Education quality and 'exam hurdles': a flawed relationship

Daryll Cahill
School of Accounting, RMIT University, daryll.cahill@rmit.edu.au

Paul Myers
School of Accounting, RMIT University, paul.myers@rmit.edu.au

Louise Robinson

This paper determines that an 'exam hurdle' can be an unreliable and unfair mechanism by which to grade students. Cahill and Hilliar (2009) found that exam hurdles are widely used across numerous disciplines in Australian universities. An exam hurdle is usually defined as a way of determining the eligibility of a student to pass a subject, where that student must obtain at least 50 per cent of the marks available and pass the end-of-semester examination. Academics state that a primary reason for their use of an exam hurdle is quality control – ensuring a basic level of capability in students who pass their subjects. However, this paper demonstrates that this quality-control role is not at all guaranteed and that an exam hurdle, per se, cannot reliably identify students who should be failed. It also outlines how the operation of an exam hurdle can involve significant procedural unfairness in the treatment of students. From 2012, the Tertiary Education Quality and Standards Agency (TEQSA) is planned to be the Australian national regulator and assurer for quality outcomes for all higher education providers. The unreliability of an exam hurdle as a quality control mechanism could prove problematic if examinations are the major method of evidencing such outcomes.

Keywords: assessment design; exam hurdles; quality control.

Introduction and background

Assessment is central to the experience of most university students, as it is a common factor across numerous disciplines. Clearly, assessment is a pivotal variable in the relationship between teaching and learning (Ramsden, 1992).

University study comes with an output – the credential. At its core, this output requires evaluative mechanisms that measure whether a student has obtained the necessary knowledge and/or skill to pass a subject and a program (Rowntree, 1977; Watty et al., 1997). Correspondingly, the critical analysis of assessment methods used at universities is longstanding, cross-disciplinary and international (see, among a vast body of work, Rowntree, 1977; Smith, Miller, & Crassini, 1998; Gammie, Paver, Gammie, & Duncan, 2003; Wood, 2007). Such research has ranged across disciplines as diverse as accounting, nursing and language studies, to name but a few. However, despite the breadth and longevity of research into assessment, there are still issues largely untouched. For example, Cahill and Hilliar (2009) found that 'exam hurdles' appear to be a prevalent form of hurdle used across numerous disciplines in Australian universities, yet little work has been done to evaluate such hurdles. (An exam hurdle is usually defined as a way of determining the eligibility of a student to pass a subject, where that student must obtain at least 50 per cent of the marks available and pass the end-of-semester examination.) Further, Cahill and Hilliar (2009) state that academics primarily used an exam hurdle because they saw it as a quality control mechanism that ensured a basic level of capability in students who passed their subjects.

Assessment occurs for a variety of reasons, for example, as a mandatory requirement of the university, to provide feedback to students and teachers, to provide motivation to students, to help in maintaining standards, and as preparation for life (Rowntree, 1977). The assessment methods employed, and the internal and external evaluation and perceptions of those methods, are influenced by a combination of educational effectiveness and resource efficiency (Thompson & Bartels, 1999; Gammie et al., 2003). Grading students has implications for their future. Students see grades as determinants of extrinsic rewards, such as future employment or postgraduate selection

(Rowntree, 1977). In addition, assessment from a student's viewpoint often defines the curriculum and, thus, strongly affects the learning mode adopted by the student (Gammie et al., 2003). However, despite the breadth of research into assessment, there are still issues and activities left largely untouched.

This paper critically evaluates this quality control role for exam hurdles. First, exam hurdles and quality controlled are defined and discussed. Next, nine hypothetical student assessment result profiles for a subject are established, and the effect of an exam hurdle assessment in the subject is tested and critiqued in each of these cases. While the student assessment result profiles in this critique are hypothetical, they are based on actual assessment result profiles of students at an Australian university. As will be outlined below, this paper concludes that the quality control role of exam hurdles does not stand up to evaluation.

What is an exam hurdle?

As outlined above, an exam hurdle is a requirement of a student to achieve a pass on an exam – usually 50 per cent of the marks available. For a student to pass an overall subject the student must obtain a least 50 per cent of marks available (outside the exam) in a semester, and the exam hurdle must also be passed. Failure to meet this standard on the end-of-semester exam means failure of the subject, irrespective of whether the total marks achieved for that semester equal more than 50 per cent of the total marks available. Commonly, though not universally, when an exam hurdle is used, if a student fails the exam hurdle but achieves marks greater than 50 per cent of those available for the semester, the final mark that student receives for the semester will be their examination score plus their other in-semester results, but set to a discounted maximum. For example, if a student fails the exam hurdle assessment but has received 55 out of 100 for other assessments, they will be awarded only 45 out of 100 for the non-hurdle assessments. Whereas if a student fails the examination but has received less than 50 per cent of the marks for the semester, the final mark that student receives is their examination score plus their other in-semester results. The assumption central to the exam hurdle is that passing the exam is evidence by an individual student of a basic level of capability in the knowledge and skills covered by a subject.

Critically evaluating the outcomes of an exam hurdle

The critical evaluation undertaken of the exam hurdles that follow was based on a range of hypothetical student assessment profiles (see Table 1). Different student result profiles for a hypothetical subject are represented in each of the columns (S1–S9). This example includes five assessment items over one semester, where 40 per cent of the possible marks were obtainable in non-exam situations, while 60 per cent of the marks were obtainable in an invigilated examination. Here the examination assessment is a hurdle assessment: students must obtain at least 50 per cent in the examination to pass the course.

Table 1: Hypothetical student assessment profiles

Assessment item	Total possible marks	Students' marks								
		S1	S2	S3	S4	S5	S6	S7	S8	S9
Essay (solo work)	5	1	5	5	5	5	0	2	5	5
Test[1]	15	4	5	7	15	15	0	6	15	15
Project (group work)	15	6	8	15	15	15	0	6	15	15
Class presentation (group work)	5	1	2	5	5	5	0	2	5	5
Examination[1]	60	15	30	18	25	6	50	54	30	50
Total raw marks	100	27	50	50	65	46	50	70	70	90
Adjusted result (if necessary)		27	50	45[2]	45[2]	46	50	70	70	90
Pass (P) or fail (F)		F	P	F	F	F	P	P	P	P

1 Invigilated, individual work

2 Exam hurdle fails recalculated to a score of 45

Hypothetical examples of a range of outcomes

Exam hurdles and standard failure

An exam hurdle is not needed to identify the standard student failures in this subject, that is, those students who do not obtain 50 per cent of the semester's available marks. For example, student one (S1) has not passed any assessment items and is a clear, standard fail for the semester. The exam hurdle plays no role at all in, and is not needed for, identifying that this student has not performed sufficiently to be passed. In this subject, all students who do not score at least 50 per cent of the marks available in the semester are failed, without recourse to an exam hurdle.

Low diagnostic power of an exam hurdle

Student two (S2) has passed some but not all assessment items in the semester and is, clearly, a marginal student. However, the exam hurdle provides no diagnostic information about which capabilities this student insufficiently achieved through the semester. Nor does the exam hurdle provide an opportunity for remedial intervention, as it occurs at the end of a semester. Also, it does not give data to the staff or student about what needs to be improved or strengthened prior to moving on to the subjects in the next semester. This should concern any program that involves prerequisite progression. Passing the examination means that S2 is allowed to include the points achieved in the other assessments in the semester and, thus, passes. The exam hurdle has played no role in identifying the overall weak performance or lack of capabilities attained by S2.

Discounting non-examination assessment scores and provisional results

Both student three (raw score 50) and student four (raw score 65) have obtained 50 per cent or more of the marks available in this semester and have passed most assessment items. However, as they failed the examination, both receive a fail for the semester. Their raw scores are adjusted to 45 for the semester: a common practice in Australian universities. In this adjustment, presumably, as the examination and the test scores are from individual, invigilated work, they are considered to be an authentic evaluation of that student's performance and, therefore, are not changed. Instead, clearly, for these students it is their non-examination and non-test marks that, at the end of the semester, have to be retrospectively discounted for S3 and S4 to receive 45 for the semester. For this to happen, the marks for non-invigilated work for S3 would need to be discounted by approximately 20 per cent (25 becomes 20)

and for S4 by 80 per cent (25 becomes 5). There is no educational rationale for having different discount rates, especially as this result suggests that S3 was a better student than S4.

Moreover, student one (S1) also failed the examination and, under the conditions of this subject, is allowed to include the points achieved in the other assessments in the semester, though will still fail. No discount is applied to the non-invigilated work of S1, though that is done to S3 and S4. It does not seem logical to use different discount rates for some students and no discount rate for others. It seems unjust to deduct marks from students who pass an assessment but fail an examination, while not doing so to students who fail the exam and the semester in the standard manner.

The truth of the operation of the exam hurdle is that the scores awarded to *all* students for their non-invigilated assessment during the semester are strictly *provisional* until their examination papers are marked. After examination marking, depending on each student's performance on the examination and throughout the semester, some students are, and some are not, given the marks apparently earned during the semester. Awarding 45 for the semester as the de facto score for an exam hurdle fail, while mathematically expedient, is logically unsound as it requires an inconsistent discount rate to be applied to the grades previously assigned to students' non-invigilated work.

Equity and exam hurdles

On equity, student five (S5) fails the examination but has not obtained more than 50 per cent of the marks for the semester, and thus will keep all points awarded for non-invigilated assessment items and receive 46 overall. Both student three (S3) and student four (S4) fail the examination but obtained 50 per cent or more of the marks for the semester, and so will receive 45 overall, a lower result than S5, despite both S3 and S4 having better total performance than S5.

Non-invigilated assessment and exam hurdles

An underlying argument supporting exam hurdles seems to be that students' results in non-invigilated assessment items can only be considered an authentic evaluation of their contribution to and participation in those items if those students can then go on to pass an examination. This argument seems to be based on a combination of some or all of the following.

1 There has been inappropriate behaviour and outright cheating by these students.

2 Teaching team staff were not consistent in their marking of the non-invigilated assessment items.

3 The non-invigilated assessment items are fundamentally flawed educational practices and cannot be (were not) operated with integrity.

The apparent disconnection between results of the non-invigilated assessment items and the examination score for student three (S3) and student four (S4) seems to be the issue which most concerns academics, and drives the desire for an exam hurdle. Both S3 and S4 do well in non-invigilated assessment items, but not on the examination. For both of these students the poor score on the examination is taken as a sign by many academics that the assessment results they had obtained prior to the examination – which were, to this point, unquestioned – now have become less 'trustworthy' measurements. If this is the case, then student eight (S8) has a poor examination score (30 out of 60) that does not match this student's apparently excellent performance on the non-invigilated assessment items in the semester. The exam hurdle does not investigate or identify this anomaly. Logically, S3 and S4 might have the right skill sets to undertake non-invigilated assessment, but be poorly skilled to deal with examination conditions.

Failing an exam hurdle provides no proof at all that such students were cheating, 'free-riding' or committing plagiarism in the non-invigilated assessment items and cannot resolve flaws in non-invigilated assessment activities. An exam hurdle does not give any clear information concerning those students who have received inappropriate results for their non-invigilated work. If the goal is quality control over the marks awarded for non-invigilated assessment items, and if such items are suspected to be an inauthentic evaluation of student performance

in the cohort, then *all* students in the semester should have those results discounted, not just the over-50 marks, hurdle-fail students. It could be argued that this is a shortcoming of quality control.

Capabilities, standards and exam hurdles

Student six (S6) and student seven (S7) perform poorly (in S6's case there is no performance at all) on all assessment items, except the examination, where both did very well. However, unless the examination is a measurement of the entire material covered and the knowledge and skills sets of the semester, then S6 and S7, having failed some assessment items during the semester, would not have demonstrated the appropriate level of achievement, despite their overall score in the subject. On quality control, this exam hurdle plays *no* role in identifying the poor or non-performance of S6 and S7 in the semester's non-examination assessment activities. These two students have not reached a satisfactory level in all the capabilities assessed in this subject and, it may be assumed, have not satisfied the learning objectives of this subject. Thus, this exam hurdle has failed in its basic role as a quality control mechanism to ensure a basic level of capability in students who pass this subject.

Clearly, the combined interaction of the examination score and the exam hurdle rule significantly devalues the worth of performance by students in the other assessments of the semester. Academics, not students, design and operate their subjects. If a student can obtain 50 per cent or more of a subject's assessment marks in a semester, then they have demonstrated that they have a passing level under the educational structure put in place. To go on to add to a subject another filter, such as the exam hurdle, seems in some ways an admission of defeat by the academic. The use of an exam hurdle appears only focused on avoiding a type one error (not passing some one who should fail); there seems to be no attempt at all to investigate or eliminate any type two error (failing students who should be passed).

Invigilated versus non-invigilated assessment

The logic of an exam hurdle is that the examination score, involving the marking of invigilated individual work, is considered to be the authentic evaluation of that student's performance. For this statement to hold true means that the writing and marking of examination papers does not suffer from variation in approach, emphasis and scoring among drafters and markers.

There is a large body of research that has shown conclusively that assessment scoring is strongly, persistently and inescapably open to inaccuracy, and this holds true for exam marking as well as any other form of assessment (see, among many others, see Lane & Sabers, 1989; Stuhlmann, Daniel, Dellinger, Denny & Powers, 1999; Swartz, Hooper, Montgomery, Wakely, DeKruif & Reed, 1999). The process of marking any form of assessment is far from infallible, and the quality of marking is highly dependent on factors independent of and of no connection to the goals of the assessment item (Hamp-Lyons, 1989; Vaughan, 1991; Lumley, 2006). Among markers scoring the same item or task, variation has been shown to occur not because of difference in students' work but in the markers' background and personal predispositions (Barnwell, 1989; Shohamy, Gordon & Kraemer, 1992; Brown, 1995) and their individual approach to the weight of penalty to be applied to errors and mistakes (Englehard, 1994; Lumley & McNamara, 1995). Moreover, across time, markers have shown instability in their scorings of the same level of work from students (Congdon & McQueen, 2000).

Also, inconsistencies are apparent when S2 is compared to S4. The total individual invigilated work in the semester amounts to 75 points: test (15) and examination (60). Of this, S2 scores 35 out of 75, passing 46 per cent of this assessment mode. S4 scores 40 out of 75, being 53 per cent of this mode. If individual invigilated work is the most authentic mode, then S4 has performed in this mode better than S2; that is, S4 has 'passed' the total, individual, invigilated assessment mode for the semester; S2 has not.

Capabilities and passing an examination

An exam hurdle basically involves numerically obtaining at least 50 per cent of the marks available on an examination paper but it is entirely possible that such 'passing' of the examination is not reliable evidence of a demonstration of an appropriate level of capability by students. For example, an issue to be resolved is whether the

exam hurdle's requirement of passing the examination actually needs to be more specific; that is, a student must pass all questions on the examination. Teaching teams using an exam hurdle should consider the implications arising if students fail some questions on the examination, but still 'pass' the examination numerically and, thus, are not identified by an exam hurdle. As Table 2 shows, there are numerous hypothetical scenarios for numerically 'passing' an examination, though some scenarios appear to indicate a less-than-adequate coverage of a subject's knowledge and skills by the student.

Table 2: Student two examination performance scenarios

Exam question	Total possible marks	S2 performance scenarios			
		S2a	S2b	S2c	S2d
Question 1	13	6.5	5F	9	13
Question 2	10	5	5	3F	0F
Question 3	12	6	5F	8	0F
Question 4	8	3.9	5	3F	0F
Question 5	5	2.5	5	1F	5
Question 6	12	6	5F	6	12
Total	60	30	30	30	30

F Failure of question

The quality control role of an exam hurdle seems to imply that examination questions must be denoted as essential (that is, must be passed) and non-essential (need not be passed). Thus, clearly, the points' allocation on an examination involving an exam hurdle must ensure that passing all essential examination questions also means that at least 50 per cent or more of the available marks have been obtained. Not to do so leads to some difficult questions. For example, has an examination been 'passed' if a student has passed some, but not all of the essential questions and some, but not all, of the non-essential questions, and still obtained 50 per cent or more of the available marks?

Also, a teaching team using exam hurdles would need to indicate which examination questions do and do not canvass and evaluate essential capabilities. Obviously, if an exam hurdle is needed to force students to demonstrate a set level of capability in the core knowledge and skills covered by a subject, then the examination that involves a hurdle must assess all those items. Otherwise, an exam hurdle is not working as many academics assume it is, that is, as a de facto exit test. This opens up the issues: what capabilities can and cannot be examined on an exam? Which topics were questioned? If knowledge or skill capabilities have been assessed already in the semester, can they be left off the exam? All the above questions need to have been considered, answered and integrated in the subject's and the examination's design if an exam hurdle is to act as a quality control mechanism.

In fact, simply numerically passing an examination might not indicate a satisfactory level of achievement in a subject. An exam hurdle, which focuses on a raw overall score, can mislead by not taking into account the performance of students on all questions. For example, consider S2, who scored 30 out of 60 on the examination. Table 2 shows several scenarios for obtaining 30 out of 60, a few of which might not be an acceptable demonstration of all of a subject's basic knowledge and fundamental skill.

In scenarios S2b, S2c and S2d, this student fails 50 per cent of the questions on this examination. It is difficult to see how this is an adequate demonstration of acceptable achievement, especially in S2d where no points are obtained for three of six examination questions. In fact, as shown in Table 3, it might argued that while an examination score of 30 by S2 is higher than the score of 25 by S4, this 25 is a better achievement – a more acceptable performance.

Table 3: Student two and student four examination performance scenarios

Exam question	Total possible marks	S2 performance scenarios				S4 performance scenarios	
		S2a	S2b	S2c	S2d	S4a	S4b
Question 1	13	6.5	5^F	9	13	1^F	4^F
Question 2	10	5	5	3^F	0^F	5	5
Question 3	12	6	5^F	8	0^F	6	6
Question 4	8	3.9	5	3^F	0^F	4	4
Question 5	5	2.5	5	1^F	5	3	3
Question 6	12	6	5^F	6	12	6	3^F
Total	**60**	**30**	**30**	**30**	**30**	**25**	**25**

F Failure of question

Scenario S4a (failure of one of six examination questions) and S4b (failure of two of six examination questions) are arguably a better achievement, though a lesser score, than S2b, S2c and especially S2d, as both S4a and b demonstrate knowledge and skill to pass more questions in total.

In fact, as outlined in Table 4, it might be argued that the best performance for student four (exam score 25 out of 60) is a better overall demonstration of a subject's basic knowledge and fundamental skill, than the worst scenario for student two or even one possible scenario for student six (S6).

Table 4: Student two, student four and student six examination performance scenarios

Exam question	Total possible marks	S2d performance scenarios	S4a performance scenarios	S6 performance scenarios
Question 1	13	13	1^F	13
Question 2	10	0^F	5	10
Question 3	12	0^F	6	12
Question 4	8	0^F	4	2^F
Question 5	5	5	3	1^F
Question 6	12	12	6	12
Total	60	30	25	50

F Failure of question

Student six (S6) does perform exceptionally well on the examination but if exam question four and five cover core knowledge and/or skill, then this score does not reflect demonstration of all that is needed to be awarded a pass a subject. This is even more relevant for student two scenario d, which involves a score that is a pass of the examination, but has failure in three questions.

Exam hurdles and comparability of subjects

Cahill and Hilliar (2009) found that exam hurdles are widely used but not universally, across numerous disciplines in Australian universities. This raises an interesting point regarding the quality control role ascribed to exam hurdles. Those subjects that use an exam hurdle do so largely to mitigate against inauthentic assessment measurement in the non-invigilated assessment items. However, there are numerous common subjects across programs in universities. Statistics, for example, is a subject common to many Australian degrees. Some of these statistics subjects would operate an exam hurdle; some would not. As numerous, but not all, universities allow the use of exam hurdles as a quality control, this policy variation should mean that these subjects cannot be considered equivalent for exemptions and credit transfer. Unless shown the examination results for students, it is reasonable to question whether staff who teach statistics subjects with exam hurdles should accept the results from statistics subjects which do not include exam hurdles, as quality control has not been validated by an exam hurdle.

Conclusion and implications

This paper shows that reliance on an exam hurdle, which requires simply obtaining 50 per cent of the marks on an examination, is educationally unsound and does not deliver reliably any assurance of quality control.

An exam hurdle implies that those capabilities not tested on the examination are of secondary value. There is no doubt that an exam hurdle identifies the usually small number of students who can pass a subject but who cannot pass the examination in that subject (for example, S3 and S4 in the data above). On the other hand, an exam hurdle is not needed for the majority of standard failure students (for example, S1). It does not identify the marginal students, those barely passing each assessment item (for example, S2). It does not always identify students who have passed a subject but who have not demonstrated a basic level of knowledge and skill in all core capabilities of that subject (for example, S6 and S7). More importantly, it might fail students who deserve to be passed (for example, S4). Added to these factors, the operation of an exam hurdle can involve significant negative issues of procedural inequity and student motivation.

Considering their widespread use in Australian universities and the impact they have on students, exam hurdles should be used only if they are demonstrably based on sound educational reasons relating to the evaluation of critical capabilities in a subject. Much more thought is required in defining what is meant by ‘passing the examination’, and the functional relationship between an examination and an exam hurdle. Further, an exam

hurdle, per se, is not the solution to concerns about non-invigilated assessment activities. An exam hurdle does not identify whether students have, and specifically which students have, behaved inappropriately in those activities.

References

Barnwell, D. (1989). Naïve native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Biggs, J. (1997). Individual differences in study processes and the quality of learning outcomes. *Higher Education, 8*, 381-394.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

Cahill, D., & Hilliar, K. (2009, November). *Assessment Hurdles: more about workload than education*. Paper presented at the ATN Assessment Conference, Melbourne.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Englehard, G. (1994). Examining rater errors in the assessment of written compositions with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Gammie, E., Paver, B., Gammie, R., & Duncan, F. (2003). Gender differences in accounting education: an undergraduate exploration. *Accounting Education: An International Journal, 12*(2), 159-196.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp 229-244), Tubingen: Gunter Narr.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied Measurement in Education, 2*(3), 195-205.

Leveson, L. (2000, July). *A phenomenographic study of individual approaches to teaching in higher education*. Paper presented to the annual conference of the Accounting Association of Australia and New Zealand, Hamilton Island, Queensland.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*(1), 54-71.

Rowntree, D. (1977). *Assessing Students: How Shall We Know Them?* London: Harper & Row.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Smith, S. N., Miller, R.J., & Crassini, B. (1998). Approaches to studying of Australian and overseas Chinese university students, *Higher Education Research and Development, 17*(3), 261-276.

Stuhlmann, J., Daniel, K., Dellinger, A., Denny, R., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology, 20*(2), 107-127.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., DeKruif, R. E. L., Reed, M., et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(3), 492-506.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125), Norwood, NJ: Ablex,.

Watty, K., Cahill, D., & Cooper, B.J. (1997, November). Graduate attributes: perceptions of accounting academics. Paper presented at the 8th World Congress of International Association for Accounting Education and Research, Paris.

Wood, D. (2007). Teachers' learning communities: catalyst for change or a new infrastructure for the status quo? *Teachers College Record, 109*(3), 699-739.