Higher Education Research and Development Society of Australasia, Inc

Research and Development in Higher Education: Reshaping Higher Education Volume 33

Refereed papers from the
33rd HERDSA Annual International Conference

6–9 July 2010
Melbourne, Australia

Bryant, D. & Zhang, F. (2010). The impact of assessment on learning achievement. In M. Devlin, J. Nagy and A. Lichtenberg (Eds.) *Research and Development in Higher Education: Reshaping Higher Education*, *33* (pp. 145–155). Melbourne, 6–9 July, 2010.

Published 2010 by the
Higher Education Research and Development Society of Australasia, Inc
PO Box 27, MILPERRA NSW 2214, Australia
www.herdsa.org.au

ISSN 0 155 6223
ISBN 0 908557 80 9

This research paper was reviewed using a double blind peer review process that meets DIISR requirements. Two reviewers were appointed on the basis of their independence and they reviewed the full paper devoid of the authors' names and institutions in order to ensure objectivity and anonymity. Papers were reviewed according to specified criteria, including relevance to the conference theme and sub-themes, originality, quality and presentation. Following review and acceptance, this full paper was presented at the international conference.

The impact of assessment on learning achievement

Dennis Bryant
University of Canberra, Canberra, Australia
dennis.bryant@canberra.edu.au

Felicia Zhang
University of Canberra, Canberra, Australia
felicia.zhang@canberra.edu.au

In an empirical study, Bryant & Zhang (forthcoming) show that the academic performance of a university subject (unit) could be understood through measuring *learning* results (operationalised from teaching outcomes). The present investigation demonstrates that assessment is important to learning results, because assessment involves selections that may have unintended effects on a unit. Understanding these effects might lead to an improvement in teaching and learning, which might in turn contribute to reshaping higher education. The assessment tasks considered were examination and essay tasks. These two tasks were frequently encountered in the preliminary pilot on feasibility. Also, significant weightings were attached to these tasks. To enable a comparison between different mixtures of assessment tasks, or different levels of weightings for each task, assessment tasks were classified according to their weighting's degree of dominance. This strategy allowed for a clear comparison between different task mixes. Essays were distinguished as generally beneficial to learning results whereas examinations were noted as generally non-beneficial to learning results. In summary, there was support for the thesis that the selection of assessment task and weighting can have an unintended impact on learning result, and a suggestion that superior selections can be made in task and weighting.

Keywords: student retention, assessment weightings

Introduction

Universities in the United Kingdom (UK), the United States of America (USA) and Australia are experiencing pressure to change their models of operation. These changes are discussed by Brennan & Teichler (2008) who see change in the UK being driven by the "growing societal importance of higher education" (p. 260). In Australia, Bradley, Noonan, Nugent, & Scales (2008) note that the "quality and performance of a nation's higher education system will be key determinants of its economic and social progress" (Executive Summary, p. xi). Bradley *et al.* (2008) propose changes to higher education through 46 recommendations. In part, their recommendation seven is built upon the "improvement of assessment practices" (p. 77). However, this recommendation makes no mention of the measurement of assessment practices. It might be concluded that the measurement of assessment practices is lacking. The present investigation is premised on the contention that the measurement of assessment practices is vital in understanding university experience, and that assessment practices are part of the:

> Comprehensive set of measures of the quality of teaching and learning [that] should be developed. These should include measures of the student experience and form part of a broader framework that is focused on the achievement of outcomes (Bradley *et al.*, 2008, p. 79).

It is the contention of this paper that a major part of academic 'student experience' happens within assessment at the unit level.

Literature review

In a study using questionnaire data from 25 departments in a large public Canadian university, research by Michelson (2004, p. 4) noted that "In-class tests, final examinations, and essays account, almost equally for three quarters of the marks" in 1012 courses in Arts and Science. In another study at the University of Southern Queensland involving 120 courses, Taylor (2006) found that examinations were used in 43 percent of Education courses; 47 percent of Arts courses; 85 percent of Science courses; 86 percent of Engineering courses; and 100 percent of Business courses. Examination weightings for Business, Engineering and Science were very high, according to Taylor, but low (usually 30 percent or less) for Arts.

Taylor (2006) reported that "if the number of assignments separate from examinations were examined … forty-five percent of courses had one or two assignments, thirty-three percent had three assignments, eleven percent had four and eleven percent had more than four assignments" (p. 3). These figures indicate that four or more assessment instruments were used in a majority of courses. Another instance of variety can be seen in an Australian Universities Teaching Committee (2003) project, entitled 'Teaching large classes project 2001', which noted that 54 of 64 (or 84 percent) of the respondents from the 24 Australian universities that took part in the AUTC project self-reported that they used a range of three or more assessment items (p. 39). Range was interpreted to mean variety, and not total count of assessment tasks. This means that four assessment tasks was possibly the norm.

On the topic of student satisfaction with assessment tasks, a Canadian study, using survey data, found an inverse (or negative) relationship (Pearson r = – .33) between satisfaction with final examinations but found a positive relationship (r = + .28) between satisfaction with essays (Michelson, 2004). In other words, students expressed decreasing satisfaction levels as final examination weightings rose, but expressed rising satisfaction levels as essay weightings increased. Michelson (2004) reported that student satisfaction was significantly and positively "related to the number of types of assessment chosen (r = + .12)" (p. 13).

On the importance of assessment, Biggs (1996) maintains that "learning is driven by assessment" (p. 10). Biggs also claimed that the quickest way to change student learning is to change the assessment system. Biggs' comments support measuring a unit's learning result.

Research questions

1. Is it possible to demonstrate that a unit's selection of assessment tasks, weightings and the number of tasks can have an unintended impact on learning results?
2. To what degree, if any, might a comparison be made between unit teaching outcomes (existing knowledge used by universities and built from unit satisfaction responses) and unit learning results (knowledge built from unit grades)?

Methodology

Unit outlines were used as the primary source of data, as they contained a description of the assessment method for each unit. All 63 semester one, 2008, outlines were collected from the print holdings of a regional university where this research took place. This constituted a convenience sample (Moore, 2007). Although there are in excess of 800 units, only 63 outlines were available at the time of the research. Recorded from each outline were the following data: whether there were formal written examinations and/or academic essays; the weighting of each task; and the number of tasks. Outlines detailed summative assessment only (Yorke, 2001). The examination task was described in outlines as timed, formal in nature, and included multiple choice or open-ended questions. Essays were described as academic in nature, with a prescriptive emphasis towards two thousand or greater word lengths, and bibliographies were required with appropriate citing of the literature. Reflective essays were not included as their emphasis on reflecting personal experience lacked academic rigour, requiring limited or no reference to the literature.

Learning results for each unit were calculated in a variable referred to as unit grade average or UGA (Bryant & Zhang, forthcoming). While UGA uses grade point average (GPA) as a model, it has major departures in calculation method and philosophy. GPA is a combined score for one student across units, and represents the learning result for one student. UGA is a combined score for one unit across students, and represents the learning result for one unit. The calculation departure lies in the values assigned to grades (Bryant & Zhang, forthcoming). In calculating UGA, a negative point is assigned for each unit fail; one point is assigned for a pass; two for a credit; three for a distinction; and four for a high distinction.

Table 1 shows a fictitious unit with fifteen students with learning outcomes of five fails (5*–1= –5); with two passes (2*1=2); with three credits (3*2=6); with three distinctions (3*3=9); and two high distinctions (2*4=8). The UGA is the sum of all outcomes (20) divided by the total number of students (15), which is a learning result of 1.33. A UGA of 1.00 or more represents a pass learning result for the unit; a UGA of 2.00 or more represents a credit learning result; UGA of 3.00 or more a distinction learning result; and UGA of 4.00 is a high distinction learning result. In Table 1, the UGA of 1.33 represents a pass result for the unit, and does so more clearly than a collection of grade descriptors.

Table 1: Calculating learning result (UGA)

Fail	Pass	Credit	Distinction	High Distinction	Total	UGA
5	2	3	3	2	15 students	
–5	2	6	9	8	20 points	1.33

It may be surprising that each failure grade is assigned a value of minus one by Bryant and Zhang (forthcoming), as this is starkly different from GPA scoring which allocates positive points for failure. However, if the learning goal is to not fail (that is, to succeed), then instances of failure have importance to the goal, and that importance is reflected by a negative assignment in the UGA calculation. Using each unit's learning result, faculty UGAs were calculated.

A set of weighting ranges were designed to represent the degree of dominance that a weighting is likely to have exerted. The ranges were:

- zero dominance, meaning that the task (such as examination) was not used and could not have any dominance;
- mildly dominant, where the task was indeed used in the unit assessment but the weighting was up to 45 percent, that is, below a dominant 50 percent;
- equally dominant, where the task (say, examinations) was used in the unit's assessment with a weighting of 50 percent but was not entirely dominant because another 50 percent weighting (say, for essays) also existed;
- dominant, where the task was used in the unit's assessment with a dominant weighting of 50 percent but at least two other weightings adding to 50 percent also existed, neither of which could be considered dominant; and
- very dominant, where the weighting was at or above 55 percent, making any other weightings not dominant.

Next, each unit in a faculty was grouped into a dominance range according to its examination weighting. This allowed calculation of learning result by faculty by weighting range. The process was then repeated to determine the learning results for essays by faculty by range.

The dominance groupings provided the basis for a clear comparison of faculty learning results by assessment tasks. This combination overcame one of the enduring difficulties of managing a comparison where there are mixes of assessment tasks in varying proportions. For example, this technique allowed a comparison of units with a mix of 80 percent examination and 20 percent presentation against units with 40 percent examination, 30 percent essay and 30 percent presentation. The final piece of information obtained from unit outlines was the number of assessment tasks that were required to fulfil the learning outcome.

Regarding research question two, unit satisfaction survey summaries of the question "Overall I was satisfied with this unit" were obtained and entered into SPSS to enable a correlation to be tested between unit teaching outcome (USS) and unit learning result (UGA). Slightly more than half of the 63 units did not have entries in the summaries, which is due to those units having less than ten survey respondents. In the absence of a unit reading, the faculty average was substituted, and in its absence, the question mean was used.

Examination findings and discussion

Figure 1 portrays the learning result for each faculty as the examination task increases in dominance, moving through the dominance groups used by each faculty. It should be noted that not all faculties used all five possible dominance groups. For example, Information Science units used only one examination weighting (very dominant) and because they have all units in one range, did not appear in Figure 1. However, any faculty which had units using two or more ranges did appear. For example, the Law faculty had units using zero dominance and the very dominant ranges.

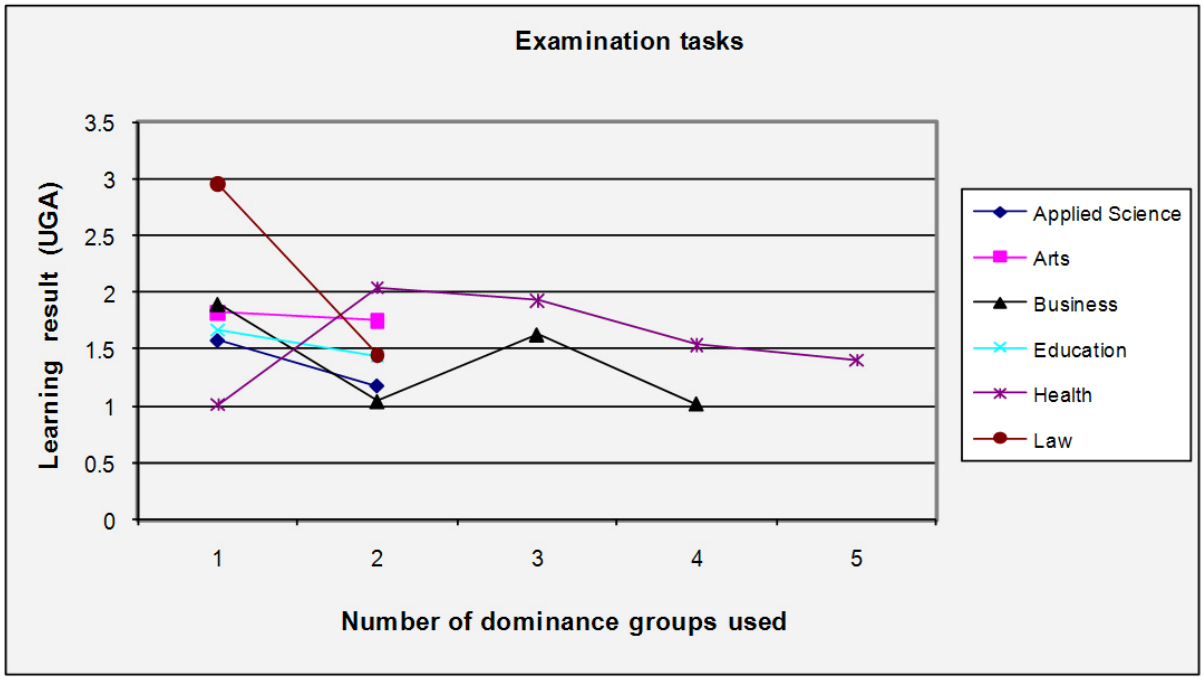

Figure 1: Learning result by faculty and weighting for examination tasks

The significance of Figure 1 is that it shows a steady decline in learning results in four faculties (Applied Science, Arts, Education and Law) as examination weighting increased. Further, it showed a fluctuating, but overall decline, for Business results. These declines suggest that rising examination weightings are correlated to learning result declines.
The only faculty to resist the downward trend was the Health faculty which showed an initial surge in learning result, followed by a steady decline. Upon investigation of Health's atypicality, one of the two Health units with zero dominance was found to be competency based, awarding only pass and fail grades. The absence of credits, distinctions and high distinctions meant that its learning results reported at an unrepresentative low level. Had a typical system been used for grading, Health results would probably have exhibited the same downward trend seen for the other faculties.

Table 2 is the numerical equivalent of Figure 1. Table 2 lists learning results by weighting range for examinations in the 63 units spanning seven university faculties (column one). The significance of Table 2 is that it suggests that the selection of examinations by faculties as an assessment task has a negative impact on faculty learning results; and by association, on unit results. In other words, Table 2 provides evidence that the selection of examination task can impact on learning results, in this case, non-beneficially.

While an examination trend can be seen, it is known that trends can be caused by chance in the dataset. To address this point, a linear regression analysis was undertaken using learning result and examination weighting, and confirmed that examination weighting does have an inverse relationship to learning result, and was significant ($p = .001$).

Table 2: Learning result by faculty and weighting for examination tasks

Faculty name	Zero dominance	Mildly dominant	Equally dominant	Dominant	Very dominant	N (units)	Faculty UGA
Applied Science	1.58				1.17	4	1.27
Arts	1.82	1.75				17	1.81
Business		1.90	1.04	1.63	1.02	16	1.23
Education	1.67	1.44				5	1.63
Health	1.01	2.04	1.93	1.54	1.40	12	1.68
Info Science					1.48	2	1.48
Law	2.95				1.45	7	1.88
Average UGA	1.81	1.87	1.22	1.59	1.24		1.59
N (units)	22	11	5	7	18	63	

Further checking of the findings discussed revealed some interesting points. In the data sample, there were students from six academic year levels: first year to fourth year, graduate and post-graduate level. The falling trend was seen statistically at all levels, but learning result rose slightly as year levels rose. This rise might be expected as students become more conversant with the university discourse and seemed to give credence, in a small way, to the overall findings. When checks were done on size, they seemed to show that size is significant but size did not appear to change the overall trend.

Essay findings and discussion

Figure 2 portrays the learning result for each faculty as the essay task increased in dominance. It should be noted that not all faculties used all the five possible dominance groups. For example, Information Science units used only one essay weighting (zero dominant) and for because they have all units in one range, did not show a trend in Figure 2. However, any faculty that had units using two or more ranges did appear. For example, the Law faculty had units using mildly dominant range and very dominant range.

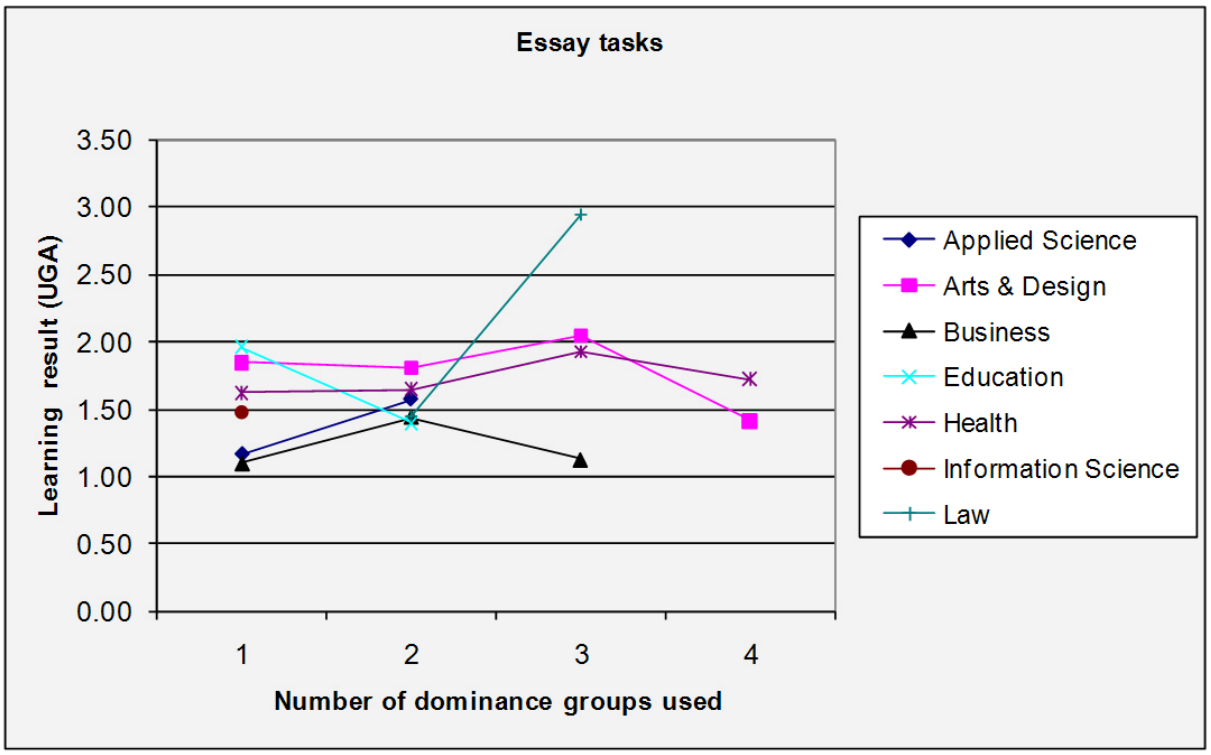

Figure 2: Learning result by faculty and weighting for essay tasks

The significance of Figure 2 is that it shows a propensity for learning result to increase as essay weighting increased, although there were riders to this interpretation. A clear rise can be seen for both Law, which rose from a mid level pass and almost reached a distinction result, and for Applied Science, which rose from a low level pass to a strong pass result. On the other hand, a clear fall was observable for Education, which dropped from almost a credit to a mid level pass. Fluctuations can be seen for Arts, Health and Business, which all had initial rises followed by declines. This pattern was suggestive of a decline in learning result as essay weighting moved towards higher dominance ranges. Another possibility is that very high dominance of a task precluded other tasks, thereby reducing the number of tasks and the chance to recover from a poor performance on any one task.

Table 3: Learning result by faculty and weighting for essay tasks

Faculty name	Zero dominance	Mildly dominant	Equally dominant	Dominant	Very dominant	N (units)	Faculty UGA
Applied Science	1.17	1.58				4	1.27
Arts	1.85	1.81	2.05		1.42	17	1.81
Business	1.10	1.44	1.13			16	1.23
Education	1.96	1.40				5	1.63
Health	1.62	1.65	1.93		1.72	12	1.68
Info Science	1.48					2	1.48
Law		1.45			2.95	7	1.88
Average UGA	1.55	1.53	1.47		2.03		1.59
N (units)	28	24	5	0	6	63	

Table 3 is the numerical equivalent of Figure 2. Table 3 lists learning results by weighting range for examinations in seven university faculties (column one). The significance of Table 3 is that it suggests, with reservations, that the selection of essays as an assessment task had a positive impact on faculty learning results; and by association on unit learning results. In other words, Table 3 provides evidence that it was possible to show that the selection of essay task can impact on learning results beneficially.

As Table 3 indicates, slightly under half of the sample did not use essays (zero dominance in column one shows that 28 units chose not to use essays). In other words, slightly over half of the units used essays, and this compares to Table 2 where two thirds of the sample used examinations. When essays were used, there were increases in learning results in a majority of cases. For example, Applied Science gained, moving from a pass (1.17 UGA) to a stronger pass (1.58 UGA), as did Law from a mid range pass result to almost a distinction result (1.45 to 2.95 UGA). However, a rising trend was not always seen, with Health moving from a mid range pass result to almost a credit result and back to a mid range pass result (1.62 to 1.93 to 1.72 UGA).

At the beginning of this project, essays were believed to have a similar degree of difficulty to examinations. Should that perspective change in the light of the contrasting findings of essays enabling learning result growth and examinations impeding learning result growth? Although a full discussion is outside the scope of this paper, we suggest briefly that this is not the case. Essays and examinations are likely to be different kinds of difficulty: one concerned with construction, planning and execution or acquiring the academic writing skill (Krause, 2001); the other with managing pressing time constraints, among other matters. Faced with the difficulty posed by either an essay or examination, a strategy of cheating or guessing (plagiarising, guessing multiple choice answers) may be employed to achieve a result. Both strategies might be conceding the difficulty.

Assessment task numbers

A second potential instance of institutional impact was a unit's interpretation of university policy which states that assessment tasks "should be several in number" according to key principle 5 (University of Canberra, 2007, p. 1) and "over-assessment should be avoided" according to principle 5.6 (University of Canberra, 2007, p. 12). The number of assessment tasks ranged from two to 15 in this sample. Four was taken from the literature as an ideal number of tasks to be administered (AUTC, 2003; Taylor, 2006) and an analysis was undertaken on how many units would have met this "ideal".

Table 4 shows that eleven of the 63 units used four tasks (column two), 38 used two or three tasks (column four), and 14 used five to 15 tasks (column six).

Table 4: Learning result by faculty and policy ideal (4 tasks) divergence

Faculty name	Ideal	UGA	Below ideal	UGA	Above ideal	UGA	Total
Applied Science	2	1.089	2	1.086			4
Arts	3	1.53	6	1.24	8	1.03	17
Business	2	0.99	11	0.62	3	0.94	16
Education			5	1.28			5
Health	4	1.20	8	0.77			12
Info Science					2	0.88	2
Law			6	1.07	1	1.24	7
	11		38		14		63

Key: Ideal = 4 assessment tasks; below policy < 4 tasks; above policy > 4 tasks.

Table 4 learning results (columns three, five and seven) might give support to the notion that four tasks is an ideal number. For example, Arts had a learning result of 1.53 when at the ideal (no. of items = 4); a learning result of 1.24 when below the ideal; but 1.03 when above the ideal. Health, Business and Applied Science also had their better results when at the ideal. While Table 4 suggested that institutional divergence from an ideal had learning result impacts, the pattern's persistence across task type was unclear. Table 5 shows findings when divergence from the "ideal" was contrasted against dominance range for examinations. The significance of Table 5 was that it showed the policy "ideal" of four assessment tasks remained valid across different assessment weightings, for examinations.

Table 5: Learning result by dominance range and policy for examinations

Dominance	Ideal	UGA	Below ideal	UGA	Above ideal	UGA	Total
Zero dominance	3	1.53	13	0.77	6	1.04	22
Mildly dominant	2	1.89	7	1.39	2	1.61	11
Equally dominant							0
Dominant	1	1.20	9	0.65	2	1.32	12
Very dominant	5	0.99	9	0.62	4	0.88	18
	11		38		14		63

Key: Ideal = 4 assessment tasks; below policy < 4 tasks; above policy > 4 tasks.

Table 5 learning results endorse the notion that four tasks might be ideal. Three of the four dominance ranges had their best results when using the ideal. For example, zero dominance range had a learning result of 1.53 when at the ideal (no. of items = 4); a learning result of 0.77 when below the ideal; but 1.04 when above the ideal. When the same calculation was undertaken for the essay type, again three of the four ranges had their best result when using the ideal number of tasks.

Teaching–learning relationship

There is a weak relationship between teaching outcome (USS) and learning result (UGA) (Pearson r = .171) but it is not statistically significant (p = .182). The safest conclusion to make is that the two measurements are not related. This might not be surprising since the two measures appear to have different geneses.

Teaching outcome was derived from a survey response of student perceptions whereas learning result was a restatement of the learning outcomes issued by a lecturer for tasks completed by students. Student participation in teaching surveys was voluntary and featured low participation rates, with the response rate for semester one 2008 varying from 16 to 25 percent; whereas participation in learning result was automatic and included the total number of students who completed a unit. Another difference was that teaching outcome was derived from self-reported questionnaires and this questionnaire design is known to have bias problems (Moore, 2007), whereas learning result was not self-reported. Yet another difference existed in the degree of skewing seen in the two measures. When teaching outcome was graphed it showed strong skewing to the upper end of the seven point scale and this is believed to be unrepresentative of a normal population (Moore, 2007). When learning result was graphed, it was more attuned to a normal curve which is an interesting observation since the university policy on assessment advocates criterion-referencing, which might be expected to exhibit upper-end skewing.

In the light of the absence of a relationship in our dataset between teaching outcome and learning result, a question arises: What is the basis for accepting and using teaching outcome as a measure for lecturer performance and unit achievement, given it has no correspondence to unit learning UGA? If there is no clear answer to that question, does it mean that the innovative learning result measure should be adopted as an important metric for unit achievement? Or does it mean that more research on unit achievement is needed?

Limitations

Limited data was made available to this research, but since the underlying questions were exciting and challenging, we proceeded on a ‘data shoestring’. While this approach weakened the certainty of the results, we hope that it did show the possibilities for unit achievement research. Data might be considered limited because there were only 63 units’ data available for analysis. When this was spread across seven faculties, it meant that observations were taken based on an average of nine units in a faculty. In fact, the distribution ranged from as little as two units in Information Science to 17 units in Arts.

Conclusion

It was possible to show that the selections of assessment task, weighting and number of assessment tasks might have unintended impacts on assessment results. In this regard, findings confirmed concerns found in the literature (Michelson, 2004). Some interesting observations came out of the investigation. Interpreting the number of assessment tasks was found to be chanceful with the policy word “several” being interpreted as meaning as low as two or as high as fifteen assessment tasks. The university might consider presenting greater clarity, based on research, in policy documents on assessment.

Secondly, some future research questions have been foreshadowed. During the research there was debate on the degree of implementation, if any, of the policy of mandatory criterion-referenced marking. One suggestion made was that various interpretations of criterion referencing might exist amongst teaching staff. Another suggestion was that norm referencing is entrenched and might be assumed because the official policy might not have been well circulated. University policy defined what appeared to be criterion referencing for learning outcomes, interspersed with a norm referencing definition, finally 'advocating' criterion referencing in key principle four (University of Canberra, 2007), when 'mandating' criterion referencing might have better clarified the policy direction.

Further research could investigate differences between university assessment policy and the interpretation of terms contained in the policy such as 'summative' and 'formative' assessment, as well as the meaning of "include different types of tasks", which is part of key principle 5 (University of Canberra, 2007). Further, the recent unexpected provision of quality data, through the university having now created an electronic repository exceeding 800 unit outlines, will facilitate a future replication study.

The final goal of comparing teaching outcome and learning result for a unit was only reached in the sense of opening a discussion on the topic – questions seem to exceed answers at this early stage. For example, if there is no relationship between learning results and surveyed satisfaction of teaching, are they reliable indicators of the teaching–learning process? An answer does not imply the dismissal of other indicators, such as graduate employability. In summary, this investigation's replicable methodology, and its generalisability to all universities, should be apparent.

Acknowledgements

Mr David Pederson for his quiet encouragement to use statistics to help solve problems.

References

AUTC. (2003). *Teaching large classes project 2001 final report*. St Lucia: The University of Queensland.

Biggs, J. (1996). Assessing learning quality: Reconciling institutional, staff and educational demands. *Assessment & Evaluation in Higher Education, 21*(1), 5.

Bradley, D., Noonan, P., Nugent, H., & Scales, W. (2008). *Review of Australian higher education*. Retrieved. from http://www.deewr.gov.au/he_review_finalreport

Bryant, D., & Zhang, F. (forthcoming). *Measuring unit level performance for first year units*. Paper to be presented at the First Year in Higher Education Conference 2010, Adelaide.

Krause, K. L. (2001). The university essay writing experience: A pathway for academic integration during transition. *Higher Education Research and Development, 20*(2), 147–168.

Michelson, W. (2004). *Enrolment matters: University course characteristics.* Paper presented at the American Sociological Association, San Francisco, CA.

Moore, D. S. (2007). *The basic practice of statistics* (4th ed.). New York: W. H. Freeman and Company.

Taylor, J. A. (2006). *Assessment: A tool for development and engagement in the first year of university study*. Queensland: Language & Teaching Support Unit, University of Southern Queensland.

University of Canberra (2007). *University of Canberra assessment policy*. Canberra: University of Canberra.

Yorke, M. (2001). Formative assessment and its relevance to retention. *Higher Education Research & Development, 20*(2), 115–126.