Portfolio: Is it a valid and reliable assessment instrument for academic writing skills?

Deirdre McClay, Letterkenny Institute of Technology

Abstract

This paper critiques one assessment strategy of a level 9 academic writing module. The module supports masters students in a School of Business by utilising a number of innovative approaches to teaching and assessing of academic writing skills. The assessment instruments include a portfolio of evidence of improvement in writing skills and a series of online discussion postings; the module is 100% continuously assessed. The use of portfolio is critiqued based on both validity and reliability of the assessment instrument, and also on conformity with the concept of formative assessment.

1: Introduction

The system for critique is an assessment strategy of a masters level, 5 credit, academic writing module. The module is designed to support taught masters students in a School of Business, and is 100% continuously assessed. It provides a novel approach to the teaching and assessment of writing skills, using a virtual learning environment (VLE) with assessed discussion postings and a portfolio. Investigation of this assessment strategy may add to the body of work on supporting student writing (Wingate, 2006; Dowling and Ryan, 2007; Cleary et al, 2009).

The focus of this critique is a portfolio of evidence of improvement in writing skills. It includes:

1. Tutor and peer review of a draft assignment from another module, usually the dissertation proposal.
2. Final, re-drafted assignment based on feedback from peer and tutor review (Bharuthram and McKenna, 2006).
3. Reflective account of improvement in writing skills.
4. In-class reflections on current state of knowledge, explicitly linked to improvement in writing skills within the reflective account.

The criteria of critique are validity, reliability, and assessment for learning. Validity and reliability are chosen as this is a higher education assessment with an inherent level of subjectivity (Bloxham, 2009); consequently, there are issues to interrogate around innovative assessment. The third criterion is a core motivation of the module designer, and investigation of how the portfolio fits with assessment for learning would aid future development of this module.

2: First principles: problems of validity and reliability in the higher education context, assessment paradigms and portfolio definition

Prior to critiquing the portfolio, certain issues need clarification. Firstly, there are recognised problems in higher education of using traditional concepts of validity and reliability which are largely based on standardised testing (Gipps, 1994; Bloxham, 2009). Therefore, it is important to define validity and reliability in a way that facilitates critique of a small scale assessment. Secondly, it is arguable that there are two paradigms of assessment with different emphasis (Gipps, 1994), and different outlooks on validity and reliability apply. Finally, the critiqued assessment is described as a portfolio, and literature on portfolio has been valuable; however, it raises questions around the definition of portfolio.

In higher education there are contextual issues around validity and reliability of assessment. Validity is traditionally defined as an assessment that "measures what it purports to measure" (Wiliam, 1992, p14; Gipps, 1994, pvii), which is based on large scale standardised testing regimes that purport to be objective. Whereas, validity in higher education assessment is more "judgment based", comprising content review by subject expert, high in subjectivity (Kane, 2001). Broadfoot (2007, p180) defines valid assessment as testing that "faithfully reflects the level of achievement or skill that it is designed to measure." This emphasises assessment design, and skills of student user, and is more suited to the critique of this portfolio.

As with validity, higher education assessment can face reliability problems. Reliability is defined in terms of accuracy and consistency; in other words, the reliability of achieving the same score twice from administering similar tests twice, or marking by different assessors (Gipps, 1994). In the context of reliability, Broadfoot (2007) uses the

term ‘dependability’, also, Wiliam (1992) argues that dependability requires accuracy, reliability and confidence in what a result discloses about attainment. However, the sophisticated tasks and higher level skills associated with tertiary level are difficult to test with a high degree of reliability (Wiliam 1992; Broadfoot, 2007). For example, studies reveal difficulties with marker reliability of essays (Bloxham, 2009; Knight, 2006). Therefore high reliability can be difficult to achieve in higher education, and valuing accuracy can lead to standardised rather than innovative assessment solutions (Parkes, 2007).

At this stage, it is also helpful to describe examples of two assessment paradigm (Gipps, 1994): the traditional psychometric model, and educational assessment. The former encompasses aspects of positivistic thinking, based on IQ testing, fixed intelligence which is measureable, and uses norm-referencing. Overall, it places high value on the reliability of tests, but validity is more problematic, as only certain types of outcome can be tested with high levels of standardisation. In contrast, educational assessment concentrates on competence, not intelligence, that is, something changeable with time and experience. It is based on criterion-referencing, and seeks to use assessment constructively to aid learning rather than measure it. According to Gipps (1994), it displays characteristics which are often the reverse of the traditional model, showing high validity, but lower reliability than standardised testing.

The portfolio under critique fits with the educational assessment paradigm. Aspects of it are ipsative - student performance is related to student past performance (Wiliam, 1992). In the draft–re-draft, improvement in writing skills is assessed against comparison with the student’s own work, not peers. Criterion referencing is also used, particularly with peer review, but evidence and documentation on that aspect of the portfolio could be improved. Also, competence is tested rather than intelligence, as the underlying focus of the assessment strategy is to evidence improvement in writing skills using examples of work over time. Assessments are not under controlled circumstances, and there is relaxing of rules of standardisation. However, there is one discrepancy, which is identification of best rather than typical work. This is not evident in the portfolio, which is also relevant to issues around definition of portfolio discussed below. Currently the portfolio evidence is mandated by the module tutor, and all work generated

is assessed. Therefore, a development of this model, which is also more consistent with traditional definitions of portfolio, would be to allow student choice of work, moving to a model of best work.

As regards portfolio, there is a lack of definition as portfolio are rarely standardised (Yaunkun et al, 2008; Meeus et al, 2009)**.** However, Klenowski et al (2006) emphasise a collection of work which includes a reflective account, stating that all work is student selected. This and other studies (Baume et al, 2004; Yaunkun et al, 2008; Meeus et al, 2009) use student choice in definition. Undoubtedly this does not fit with the critique portfolio, as most tasks are tutor mandated. However, there is student choice in the assignment for draft-re-draft, and there is a reflective account. Consequently, this assessment strategy is a portfolio-hybrid, making studies of portfolio relevant. An interesting question might be whether or not the integral issue of more student choice, and an ethos of best work, and might improve this assessment strategy? This is explored further under assessment for learning.

3: Validity

Validity is not a simple concept, evidenced by Wiliam (1992) identifying nine different aspects, demonstrating how it is multi-faceted. A number of these can be used to critique the validity of the critique portfolio, whilst bearing in mind that it is a small scale assessment in the educational assessment paradigm and no empirical analysis is done. Messick (1989) identifies two major threats to validity, as described by Gipps (1994). These are construct underrepresentation, when things are underrepresented that should be assessed, and construct irrelevance variance, when things are assessed that need not be assessed. Based on this analysis, the validity critique should ensure underlying evidence that the assessment is a good measure of what it is supposed to measure, being mindful of the use of assessment information. In order to critique the portfolio based on validity, it is useful to consider aspects of validity under Wiliam (1992), the two major threats to validity described by Messick (1989), and where there are gaps apparent, portfolio validity studies.

Firstly, face validity, is stated as "…the result of the assessment *looks* as if it will mean what it is supposed to mean" (Wiliam, 1992, p14). The end result of the critiqued portfolio assessment is a combination of marks from a range of writing tasks. For example, these tasks demonstrate use of higher level skills in the peer review, improvement in the use of academic conventions through the draft-re-draft exercise, and a reflective account of improvement in writing. Therefore, the composite mark on writing skills has high face validity.

Secondly, Wiliam defines content and descriptive validity as similar but with different emphasis. Content validity is that "…the test does indeed assess the content that it claims to address" and descriptive validity requires the test "…is actually measuring what its descriptive scheme contends it is measuring" (Wiliam, 1992, p14-15). Consequently, content must be addressed, but also issues such as compatibility with syllabus, and skills and processes. Content is often assessed by experts looking at both task criteria and student answers, and with the critique portfolio, an external examiner moderates. The fit with syllabus, skills and processes is also high, as the variety of tasks cover all learning outcomes. A mapping exercise of all assessment tasks of the portfolio to the module learning outcomes reveals that all module learning outcomes are covered by the portfolio, some more than once. This would suggest good validity but possible over assessment.

Thirdly, convergent and discriminant validity are developments of intrinsic validity involving comparison of assessments of different topics measured in different ways (Wiliam, 1992). Convergent involves the same topic being measured by two different methods. This is done through the portfolio, as for example, effective writing skills are assessed through tasks designed to practice directly that skill (draft–re-draft), and also through a reflective account which applies writing skills, but demands reflection. In practice, the marks on the reflective account are lower than on the draft-re-draft; however, one explanation for this is lack of training in reflective analysis, and also, a lack of detailed criteria on the reflective account. Thus, more explicit criteria and training might improve this outcome.

Finally there is the overarching concept of construct validity. It encompasses the idea that validity is about collecting evidence to support the declared meaning of

assessment, being mindful of consequences of assessment results (Messick (1989) as cited by Gipps (1994)). In the instance of this portfolio, it is arguable that the portfolio assessment is moderate to high in validity, because there are a wide range of assessment tasks. Moreover, this portfolio is only one assessment strategy of the module, and also module results are a small part of a composite programme mark (Bloxham, 2009). However, it is also useful to look at Messick's threats to validity. Construct underrepresentation is unlikely as all learning outcomes are covered by assessed tasks; it is more likely that there is construct overrepresentation. Also for the same reason, there is no finding of construct irrelevance variance, except to emphasise that some skills may be over assessed.

In summary, the portfolio is high in face validity, and mostly high in content, descriptive and intrinsic validity. Wiliam (1992) also looks at criterion-related, curricular, and instructional validity, but these are difficult to assess without empirical analysis**.** However, importantly, there are no significant weaknesses under construct validity. Nevertheless, due to unknowns in the critique, it is worthwhile considering studies on portfolio validity. Two such studies are Yaunkun et al (2008) and Meeus et al (2009) which analyse validity of teacher training portfolios. Both studies encounter portfolio definition problems, and one suggestion is to outline assessment objectives of the portfolio, aiding interrogation of validity (Meeus et al, 2009). This would improve the portfolio under critique - a clear statement of objectives leading to a dedicated and more detailed marking scheme. Also, both studies have similar findings on the type of competency that can be validly assessed by portfolio. Validity is low on tasks that are indirectly assessed by portfolio alone. For example, teacher performance is described as a competency that requires other evidence, such as expert observer. However, reflective accounts are found to be valid, because worst work is chosen in addition to best in order to demonstrate improvement.

In relation to the academic writing portfolio, since writing skills are being assessed, the portfolio tasks are directly assessing the desired competency, thereby overcoming a portfolio weakness (Yaunkun et al, 2008; Meeus et al, 2009). Also, the reflective account, which is part of the portfolio, is described as a valid method of assessing professional development, and is valuable to development of writing skills.

However, Meeus et al (2009) stress the importance of well structured, deep, broad and supported reflection. This is a weakness of the critiqued portfolio; it needs more supporting materials on reflective accounts.

Finally, Gipps (1994, p98) categorises portfolio under performance assessment, which is defined as "assessment carried out using tasks which are performance based." Performance assessment gives enhanced validity, especially construct and consequence validity, and helps to assess higher level skills. All of these factors fit with the critiqued portfolio. However, Gipps (1994) recommends training of raters, and moderation of results (also, Bloxham, 2009), and declaration of underlying cognitive requirements. The critiqued portfolio is moderated by an external examiner, but not a second marker; there is no specific training of raters, however, there is also only one internal marker. Perhaps a more realistic improvement would be a statement of cognitive requirements which fits with the suggestion for enhanced marking criteria.

4: Reliability

According to Wiliam (1992), there is a trade-off at the core of reliability, that is, a decision in relation to a particular assessment on acceptable levels of precision. Thereby, the finer the grading, the less accurate it will be, and the acceptable level of inaccuracy depends on the purpose of assessment. Also, there is argument that complex tasks, common at tertiary level, increase the variation in marks among assessors (Knight, 2006; Bloxham, 2009). Since academic writing skills are complex tasks, reliability may be problematic, especially where numeric grading is used, and with this particular module, grades are numeric. In addition, the practice of writing, and the goal of achieving a high level of tutor feedback, limits class size, and smaller group assessment by teacher is more open to abuse than large scale standardised testing (Broadfoot, 2007). Therefore reliability of the portfolio is more likely to be a problem than validity.

Wiliams (1992) describes reliability as consistent results, and relates three aspects of consistency with different types of reliability. They are test-retest reliability, split-half reliability and mark-remark reliability, and traditionally these are measured statistically, and associated with large scale standardised testing. However, lack of statistical reliability testing may not be problematic, as Wiliam (1992) concludes that such tests

show high procedural reliability, but can still, in a variety of ways, leave an individual with an unreliable mark. Moreover, in a similar argument on higher education assessment, Bloxham (2009) claims that some traditional features of the university system concentrate on procedures of assessment with no obvious increase in reliability.

There are small scale higher education studies of portfolio mark-remark reliability (Meeus et al, 2009; Baume and Yorke, 2004), but they involve paying experts to remark work. The results are useful to this critique, but none of the reliability measures are feasible. What is realistic, are findings and suggestions from portfolio reliability studies (Meeus et al, 2009; Baume and Yorke, 2004).

Meeus et al (2009) argue that problems of reliability are inherent in portfolio, as they are not standardised, yet too much standardisation would damage the tool. There are five suggestions for aiding reliability of portfolios:

1. One assessment protocol for all assessors.
2. A checklist of assessment criteria.
3. Use of holistic marking, not analytical marking.
4. Training of markers.
5. More than one marker.

Applying these suggestions, there is protocol for the critique portfolio but it needs systematic documentation. Also, there are assessment criteria (peer review criteria, and a short marking scheme), but they could be developed. With respect to marking, analytical marking grades separate sections deriving a composite mark; whereas, holistic marking derives a global mark. Under holistic marking, different criteria can be used to assess the various sections, but they are qualitative in nature. The critique portfolio is marked in separate sections, and a composite numerical mark awarded; this is analytical marking. Some consideration might be given to holistic marking based on a qualitative marking framework. However, this marking suggestion from Meeus et al (2009) is based on a single study (Baume and Yorke, 2004), where mark-remark reliability is low, and markers often circumvent a detailed analytical scheme to achieve a holistic mark. The critique portfolio has a much simpler structure.

A further consideration is the training of markers, which may be less of an issue here, as there is only one marker, and an external moderator. However, Bloxham (2009)

argues that issues of double marking and external review of assessments may achieve procedural reliability, without generating enhanced accuracy. Similarly, Partington (1994) identifies double marking as helpful for student confidence in the system, but then explores problems of second markers. These arguments acknowledge that many skills at tertiary level are difficult to assess accurately, but standards are maintained through expert review.

Therefore, in conclusion, it is difficult to estimate the reliability of this form of assessment, especially without resources to ensure mark-remark reliability. However, the range of assessment tasks given is helpful, and there is documentation available to students and the external examiner on protocols and grading. The main suggested improvement would be a more detailed marking scheme, and the possibility of moving to holistic marking.

5: Assessment for Learning

The final critique looks at how the academic writing portfolio fits with assessment for learning. Aspects of summative and formative assessment are explored, followed by literature on portfolio and assessment for learning. Finally, there is an evaluation of the usefulness of further formative assessment strategies.

Newton (2007) describes three distinguishing characteristics of formative and summative assessment. They are purpose, timing and level of generalisation. Under purpose, formative assessment is "learning to learn", whereas summative is grading. Timing is often distinguished, with formative being during course of study, while summative is end stage. Then generality describes a difference in focus, with formative showing narrow focus on specific areas, and summative, broadly focused. Moreover, a further development in definition has been the ongoing importance of feedback loop under formative assessment. Broadfoot (2007) describes feedback to modify both teaching and learning; corrective feedback helps students develop, and eventually to self-monitor, and the teacher to adapt teaching to student needs. This is consistent with the views of Black et al (2003), formative assessment and feedback must influence teaching and learning, and are mostly informal. Therefore, formative assessment is essentially part of teaching, and in some circumstances it is ipsative (Harlen and James, 1997). This

becomes another distinguishing factor (Harlen and James, 1997), summative assessment can be norm-referenced (assessment with reference to peers), or criterion referenced (assessment with reference to criteria). However, formative assessment is criterion referenced or ipsative with ideas around best rather than typical work.

In evaluation of the academic writing portfolio, all tasks are summatively assessed for grading. Also, students may adapt learning based on peer and tutor feedback in order to complete a better assignment, but the tutor does not formally adapt teaching methods. The portfolio is also mostly tutor mandated, thereby fitting better to models of typical work. None of the above conclusions conform well to formative assessment.

Conversely, underlying all aspects of the portfolio are either criterion referencing (e.g. peer review), or ipsative (e.g. draft-redraft, and reflective account). Additionally, there is a feedback loop which aids learning to learn, and influences the final assignment document which is assessed in the portfolio, and is also graded in another module. The portfolio draft is ipsatively assessed (Wiliam, 1992), reflecting student achievement from first draft, through feedback to final draft. All students complete a peer review which is criterion referenced and tutor mandated and summatively assessed. However, that same peer review, is given to the peer as feedback, along with a separate tutor review; these reviews are formative, the purpose being learning to learn and qualitative in nature. Use of a feedback loop, improves work that is summatively assessed.

Therefore, the portfolio has aspects of formative assessment built into its creation (tutor and peer review), but it is summatively assessed. Overall, it is influenced by ideas that are common to formative assessment, for example, ipsative and criteria referenced assessment.

In addition, some investigation of portfolio and assessment for learning is merited. However, Klenowski et al (2006, p268) state that there is little research in the area of portfolios for ‘formative and learning purposes’ at postgraduate level (learning portfolios). The particular study looked at three different tools with some formative purpose. Resulting from a cross case analysis, the following advice is proffered. Tutors must explain carefully the portfolio purpose, whether formative, summative or both. Students must be facilitated towards an understanding that portfolio goes beyond a collection of evidence; there must be ‘meta-learning’, moving beyond content and

reflecting on the process of learning. Also, the tool must be generative, rather than unconnected evidence, thoughts and reflections, they must be integrated. Invariably, a move to portfolio, whether formatively or summatively assessed, has an impact on pedagogic practice (Klenowski, 2000). There is more emphasis on independent learning, growth of learning over time, collaborative practice, self-evaluation and reflection (Klenowski, 2000). Students need support in making this learning shift particularly in the early stages. There must be strong course documentation and underlying facilitation strategies, and group support is helpful (Klenowski et al, 2006, p268). However, one weakness of this approach for the academic writing portfolio, is that the tools described by Klenowski et al (2006) are compiled by working professionals learning in their own field. In other words, they may be more confident and situated learners at the outset, compared with taught masters' students facing a dissertation for the first time.

In evaluating this approach for revision of the critique portfolio, one obstacle would be resources in terms of staff time, as there is only one tutor. Nevertheless, the module currently has two strands of assessment, as there are eight discussion postings in addition to the portfolio. However, since all learning outcomes are covered by the portfolio alone, it is arguable that the students are over assessed, and a move towards a single assessment portfolio is merited. Some of the current discussion posting tasks might be reconfigured as class based tasks, with feedback from tutor and peers. Thereby a collection of a broader range of evidence, formatively assessed, would aid movement towards a student choice portfolio, with more meaningful reflective accounts. The previous suggestions under validity and reliability of more detailed assessment criteria would also facilitate better communication to students.

6: Conclusion

This critique of a higher education, masters level portfolio, has been mindful that the assessment fits within the educational assessment paradigm, and is open to problems of subjectivity. It is critiqued on validity and reliability, in so far as this is possible without empirical analysis. Validity is found to be acceptable, with all learning outcomes assessed, and a range of methods used; in fact there is argument for over assessment. Portfolio literature is also investigated, contributing to the conclusion that validity could

be strengthened by better documentation of marking criteria, and explicit articulation of objectives of the tool. Under reliability, traditional methods are not feasible, and portfolio literature is explored. Consequently there are suggestions for a protocol of assessment issues, an enhanced marking scheme, and the possibility of holistic marking. Finally, assessment for learning is researched for possible improvements to the portfolio. This is a fruitful exercise, as enhancing the formative aspects of the overall module and restructuring the portfolio to include student choice, would help with issues of over-assessment, and allow more tutor time for developing student understanding of marking criteria and reflective learning skills.

References

Baume, D and Yorke, M. (2004) What is Happening When We Assess, and How Can We Use Our Understanding of this to Improve Assessment, *Assessment and Evaluation in Higher Education*, 34(4) pp.401-413.

Bharuthram, S, and McKenna, S. (2006) A Writer-Respondent Intervention as a Means of Developing Academic Literacy, *Teaching in Higher Education,11*(4) pp.495-507

Black, P, Harrison, C, Lee, C, Marshall, B and Wiliam, D. (2003) *Assessment for Learning: Putting It into Practice*, Buckingham: Open University Press.

Bloxham, S. (2009) Marking and Moderation in the UK: False Assumptions and Wasted Resources, *Assessment and Evaluation in Higher Education*, 34(2) pp.209-220.

Broadfoot. P. (2007) *An Introduction to Assessment*, London, Continuum.

Cleary, L., Graham, C., Jeanneau, C. and O'Sullivan, I. (2009) Responding to the Writing Development Needs of Irish Higher Education Students: A Case Study, *All Ireland Journal of Teaching and Learning in Higher Education*, 1(1) pp.4.1-4.16. Available online at: http://ojs.aishe.org/index.php/aishe-j/article/view/4 [accessed 05 January 2011]

Dowling, L, and Ryan, O (2007) Academic Skills Development and the Enhancement of the Learning Experience, Paper presented at AISHE conference 2007 available online at: http://www.aishe.org/events/2006-2007/conf2007/proceedings/paper-32.pdf [accessed 05 January 2011]

Gipps, C. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*, London: Falmer Press.

Harlen, W. and James, M. (1997) Assessment and Learning: Differences and Relationships between Formative and Summative Assessment, *Assessment in Education: Principles, Policy & Practice*, 4(3), pp. 365-379.

Kane, M.T. (2001) Current Concerns in Validity Theory, *Journal of Educational Measurement*, 38(4) pp. 319-342.

Knight, P.T. (2006) The local practices of assessment. *Assessment & Evaluation in Higher Education*, 31(4) pp. 435–52.

Klenowski, V. (2000) Portfolios: Promoting Teaching, *Assessment in Education: Principles, Policy & Practice*, 7(2), pp. 215-236.

Klenowski, V., Askew, S. and Carnell, E. (2006) Portfolios for Learning, Assessment and Professional Development in Higher Education, *Assessment and Evaluation in Higher Education*, 31(3) pp.267-286.

Meeus, W., Van Petegem, P., and Engels, N (2009) Validity and Reliability of Portfolio Assessment in Pre-service Teacher Education, *Assessment and Evaluation in Higher Education*, 34(4) pp.401-413.

Messick, S. (1989) Validity, in Linn, R. (Ed) *Educational Measurement* (3rd edn) American Council on Education, Washington, Macmillan, cited by Gipps, C. (1994) *Beyond Testing: Towards a Theory of Educational Assessment*, London: Falmer Press.

Newton, P.E. (2007) Clarifying the Purposes of Educational Assessment, *Assessment in Education: Principles, Policy & Practice*, 14(2) pp. 149-170.

Parkes, J. (2007) Reliability of Argument, *Educational Measurement: Issues and Practices*, Winter, pp.2-10.

Wiliam, D. (1992) Some Technical Issues in Assessment: A User's Guide, *British Journal of Curriculum & Assessment*, 2(3), pp. 11-21.

Wingate, U (2006) Doing Away with Study Skills, *Teaching in Higher Education*,11(2) pp.457-469.

Yuankun, M., Thomas, M., Nickens, N., Anderson Dowing, J., Burkett, R.S. and Langon, S. (2008) Validity Evidence on an Electronic Portfolio for Pre-service Teachers, *Educational Measurement: Issues and Practice*, Spring pp. 10-24.