want	.48	.66						
Show me how much of the course content I understand			-.73	.87				
To work out what are the expected achievement standards			-.63	.82				
I use assessment to work out how well I am doing			-.53	.71				
I use assessment to figure out what is important to learn			-.54	.57				
I learn more when I do the assessment tasks			-.51	.61				
My aim is to pass the course with as little work as possible					.73	.87		
Assessment to work out the minimum work needed to pass					.73	.80		
I do assessment because I have to					.44	.45		
Judge my own work using my knowledge of the criteria							-.83	.72
Judge my peer's work using my knowledge of the criteria							-.79	.88
Use the criteria provided to improve my work							-.31	.51

Note. Pre-test measure (both campus locations), (N = 317); post-test measure (Campus A only), (N = 158); Item loadings < .3 have been suppressed; AU=Assessment literacy (understanding); AL=Assessment for learning; MEO=Minimum effort orientation; AJ=Assessment literacy (judgement)

As anticipated the MEO scale scores were negatively correlated with AL (r=-.26, p<001), AU (r=-.30, p<001), and AJ (r=-.27, p<001). The assessment for learning scale (AL) was positively correlated with AU (r=.42, p<001) and AJ (r=.33, p<001). The other two assessment literacy scales, understanding (AU) and judgement (AJ) were also positively correlated (r=.50, p<001). This is evidence of appropriate and predicted patterns of convergent and discriminant validity (Campbell & Fiske, 1959).

Scale scores were then computed for the pre-test (both campuses) and post-test data (Campus A only) for the four factors (the time-dependent measures are indicated with subscripts e.g., MEO_{T1} and MEO_{T2} refer to the pre-test and post-test measures of the MEO scale respectively).

Further data screening was undertaken to detect the presence of any univariate or multivariate outliers. One case with a high Mahalanobis' Distance statistic was removed in addition to five cases subject to a "self-inflation" effect. Such participants appeared to have an inflated sense of their understanding of assessment and their judgment ability in comparison to their generally low academic performance (i.e., these participants rated themselves highly in terms of AU_{T2} or AJ_{T2}, but had consistently poor academic performance as measured through other assessment tasks). Independent t-tests and effect sizes revealed that these students had significant or largely poorer general performance as measured via: the pre-intervention quiz $t(320)=1.64, p=.1, d=.75$, the report $t(315) = 9.45, p = .00, d = 2.2$, and the final exam $t(315) = 7.87, p = .00, d = 1.87$. Furthermore, analysis of response patterns demonstrated an acquiescent response bias in these cases therefore they were removed from further analysis.

Assessment of group differences (pre-test). To determine the presence of any significant difference between Campus A (intervention) (n = 162) and Campus B (control) (n = 177) participants on the pre-test measures, an independent samples *t* test (two-tailed) was conducted. There were no significant differences between the groups on minimum effort orientation, assessment literacy (judgement) and assessment literacy (understanding) at pre-test [MEO_{T1} $t(337) = .67, p =.51, d = .08$, AJ_{T1} $t(335) = -1.11, p =.27, d = .12$, and AU_{T1} $t(331) = -1.31, p =.19, d = .15$]. While differences were not significant, Campus A (intervention) scored more poorly than Campus B (control) students in terms of *effort* (higher MEO_{T1}, *Ms* =

2.88, and 2.81 respectively), *judgement* (AJ_{T1}, *Ms* = 3.24, and 3.33 respectively), and *understanding* (AU_{T1}, *Ms* = 3.61, and 3.70 respectively).

Campus A participants scored significantly lower *assessment-for-learning* scores (AL_{T1} *M* = 3.71) than Campus B participants (*M* = 3.92, *t*(334) = -3.02,*p* = .003, *d* = .33). These results point towards a general trend of slightly poorer baseline motivation and assessment literacy levels for Campus A (intervention) students, especially in terms of using assessment for learning purposes. Consideration must be paid to these baseline differences in any later comparison of average report marks between campus groups, in addition to determining the intervention's impact on assessment literacy levels and any associated improvement in report mark for Campus A. All means, standard deviations and effect sizes are shown in Table 2.

Table 2

Mean Scores by Campus Location

Factor	Campus A (intervention)		Campus B (control)		Mean Difference	Effect Size
	M	*SD*	*M*	*SD*		*d*
AL_{T1}	3.71	.68	3.92	.61	.21*	.33 (SM)
MEO_{T1}	2.88	.89	2.81	.90	-.07	.08 (S)
AJ_{T1}	3.24	.75	3.33	.73	.09	.12 (S)
AU_{T1}	3.61	.59	3.70	.61	.09	.15 (S)

* Significant difference p < .05

Note. S = small effect size, SM = small to medium effect size as per Cohen's (1988) conventions.

Intervention impact upon assessment literacy. Paired sample *t* tests were conducted to investigate the impact of the intervention on participants' effort, use of assessment and assessment literacy levels. The intervention was effective in producing positive and significant change across the three assessment literacy factors: AU *t*(153) = -10.21,*p* =.00, *d* = .73, AJ *t*(157) = -6.51,*p* =.00, *d* = .52, and AL *t*(155) = -6.03,*p* =.00, *d* = .39. These effect sizes reveal the intervention resulted in medium to large changes in the three assessment literacy levels (see Table 3). Considering the brevity of the intervention, the magnitude of this impact may therefore be considered a hefty return on a small investment. No significant change occurred

in the attitudinal measure of MEO $t(157) = 1.67, p = .10$, $d = .08$. All means, standard deviations and effect sizes are shown in Table 3 as follows.

Table 3

Change in Assessment Literacy and MEO scores across time (Campus A - Intervention group)

Factor	Time 1		Time 2		Difference	Effect Size
	M	*SD*	*M*	*SD*		*d*
AU	3.62	.59	4.08	.56	.45*	.73 (ML)
AJ	3.26	.75	3.65	.77	.39*	.52 (M)
AL	3.70	.69	3.97	.69	.27*	.39 (SM)
MEO	2.86	.89	2.78	1.00	-.08	.08 (S)

* Significant difference $p < .05$

Note. S = small effect size, SM = small to medium effect size, ML = medium to large effect size as per Cohen's (1988) conventions.

Impact of assessment literacy on assessment results. Overall, an independent *t* test revealed there was no significant difference in average report marks between the intervention group (Campus A, $M = 18.08$) and the non-intervention group (Campus B, $M = 18.2$), $t(157) = -6.51, p = .00$ (two-tailed), $d = .02$. However, given the identified differences in baseline levels of motivation and assessment literacy between the campus groups, further analysis was required to fully determine the impact of the intervention on learning outcomes. Consequently, correlational analysis and a regression model were developed to examine Campus A (intervention) students' report mark as a function of changes in their motivation or assessment literacy levels due to the intervention. As such, change scores were calculated for the four variables, indicated by the use of subscripted "Ch" (i.e., MEO_{Ch}, AL_{Ch}, AU_{Ch} and AJ_{Ch}).

For Campus A, *improvements* in each assessment literacy factor (AL_{Ch}, AU_{Ch} and AJ_{Ch}) were *significantly and positively related* to report mark. While the strength of these relationships were small to medium at best ($r < .3$), they were nonetheless in the expected direction, such that improvements in assessment literacy levels following the intervention were related to higher report marks. The relationship of most importance was between AJ_{Ch} and report mark ($r = .27$), where increases in AJ_{Ch} were related to

better report marks. As per Cohen's (1988) conventions, the magnitude of this effect is considered medium.

No significant relationship was found between changes in MEO and report mark. This is most likely due to the intervention not leading to any significant change in MEO levels (refer to Table 3). Table 4 describes outcomes from the correlational analysis.

Table 4

Descriptive Statistics for Attitudinal, Assessment Literacy and Learning Outcome Measures

Variable	*M* or Mean difference ($T_2 - T_1$)	*SD*	*N*	Correlation with Report Mark (Pearson's *r*)
AJ_{Ch}	.39	.76	158	.27*
AU_{Ch}	.45	.55	154	.16*
AL_{Ch}	.27	.56	156	.15*
MEO_{Ch}	-.08	.59	158	.11
Report Mark	18.1	5.05	155	-

* Significant difference p < .05 (one-tail)

As the change in MEO for Campus A was not significantly correlated with report mark, this variable was not entered into the multiple regression model which comprised the change in scores for: AL_{Ch}, AU_{Ch} and AJ_{Ch}, along with report mark.

Combined, changes in AL, AU and AJ accounted for 9.5% (R^2) of variance in report mark, which may be considered a small to medium effect by Cohen's (1988) conventions, $F(3,135) = 4.72, p = .00$. Of the three assessment literacy factors only the change in AJ made a significant unique contribution to report mark, contributing 5.3% of the variance, $t(135) = 2.78, p = .01$. Thus of the three assessment literacy factors, improving students' ability to judge the standards of their own and others' work appears to be the most critical to enhanced learning outcomes. Unstandardised (*B)*, and standardised (β) regression coefficients, and squared part correlations (sr^2) for each predictor in the model are shown in Table 5.

Table 5

Unstandardised (B) and Standardised (β) Regression Coefficients, and Squared Part Correlations (sr^2) for Each Predictor in the Model Predicting Report Mark

Variable	*B*	β	sr^2 (%)
AJ_{Ch}	1.71*	.24	5.20%
AL_{Ch}	1.17	.12	1.28%
AU_{Ch}	.58	.06	.28%

* Significant difference $p < .05$

Unique contribution of improved assessment literacy. To determine whether the improved assessment literacy levels predicted report marks over and above participants' post-test motivation (MEO_{T2}) and pre-existing academic ability (pre-intervention quiz), a hierarchical multiple regression analysis was conducted for the intervention group.

At step one, MEO_{T2} and the proxy for academic ability (pre-intervention quiz mark) were entered. The three assessment literacy change factors (AU_{Ch}, AJ_{Ch}, and AL_{Ch}) were entered at step two. Overall this regression analysis accounted for 16.8% (R^2) of the variance in report mark, which may be considered a medium effect by Cohen's (1988) conventions, $F(5,131) = 5.28, p = .00$. At step one, pre-intervention quiz and MEO (post-test) accounted for 7.8% (R^2) of the variance in report mark which was significant, $F(2,134) = 5.68$, $p = .00$. At step two, the inclusion of the three assessment literacy change factors (AU_{Ch}, AJ_{Ch}, AL_{Ch}) resulted in an additional 9.0% of the variance in report mark being explained which was significant, $\Delta F(3,131) = 4.71, p = .00$. Of all the variables, only participants' pre-existing academic ability (pre-quiz mark) and change in AJ (AJ_{Ch}) were significant individual predictors. Standardised regression coefficients show that changing students' judgment ability (β = .21) is slightly more important than their pre-existing academic ability (β = .20) in explaining their report marks, though in the same order of magnitude. This finding is practically significant given the brevity of the intervention; it seems that from an intervention of just 50 minutes duration, designed to develop students' judgement abilities, their marks on a related task can be significantly increased. A summary of the hierarchical regression for predicting report mark is shown in Table 6.

Table 6

Summary of Hierarchical Regression Model Predicting Report Mark

	R^2	R^2Ch (%)	*F*Ch	*df*	*B*	β	sr^2 (%)
Step 1	.08	7.80%	5.68	2, 134			
Pre-intervention Quiz					.77*	.21	4.14%
MEO_{T2}					-.69	-.13	1.43%
Step 2	.17	9.00%	4.71	3, 131			
AJ_{Ch}					1.49*	.21	3.94%
Pre-intervention Quiz					.72*	.20	3.59%
MEO_{T2}					-.79	-.15	1.87%
AL_{Ch}					1.36	.13	1.69%
AU_{Ch}					.75	.07	.47%

* Significant difference $p < .05$

Between group differences in report mark as a function of assessment literacy. As discussed, no overall significant difference was evidenced in average report marks between the intervention and non-intervention groups. However, to further determine the effect of assessment literacy upon actual assessment (i.e., report mark), bivariate correlations were conducted between assessment literacy factors and report mark for the pre- and post-test data.. For both groups' pre-test data, the three assessment literacy factors (AU_{T1}, AJ_{T1}, AL_{T1}) did not significantly correlate with report mark.

Given the improvement in the three assessment literacy factors for Campus A (intervention) students following the intervention, bivariate correlations between post-test assessment literacy measures and report mark were conducted. As hypothesised, these improved assessment literacy factors of AU_{T2}, AJ_{T2}, AL_{T2} as measured at post-test, were more strongly (and significantly) related to report mark, such that higher assessment literacy levels were related to higher report marks. While the intervention did not significantly alter Campus A levels of MEO, this attitudinal factor as measured at both time one and time two, did significantly relate to report mark, with a higher propensity towards using minimal effort related to lower report marks. Table 7 presents these relationships.

Table 7

Correlations Between Pre- and Post-Test Measures and Report Mark

Assessment Literacy Survey Variables	Report Mark	
	Time 1 (Pre-test) Campus A	Time 2 (Post-test)
MEO	-.27*	-.16*
AL	.02	.15*
AU	.12	.26*
AJ	.01	.28*
	Campus B	
MEO	-.02	-
AL	.00	-
AU	-.09	-
AJ	-.12	-

* Significant difference $p < .05$ (one-tail)

As report marks were found to be higher for Campus A (intervention) students with higher assessment literacy levels, to determine if any report mark differences existed between the Campus B non-intervention group and the Campus A participants with low, medium and high assessment literacy levels, four one-way analysis of variance (ANOVA) tests were conducted. In order to conduct the one-way ANOVAs, the post-test scale scores for Campus A students were recoded into three groups (low, medium and high) for each of the assessment literacy and attitudinal factors, with low and high groups defined as one standard deviation below or above the mean. As report marks for Campus B (control) students were not related to their initial assessment literacy levels, Campus B students were not partitioned according to their assessment literacy. The three Campus A groups were then compared to the Campus B group as a whole.

There was no significant difference in report mark for groups according to *Assessment for learning* (AL), $F(302, 305) = 1.12, p = .34$. While the Campus A "high learning" group did have a higher average report mark ($M = 19.72$) than the Campus B group ($M = 18.20$), the effect size of this difference was small to medium ($d = .28$). This finding may be partly attributable to the Campus B group being significantly more likely to use assessment for learning ($M = 3.92$) than the Campus A group ($M = 3.72$) (pre-test measures)

prior to the intervention. In other words, Campus B participants had a greater average baseline level of using assessment for learning purposes.

In relation to the *understanding* dimension (AU), average report marks were significantly different across groups, $F(294, 297) = 3.67, p = .01$. Of most interest was any potential difference between the Campus A "high understanding" and Campus B group. Further analysis revealed that while the Campus A high understanding group had a greater average report mark ($M = 20.95$) than the Campus B group ($M = 18.2$), this difference was not significant ($p = .12$) using Tukey's Honestly Significant Difference (HSD) test. However further examination revealed a medium effect size ($d = .54$), indicating that this non-significant finding is instead due to poor power.

Likewise, for the *judgement* dimension (AJ), average report marks were significantly different across groups, $F(304, 307) = 3.43, p = .02$. Further analysis revealed that while the Campus A "high judgment ability" group had a higher average report mark ($M = 21.23$) than the Campus B group ($M = 18.2$), this difference was not significant ($p = .18$) using Tukey's HSD test. Again further examination revealed a medium effect size ($d = .58$), indicative of a practically meaningful effect. Like the regression results, these trends again reflect the importance of improving judgment in relation to enhancing student learning outcomes.

In regard to *minimum effort orientation* (MEO), no overall significant difference between groups was found for report mark, $F(303, 306) = 2.21, p = .09$. While the Campus A "high effort" (low MEO) group did have a higher average report mark ($M = 19.48$) than the Campus B group ($M = 18.20$), the effect size of this difference was small ($d = .24$).

Discussion and conclusions

This study theorised the notion of *assessment literacy* as multi-dimensional, and has shown how the dimensions of assessment literacy differentially contribute to the educational gains derived from this pedagogical intervention. Specifically, after controlling for prior academic ability and motivational attitude, one dimension of assessment literacy stands out as the "high-leverage" dimension – the ability to judge actual works against criteria and standards. The importance of this finding is that it was the nature of the

intervention (i.e. getting students to look at and judge actual examples of student work) that created the gains in this dimension of assessment literacy. This implies that interventions aimed at garnering enhanced learning from assessment, should target the development of *assessment literacy*. This in turn means creating an emphasis on a meta-dialogue about assessment, its purposes and how it functions. A further implication is that gains typically attributable to formative feedback could be enhanced not by a more detailed explication of the feedback by lecturers but rather by deploying assessment literacy (judgement)-enhancing protocols at the formative feedback points during the semester.

These findings support the view that helping students to develop their ability to judge their own and others' work will likely enhance their learning outcomes. This is consistent with the findings of the ASKE project (O'Donovan, Price & Rust, 2004), however the present study shows that significant gains could be expected from shorter interventions than those reported by the ASKE team, if the interventions are targeted at this dimension of assessment literacy. Further, the present study provides a theoretical explanation of how gains are made in student learning through interventions such as these, by specifying a mechanism (the development of assessment literacy) by which the gains are made: interventions which give student practise in judging work against standards, develops the judgement dimension of *assessment literacy*, which in turn allows them to perform better themselves on similar tasks.

Beyond *assessment literacy*, then, this study alerts us to another new concept as it describes a case in which there is a good educational return on the pedagogical investment made. This idea is important given that vast array of pedagogical strategies available to teachers in higher education, which vie for attention against a backdrop of increased workloads, larger and more diverse student cohorts, and diminishing resources for administrative support in universities. If we are to make a case for changing pedagogical practice, teachers themselves will want to know what return they can expect for their students from the changes being proposed. The educational return on pedagogical investment is a neat way of capturing this thinking.

Although the gains in student marks were modest in this case, the intervention was extremely brief, just 50 minutes. It is the potential leverage of the development of the ability to judge standards that

makes it worthwhile considering incorporating this kinds of assessment literacy developing protocol into regular teaching practice. In this study the gain was a 10% increase in the marks (approximately 2 in 20) on the task to which the protocol applied.

It may now be speculated that the gains from each extra hour of such an intervention may not be linear, but could well be curvilinear, boosted by practice effects over repeated occasions. Further it may be speculated that the effects of this kind of intervention may persist into later years of students' degrees if conducted in the first year of their studies. Such speculations can be tested in future studies.

Finally, It is worth noting that the students and teaching staff were generally positively disposed towards the intervention and the convenor has up-scaled its deployment into all cohorts and possibly other courses. Student comments such as "…now I understand what it's all about…" and "I think I'm starting to get this…" and "Now I know what's expected of me" indicated that students had learned from the experience. Comments from teaching staff included "What a fabulous activity…" and "…really useful" indicated that not only students but also teaching staff perceived benefits in the intervention.

References

Allen, P., & Bennett, K. (2008). *SPSS for the health and behavioural sciences*. Melbourne, Victoria: Thomson.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. . *Studies in Continuing Education, 22*(2), 151-167.

Braxton, J. M. (Ed.). (2000). *Reworking the student departure puzzle*. Nashville: Vanderbilt University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Francis, R. A. (2008). An investigation into the receptivity of undergraduate students to assessment empowerment. *Assessment and Evaluation in Higher Education, 33*(5), 547-557.

Gibbs, G., & Dunbar-Goddet, H. (2009). Characterising programme-level assessment environments that support learning. *Assessment and Evaluation in Higher Education, 34*(4), 481-489.

Gibbs, G., & Simpson, C. (2004). Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education*(1), 3-31.

Nicol, D. (2009). Assessment for learner self-regulation: enhancing achievement in the first year using learning technologies. *Assessment and Evaluation in Higher Education, 34*(3), 335-352

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218.

O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9*(3), 325-335.

Rust, C., O'Donovan, B., & Price, M. (2005). A social constructivist assessment process model: how the research literature shows us this could be best practice. *Assessment & Evaluation in Higher Education, 30*(3), 231-240.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving Students' Learning by Developing their Understanding of Assessment Criteria and Processes. *Assessment & Evaluation in Higher Education, 28*(2), 147-164.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159-179.

Sadler, D. R. (in press). Beyond feedback: Developing Student Capability in Complex Appraisal. *Assessment and Evaluation in Higher Education*.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics.* (4th ed.). Needham Heights, MA: Allyn and Bacon.

Yorke, M. (2005a). *Employability in Higher Education: what it is – what it is not*. York: Higher Education Academy.

Yorke, M. (2005b). *Issues in the assessment of practice-based professional learning*: Open University.