INTERTESTER RELIABILITY: REPORTING ON ASSESSMENT METHODS IN INTERDISCIPLINARY UNITS

DAVID NEWLYN AND LIESEL SPENCER[*]

The study which is the subject of this paper examines 'intertester reliability', with one of the stated aims to identify a marking method of optimum reliability so as to improve consistency between multiple markers. A brief survey of the literature suggests that criterion-referenced assessment implemented by means of a marking rubric is a superior form of managing formal assessment. The matters under investigation are confined to a comparison of the (intertester) reliability of marking results obtained using three different marking methods, a finding as to which method produces the greatest range of results, and a comparison of the time taken by markers using those methods. The results could, however, have secondary utility in considering the reliability of the use of marking rubrics as a means of implementing criterion-referenced assessment, particularly in large groups with multiple markers. The findings of this paper will show that there is no statistically significant ($p<0.05$) difference in reliability of results, range of results or time taken when utilising any of the 3 methods employed.

I. INTRODUCTION — WHY BE CONCERNED ABOUT RELIABILITY OF MARKING ASSESSMENTS?

Formal assessment is an integral part of higher education.[1] There is an abundance of literature on the purpose and design of assessment tasks and the integration and compatibility of those tasks with broader aims (such as graduate attributes and learning outcomes of the tertiary institution as a whole, the relevant faculty or school, and the individual student).[2] The reliability of the marking process and resulting scores or grades, however, receives comparatively sparse coverage from researchers.[3]

It is of increasing importance to tertiary institutions to optimise the accuracy and integrity of assessment. Reputation is a factor in a prospective student's choice between competing institutions.[4] Once a student has selected an institution, they incur financial liability for study, with failed subjects incurring repeated liability. There has been a consequent shift in the institution/student relationship whereby the student body is now more 'customer-oriented', viewing themselves as consumers of the tertiary product,[5] investing time and money and expecting a return on that investment.

Determining the most reliable, defensible marking method is therefore both an advantage over market competitors and a risk management tool. This knowledge can benefit a tertiary institution in managing the student appeals process, justifying the outcome of that process to upper stratum of review in the institution and, in a most extreme

* Dr David Newlyn (Lecturer) and Liesel Spencer (Associate Lecturer) are both members of the School of Law, University of Western Sydney.

1 'Few would argue with the assertion that assessment lies at the centre of the student experience and is a dominant influence on student learning': Barry O'Donovan, Margaret Price and Chris Rust, 'Know What I Mean? Enhancing Student Understanding of Assessment Standards and Criteria' (2004) 9(3) *Teaching in Higher Education* 325.

2 See, eg, John Biggs, *Teaching for Quality Learning at University* (2nd ed, 2003) 156, regarding 'authentic or performance assessment'; Sally Brown and Peter Knight, *Assessing Learners in Higher Education* (1994) ch 3; George Brown, Joanna Bull and Malcolm Pendlebury, *Assessing Student Learning in Higher Education* (1997) 10.

3 Erica Smith and Kennece Coombe, 'Quality and Qualms in the Marking of University Assignments by Sessional Staff' (2006) 51 *Higher Education* 45, 45; Christopher Dracup, 'The Reliability of Marking on a Psychology Degree' (1997) 88 *British Journal of Psychology* 691, 692.

4 Anna Reid, Vijaya Nagarajan and Emma Dortins, 'The Experience of Becoming a Legal Professional' (2006) 25(1) *Higher Education Research & Development* 85, 97.

5 Clair Hughes and Clare Cappa, 'Developing Generic Criteria and Standards for Assessment in Law: Processes and (By) Products' (2007) 32(4) *Assessment and Evaluation in Higher Education* 417, 418; Mantz Yorke, 'The Management of Assessment in Higher Education' (1998) 23(2) *Assessment and Evaluation in Higher Education* 101, 101; Smith, above n 3, 45; Reid, above n 4, 89-90

case, defending litigation by aggrieved students.[6] Sue Bloxham states bluntly that her 'hunch is that students will become increasingly litigious about marking in the future and our procedures will struggle to stand up to this kind of onslaught if we persist in claiming that marks given are completely accurate'.[7] Hilary Astor confirms that students are 'increasingly taking their universities to court', with the number of student litigation cases in Australia having 'escalated since the mid-1990s'.[8] Whilst there is authority to the effect that courts will intervene in 'disciplinary issues' but will not intervene in 'questions of academic assessment and judgment'[9], this line is not always clear and resulting litigation can, in any event, 'cost universities millions of dollars in legal fees'[10].

Continuing the analogy of the university as a quasi-corporate supplier of product to students as consumers, more effective evaluation of the product is facilitated by using a more reliable marking method. Assessment provides 'evidence regarding the success or otherwise of the programme'.[11]

Broader ethical and integrity issues of equality and fairness are also supported by a marking method with superior reliability — students are more likely to be graded only on their own performance in the piece of work being assessed, without additional reference to their prior academic performance or to other students' performance in the item being marked.[12]

Managing large groups of students with multiple academic staff responsible for marking presents specific challenges.[13] Individual markers need to ensure consistency in marking on different days and across large numbers of assessments,[14] whilst consistency between markers must also be achieved.[15] Mark Saunders and Susan Davis cite 'pressures caused by high student numbers and tight marking deadlines' as a particular challenge to reliability of marking standards.[16] A corollary of assessing large student groups and carrying substantial marking workloads is that a marking method which saves time without compromising reliability has a practical advantage.[17] The increased use of casual or sessional academic staff to carry out marking — the 'outsourcing' of marking work — presents further challenges to achieving consistent quality across a spectrum of markers.[18] Saunders and Davis note that staff changes over time are especially 'likely to alter' the 'understanding of criteria and consistency' between markers.[19]

6 Yorke, above n 5, 101; Harvey Woolf, 'Assessment Criteria: Reflections on Current Practices' (2004) 29(4) *Assessment and Evaluation in Higher Education* 479, 490.

7 Sue Bloxham is quoted in Rebecca Attwood, '"Inherently Frail" — the Verdict on Marking' *Times Higher Education Supplement* (United Kingdom), 26 October 2007 <http://www.timeshighereducation.co.uk/story.asp?storyCode=310884&sectioncode=26>at 5 December 2008. Bloxham is co-author, with Peter Boyd, of *Developing Effective Assessment in Higher Education* (2007).

8 Professor Hilary Astor of the University of Sydney Faculty of Law is quoted in Harriet Alexander, 'Student Fees Spark Rush of Grade Disputes', *The Sydney Morning Herald* (Sydney), 15-16 March 2008.

9 *Walsh v University of Technology, Sydney* [2007] FCA 880 [77] (Buchanan J) applying *Griffith University v Tang* (2005) 221 CLR 99.

10 Alexander, above n 8.

11 Yorke, above n 5, 105.

12 D Royce Sadler, 'Interpretations of Criteria-based Assessment and Grading in Higher Education' (2005) 30(2) *Assessment and Evaluation in Higher Education* 175, 178.

13 Richard James, Craig McInnis and Marcia Devlin, *Assessing Learning in Australian Universities* (2002) 31; Mark Saunders and Susan Davis 'The Use of Assessment Criteria to Ensure Consistency of Marking: Some Implications for Good Practice' (1998) 6 *Quality Assurance in Education* 162, 162.

14 Sadler, above n 12, 191.

15 Effie Maclellan, 'Authenticity in Assessment Tasks: A Heuristic Exploration of Academics' Perceptions' (2004) 23(1) *Higher Education Research and Development* 19, 25.

16 Saunders and Davis, above n 13, 166.

17 Raija Kuisma, 'Criteria Referenced Marking of Written Assignments' (1999) 24(1) *Assessment and Evaluation in Higher Education* 27, 27.

18 Smith and Coombe, above n 3, 50. More generally, on the casualisation of the workforce, see Sally Kift, 'Assuring Quality in the Casualisation of Teaching, Learning and Assessment: Towards Best Practice for the First Year Experience' (Paper presented at the 6th Pacific Rim First Year in Higher Education Conference 2002: Changing Agendas – Te Ao Hurihuri, University of Canterbury, Christchurch, New Zealand, 8 – 10 July 2002) <http://ultibase.rmit.edu.au/Articles/march03/kift1.htm> at 5 December 2008.

19 Saunders and Davis, above n 13, 4. See also Tony Taylor et al, ' A Bleak Outlook: Academic Staff Perceptions of Changes in Core Activities in Australian Higher Education, 1991-96' (1998) 23 *Studies in Higher Education* 255.

Integrity of assessment in the humanities, including law, is complicated by assessment tasks more likely to consist of 'open questions' which are more difficult to assess in a fair, valid and reliable manner.[20] The contemporary trend towards 'modular schemes', whereby students are able to construct programs and select electives across disciplines, provides added impetus to ensure equity by maximising the integrity of assessment methods.[21]

The range of results obtained by different marking methods is relevant to the integrity of those methods.[22] Where a method produces marks which are concentrated or clumped in one band, there is a failure to adequately discriminate between students, and 'the good students are not credited for their achievements' whilst 'the poorer students are not made aware strongly enough that improvement is needed'.[23] Discriminating between students' work which lies within a concentrated middle band is more difficult for markers than allocating marks to papers which lie at extreme ends of the spectrum of marks.[24] It is desirable, then, that a marking method produce both a relatively greater range of results and a valid means of discriminating between items of student work which are close in quality.

Which method of marking assessment tasks, then, achieves optimum integrity and reliability? Effie Maclellan conducted a qualitative study involving 'in-depth interviews with twelve academics' seeking their views on 'what might constitute desirable assessment tasks and scoring methods'.[25] The former concern is beyond the scope of this paper. However, the responses to the latter issue are of interest. All respondents were in favour of 'explicit assessment criteria' being accessible by both markers and students when the assessment task was issued.[26] Respondents nominated 'the provision of clear marking rubrics' as a means to achieve consistency between markers.[27] Maclellan observes that the judgments involved in assessing student performance are complex, and that 'opinion was divided on the extent to which criteria enables the judgment'.[28]

The references to criteria by the academics surveyed in Maclellan's study reflect a trend towards, and support in literature for, criteria-based or criterion-referenced assessment.[29]

II. RESEARCH PROJECTS IN THIS AREA

Raija Kuisma conducted an 'action research project' to compare the results of lecturers' marking with and without the use of a criterion-referenced marking form, with 'formally espoused criteria'.[30] The methodology involved four lecturers marking separate bundles of assignments by using first their 'own criteria through the exercise of their normal judgment which was their normal practice'[31] and then marking the same assignment bundles using the criterion-referenced marking form. The methodology selected meant that marking between the lecturers — 'intertester reliability' — could not be compared. What could be analysed was 'intratester reliability'. The results demonstrated a greater range of marks

20 Kuisma, above n 17, 2.
21 Yorke, above n 5, 104. Yorke notes that a trend to offer students modular schemes, where a degree is constructed by students' selection of subjects across faculties, raises issues of equity, citing a prior study (Mantz Yorke et al, 'Module Mark Distributions in Eight Subject Areas and Some Issues They Raise' in Norman Jackson (ed), *Modular Higher Education in the UK in Focus* (1996) 105-107) wherein 'for both means and standard deviations of assessments in eight disciplinary areas, there were persistent differences that transcended institutional boundaries'.
22 'Reliability requires not only that an individual mark or score or grade is accurate, but also that this mark or score or grade bears the appropriate relationship to any other in the same set of scores': Chartered Institute of Educational Assessors, *Assessment Reliability* (2007) <http://www.ioea.org.uk/knowledge_centre/articles_speeches/general_articles/assessment_reliability.aspx> at 5 December 2008.
23 Kuisma, above n 17, 33.
24 Catherine Haines, *Assessing Students' Written Work* (2004) 38.
25 Maclellan, above n 15, 19.
26 Ibid 25.
27 Ibid.
28 Ibid 27.
29 Sadler, above n 12, 176, 178.
30 Kuisma, above n 17, 27.
31 Ibid 29.

using the marking form, and a 'distribution of marks closer to the normal distribution' when using the marking form.[32] George Brown, Joanna Bull and Malcolm Pendlebury[33] note that 'the two main measures of reliability in assessment are measures of agreement between assessors and within assessors' or intertester and intratester reliability.

The study which is the subject of this paper examines 'intertester reliability', with one of the stated aims to identify a marking method of optimum reliability so as to improve consistency between multiple markers.

Christopher Dracup[34] conducted a study of intertester reliability between two academics double-marking the same set of papers without reference to an espoused marking criteria in order to determine whether double-marking in the subject and course studied could be justified by a significant compensation for marker unreliability. This study compares results using three different marking methods.

Saunders and Davis[35] collected data from two workshops run to address concerns that marking criteria and procedures (for undergraduate dissertations at their institution) required revision, and that the consistency of application of marking criteria by multiple markers required examination. The results of their research indicated that the revision of existing criteria, and the beneficial effects of collegial discussion and marker training, improved reliability. These are issues outside the scope of this paper. However, they would be appropriate subjects for future studies.

III. TERMINOLOGY IN ASSESSMENT LITERATURE

Based on an analysis of the literature, distinctions need to be drawn between the different terminology used in this area.

First, the concepts of 'assessment criteria', 'criterion-referenced assessment' and 'criteria-based assessment' must be distinguished from the terms 'marking guide', 'marking form' and 'marking rubric'. The former terms refer to the design and purpose of the assessment task itself, whereas the latter terms refer to a marking tool used to actually score or grade the student on their performance in the task. In formulating a marking guide, form or rubric reference is logically made to the assessment criteria. The marking guide designed for use in this study was explicitly based on assessment criteria (based on desired outcomes) made available to students in the form of assignment instructions. As a side observation, marking forms are of utility as a form of feedback to students — 'the matrix itself has considerable diagnostic value for the learner'.[36]

The second distinction between concepts which arises is that between 'criterion' and 'standards'.[37] Clair Hughes and Clare Cappa report on the process of developing 'a set of generic assessment criteria and standards, or rubric, which could be customised to the requirements of individual law subjects'.[38] They define criteria as pertaining to 'qualities of interest and utility', whereas standards are about 'a definite level of achievement … definite levels of quality'.[39]

D Royce Sadler reviews four grading models utilised by four different universities, each of which described their models as 'criteria-based': achievement of course objectives; overall achievement as measured by score totals; grades reflecting patterns of achievement; and specified qualitative criteria or attributes.[40] An advocate of the last approach, Sadler nonetheless argues that in order to judge student work 'on an absolute rather than a relative scale', and to 'realise on the aspirations for criteria-based grading', a shift is required from

32 Ibid 27.
33 Brown, Bull and Pendlebury, above n 2, 234.
34 Dracup, above n 3, 691-708.
35 Saunders and Davis, above n 13.
36 Sadler, above n 12, 185.
37 Hughes and Cappa, above n 5, 417; Sadler, above n 12, 175, 193.
38 Hughes and Cappa, above n 5, 417.
39 Ibid 418.
40 Sadler, above n 12, 179, 181, 183, 184.

criteria-based assessment to standards-based assessment. The practical implementation of this shift as it relates to marking methodology should include 'fixed reference levels of attainment' (standards) of 'attributes or properties' (criteria) measured by 'statements setting down the properties that characterise something of the designated levels of quality' together with 'exemplars' representative of various standards or grades.[41]

IV. MARKING GUIDES/CRITERIA/RUBRICS

The practical guide issued by the University of Melbourne's Centre for the Study of Higher Education recommends specific strategies for the management of volumes of marking generated by large groups of students. These recommendations include the provision of consistent criteria to markers (marking guides), the use of exemplars, and the use of a standardised feedback sheet.[42] Brown, Bull and Pendlebury state that 'specific, but manageable, criteria or marking schemes increase reliability', referring to intertester reliability, but noting that criteria or marking schemes are also important in improving the consistency of an individual marker (intratester reliability).[43] The Institute of Educational Assessors recommends 'the use of straightforward, unambiguous mark schemes which can be interpreted consistently by all examiners' as a means of ensuring 'inter-marker reliability' (intertester reliability).[44] Saunders and Davis affirm that 'what is clear from other research … and emphasised by our experience, is that criteria which are designed carefully and used with clear procedures can reduce inconsistency in assessment'.[45] Dracup, referring to the utility of his 1997 study, cautions that whilst marking criteria can reduce 'marker unreliability … reliable marking does not guarantee valid assessment'.[46] Dracup illustrates this point with a hypothetical scenario wherein multiple markers base the scores awarded on an estimate of the word count of the assessments submitted — reliability (correlation between the scores awarded to the same paper by different markers) would be high whilst validity would be nonexistent.[47] Wiggins, discussing the concept of 'authentic' assessment, addresses this topic, stating that marking criteria must be 'appropriate' as well as standardised in order to ensure both validity and reliability.[48] This paper does not seek to address the issue of the validity of the assessment task used as the subject of the research.[49]

In addition to the use of marking guides (variously referred to as rubrics, criteria, or schemes), the use of exemplars was a recurring theme in various authors' recommendations for improving the reliability of marking.[50] This study did not employ exemplars as a tool to compare the relative reliability of different marking methods. This is an area which could be investigated by a follow-up study.

Sadler[51] (as noted above) identifies four 'models' which various tertiary institutions claim 'denote criteria-based assessment or grading'. The model he appears to advocate is based on 'specified qualitative criteria or attributes', implemented by marking guides, with criteria 'elaborated into a marking grid' in a number of forms.[52] These marking grid/guide formats can include: 'a simple numerical rating scale for each criterion' (with marks tallied overall); 'a simple verbal scale for each criterion'; or a verbal scale 'expanded into verbal

41 Sadler, above n 12, 190, 192.
42 James, McInnis and Devlin, above n 13, 35.
43 Brown, Bull and Pendlebury, above n 2, 234-235.
44 Chartered Institute of Educational Assessors, above n 22.
45 Saunders and Davis, above n 13, 4.
46 Dracup, above n 3, 707.
47 Ibid.
48 Grant Wiggins, 'The Case for Authentic Assessment' (1990) 2(2) *Practical Assessment, Research and Evaluation*, 2.
49 'My subject is not the philosophy of examination, but the statistics of marks': Francis Y Edgeworth, 'The Element of Chance in Competitive Examinations' (1890) 53(3) *Journal of the Royal Statistical Society* 460, 461.
50 Catherine Taylor, 'Assessment for Measurement or Standards: The Peril and Promise of Large-scale Assessment Reform' (1994) 31(2) *American Educational Research Journal* 231, 243; James, McInnis and Devlin, above n 13, 35; O'Donovan, Price and Rust, above n 1, 332.
51 Sadler, above n 12, 176.
52 Sadler, above n 12, 184.

statements that indicate different degrees on each criterion', in which case the marker may simply 'eyeball the matrix to assign an overall grade'.[53] These criteria and results can be distributed to students on 'grading criteria sheets' or as 'scoring rubrics' at the time an assessment task is handed out to students.[54] A less explicit format noted by Sadler is a list of 'verbal grade descriptions' for each grade level. The generalised use of the terms 'marking guide, rubric, criteria or scheme', is therefore not indicative of a consistent or homogenous entity.[55] The marking guide employed in this study is annexed as an example.

V. MATTERS UNDER INVESTIGATION IN THIS STUDY — CURRENT AND FUTURE UTILITY

This survey of the literature suggests that criterion-referenced assessment implemented by means of a marking rubric is a superior form of managing formal assessment. Arguably, the trend of opinion is that criterion-referenced assessment implemented by means of a marking rubric, based on standards as opposed to criteria, is the ideal, or as close to an ideal as is possible given the inherent subjectivity of human judgment.[56] This study does not attempt to address this distinction between criteria and standards, nor to investigate the concept of criterion-referenced assessment. The matters under investigation are confined to a comparison of the (intertester) reliability of marking results obtained using three different marking methods, a finding as to which method produces the greatest range of results, and a comparison of the time taken by markers using those methods.

The results could, however, have secondary utility in considering the reliability of the use of marking rubrics as a means of implementing criterion-referenced assessment, particularly in large groups with multiple markers.

Follow-up studies would be required to examine other aspects of this area. Intratester reliability using the three marking methods could be explored by having 10 different markers mark a random selection of 10 of the 30 assignments, with each person marking the same 10 assignments using the three marking methods employed in this study. A statistically significant difference between results obtained by the same marker using different methods to repeatedly mark the same assignments would indicate levels of intratester reliability for each method.[57]

A hypothesis can be extrapolated from the distinction drawn between criteria and standards that a marking form constructed by reference to standards, rather than merely by reference to criteria, would yield a greater range of marks and a finer tool to discriminate between work of less differentiated quality in the central or average band of marks.

VI. INTRODUCTION TO BUSINESS LAW (IBL)

Introduction to Business Law (IBL) is a first-year introductory law subject for students not enrolled in a law degree (non-lawyers)[58], which has been offered at the University of Western Sydney (UWS) since 2001. The primary focus of IBL has been on providing non-lawyers with the necessary skills/information for them to recognise legal problems when they may arise in their chosen future careers.[59] Historically, the students in IBL have been

53 Ibid 184-185.
54 Ibid 185.
55 Ibid.
56 Paul Newton and Chris Whetton, 'The Effectiveness of Systems for Appealing Against Marking Error' (2005) 31(2) *Oxford Review of Education* 273, 273; Maclellan, above n 15, 28; Sadler, above n 12, 190.
57 Kuisma, above n 17, 28.
58 The current stated aims for the unit as listed in the unit outline are: 'This is an introductory law unit designed to introduce the fundamentals of law in a commercial context. The unit introduces students to the basic principles of law and the legal system as well as examining some of the major areas of law that impact on commercial dealings. This unit examines the structure of the legal system, the way law is made, legal reasoning and problem solving. The main areas of law covered include contracts, torts, consumer protection and agency.'.
59 Further significant details about the history and structure of Introduction to Business Law can be found in Susan Fitzpatrick, 'Making IBL Relevant to Gen Y', (Paper presented at the 62nd Australian Law Teachers Association (ALTA) Conference, University of Western Australia, Perth, Western Australia, 23–26 September 2007).

from a Commerce or Accountancy program, but there are also a significant number of students from a diverse range of degrees/majors including tourism, hospitality, management and engineering.

IBL is usually offered during both of the main semesters (Autumn and Spring) of the university year and regularly attracts in excess of 1200 students.[60] For formal assessment purposes, students in IBL have traditionally undertaken three items of assessment: two 'take-home' assignments and a formal final examination.[61] This research project will examine the operation of the first of those take-home assignments.

The first assignment (see attachment 1) that students undertook in the Spring 2007 session of IBL was a brief research essay which focused on the area of Alternative Dispute Resolution (ADR). It was 1200 words in length and was due in the fifth week of the semester.

VII. PROJECT METHODOLOGY

The research team collected 30 assignments randomly from the approximately 1200 assignments which were submitted by the students. Each of those assignments was electronically scanned and stored. The electronically scanned versions of the assignment contained no identification of an individual student. The original assignments were then returned to the person who was tasked to mark them for the purposes of the IBL unit. Students had no knowledge of whether their paper was selected for analysis and in no way did this research project affect or attempt to alter the marks which students were ultimately awarded for their respective assignments.

With a view to examining the integrity of the marks which the student is ultimately awarded, each of these 30 assignments would be assessed in three different ways by different markers. A high school located within the catchment region of UWS was identified and agreed to participate in this project. A high school was chosen as it provided ready access to a large number of people who could, in a relatively short period of time, perform the tasks which will be outlined below. The researchers were not of the belief that it would be reasonably possible to organise to have the assessment of the assignments undertaken by legal academics in tertiary institutions in a reasonable period of time, in sufficient numbers to be statistically valid, particularly without piloting the project first. Therefore the researchers also envisaged that by undertaking this research task in a high-school environment, the possibility to pilot the methodology for consideration for use in a much larger study involving tertiary institutions throughout New South Wales would be extremely valuable.

High-school teachers in New South Wales also have experience in using standards-referenced assessment techniques which are used widely in the marking of assessment items for the school-based assessment component of the Higher School Certificate grade, and when participating in the marking of High School Certificate examinations. This experience, as professional markers, positions this group as 'experts' in this type of assessment methodology and means that they are highly consistent and reliable markers.[62]

The authors would assert that it could be argued that, as professional teachers, these participants do have a similar skill level to academics who may have taught a similar type of course.[63] The content of the assignment chosen certainly was not so technical or specific

60 For spring session 2008, the current prediction is for more than 2000 students. The unit is normally also offered during the summer session at the university and often attracts more than 120 students at that time.

61 During the semester in which this research project operated, the first take-home assignment constituted 15% of the final mark, the second assignment 25%, and the final exam constituted 60%. Note that in 2008, a new direction for assessment purposes began. Instead of two assignments and a formal exam, students were provided with an online multiple choice test (20%), an assignment (20%) and a formal examination (60%).

62 For further discussion of standards-referenced frameworks and assessment, see Board of Studies, New South Wales, *HSC Assessment in a Standards-Referenced Framework — Guide to Best Practice* <http://www.boardofstudies.nsw.edu.au/hsc_assessment_policies/#hsc_assess_framework> at 5 December 2008.

63 The authors acknowledge that any attempt to resort to literature for a concrete definition of 'professional' or 'profession' is met with severe limitations, but offer further justification and discussion on this point via an

that a non-lawyer would have difficulty with it. No attempt to provide an analysis of the demographics/skill levels of those 30 teachers is provided within the scope of this project.

Thirty teachers from the identified high school volunteered to be part of the project. Those 30 teachers were divided into three equal groups of 10. Each of these three groups received the same 30 assignments, but each group was provided with a different method of marking. All groups were provided with a copy of the assignment question (see attachment 1) and hard copies of the assignments. The groups were labelled as 1, 2 and 3.

Each of the 10 teachers in group 1 was asked to mark their 30 assignments based on the assignment marking criteria issued by the creator of the assignment task. Each of these markers were asked to record their marks on the papers themselves and also to list the total time spent marking each paper.

Similarly, each of the 10 teachers in group 2 were asked to mark the same 30 assignments and again record the time taken for each paper. However, participants in this second group were not provided with the marking criteria issued by the creator of the assignment task. Instead this group was provided with marking criteria designed by this research team (see attachment 2). This criteria was designed to be much more specific than that which was issued by the assignment's creator. They provided a marking guide with very specific reference to items we were looking for and assigned a particular numerical value to those items, which in some instances was so specific that it went into fractions. In addition, markers in this group were also provided with a one-page summary sheet upon which they could see a summary of the marking criteria and associated marks (again see attachment 2). The following is a typical example of how this would occur. It specifically relates to the notion of identifying the four most common types of ADR and ascribes a very clear numerical value to each part of the answer:

States/explains four processes of ADR –

- Negotiation. Simple equitable discussions between the parties to the dispute
- Mediation. Neutral third party used to assist in resolving the dispute
- Conciliation. Third party plays an active role in assisting the parties to resolve their disputes
- Arbitration. A third party hears evidence and arguments from both parties, then imposes a decision on the parties. The arbitrator is usually an expert in the relevant field

> *States/explains four processes – ½ mark per process (possible 2 marks)*
>
> (From attachment 1)

Group 3 were also issued with the same 30 assignments and again requested to record the mark and time taken for each of the papers. As was the case for group 2 participants, members of this group were not provided with the marking criteria issued by the creator of the assignment. Instead this group was issued with no criteria at all. It was left entirely to their professional judgment and skill to determine how they would assess the assignment and what mark they might ascribe to each of the 30 assignments they were presented with.

So, whilst each of the three groups of 10 markers were issued with the same 30 assignments, each of the three groups were asked to mark using distinctively different methods.

examination of J A Jackson (ed), *Professions and Professionalisation* (1970), Myron Lieberman, *Education as a Profession* (1956) and C O Hoale, *Continuing Learning in the Professions* (1981).

VIII. RESULTS

The following table shows, by group, the mean and the standard deviation (SD) for the results of all 30 assignments marked in the three different formats as described above. The average time taken to mark that item is also displayed.

	Group 1 Assignments			Group 2 Assignments			Group 3 Assignments		
	MEAN	**SD**	**TIME TAKEN**	**MEAN**	**SD**	**TIME TAKEN**	**MEAN**	**SD**	**TIME TAKEN**
Assignment 1	8.7	0.86	9.2	8.7	0.63	10.1	9.55	1.62	7
Assignment 2	10.05	0.60	10.1	10.3	0.89	10.2	9.95	1.12	9.5
Assignment 3	7.65	0.47	11.1	7.75	0.42	10	9.7	1.64	11
Assignment 4	6.4	0.57	9.7	6.65	0.53	8.5	7	0.82	12.2
Assignment 5	8.6	0.32	8.5	8.8	0.26	8	8.3	0.92	8.7
Assignment 6	8.5	0.41	9	8.35	0.47	8.6	8.9	1.20	9.9
Assignment 7	8.5	0.47	12	8.35	0.47	12.9	8.95	1.17	15
Assignment 8	6.85	0.88	16	7.3	0.75	11.1	7.6	1.39	12.9
Assignment 9	8.55	0.50	9.1	9.15	0.85	7.6	9.25	0.75	8.2
Assignment 10	8.25	0.59	9	8.75	0.59	10.1	9.15	1.13	9.8
Assignment 11	6.9	0.77	7.8	7.45	0.50	9	8.15	1.16	7.9
Assignment 12	7	0.67	12	7.7	0.42	12.6	7.8	0.59	10.9
Assignment 13	6.9	0.81	15.6	6.4	0.74	14.2	7.1	1.26	12.1
Assignment 14	8.85	0.88	14.6	7.65	0.78	14	7.7	0.82	15.6
Assignment 15	8.9	0.46	9	9	0.24	9.9	9	0.24	8.9
Assignment 16	8.95	0.83	13.2	8.3	0.48	12.9	8.5	0.53	11.8
Assignment 17	9.05	0.72	15	9.15	0.63	14.6	9.15	0.63	15.2
Assignment 18	10.35	0.71	16.5	10.8	0.48	14	10.35	0.63	12.7
Assignment 19	10.35	0.78	12.3	10.55	0.37	16.5	10.75	0.59	14.2
Assignment 20	9.2	0.71	10	9.2	0.59	10.1	9.35	0.82	11
Assignment 21	9.3	0.86	10.5	9.6	0.94	11.7	10.1	0.94	12.2
Assignment 22	9.1	0.74	19.3	8.15	0.47	12.5	8.55	1.01	16.6
Assignment 23	9.4	0.66	12.4	9.1	0.21	12	8.95	0.16	12.7
Assignment 24	9.1	0.70	13.2	8.95	0.98	13.1	8.75	0.72	11
Assignment 25	9.05	0.64	12.4	9.35	1.13	11.6	10.45	0.76	12.7
Assignment 26	9.05	0.80	12.4	9.05	0.80	12	9.05	0.80	11
Assignment 27	8.75	0.59	12	8.8	0.59	11.6	8.8	0.59	12.7
Assignment 28	8.55	0.64	11.3	8.45	0.80	11.6	8.9	1.31	10.5
Assignment 29	8.3	0.75	12	8.25	0.82	12	8.6	1.15	12.7
Assignment 30	8.8	0.42	12.1	8.9	1.37	11.6	9.1	1.07	11.7

Table 1. Distribution of data

The mean and standard deviation of the marks for each of the 30 assignments for the three different groups reveals some interesting data. The most significant difference in mean between the three groups occurs with Assignment 3 for groups 1 and 3, where the mean for group 3 is 9.7 and the mean for group 1 is 7.65. This is a difference of 2.05.

The most significant difference in standard deviation, perhaps unsurprisingly, also occurs with Assignment 3, but this time for groups 2 and 3, where the standard deviation for these groups are 0.42 and 1.64 respectively. This is a difference of 1.22.

The following graphs show the distribution of the assignment means and standard deviation by group.

Chart 1. Mean of assignments by group

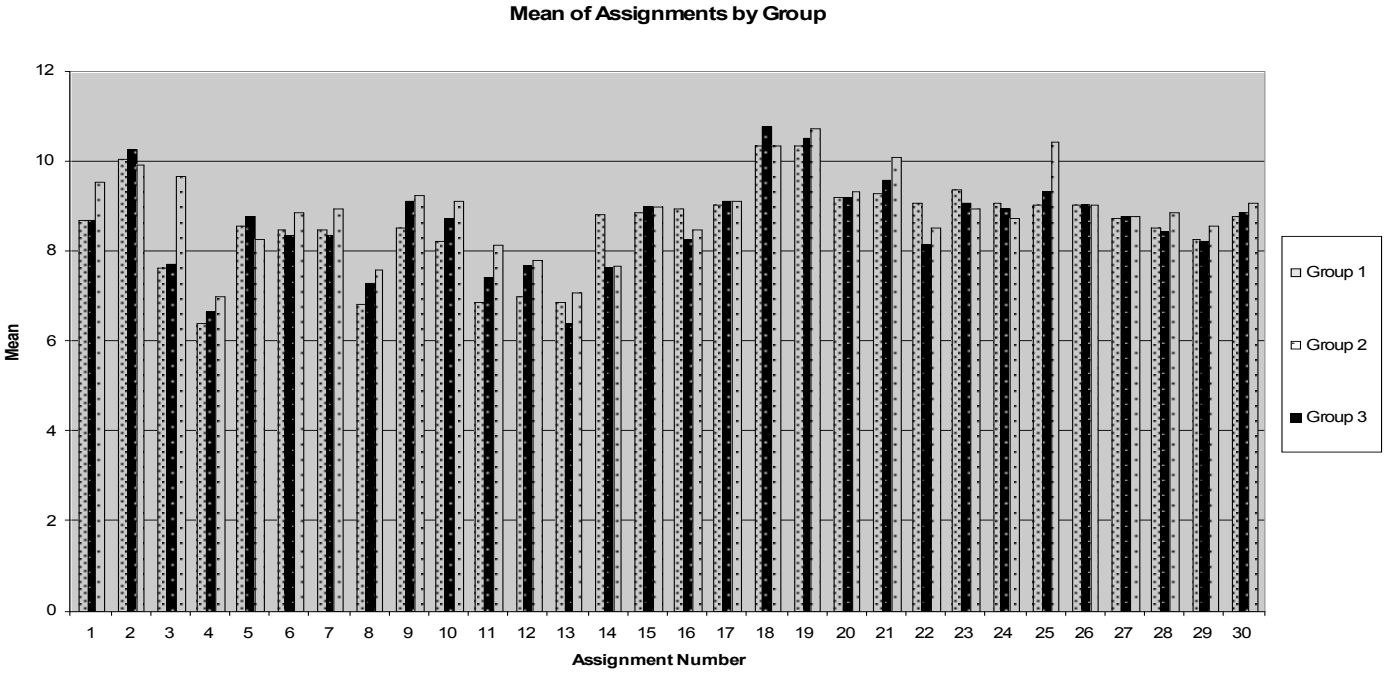
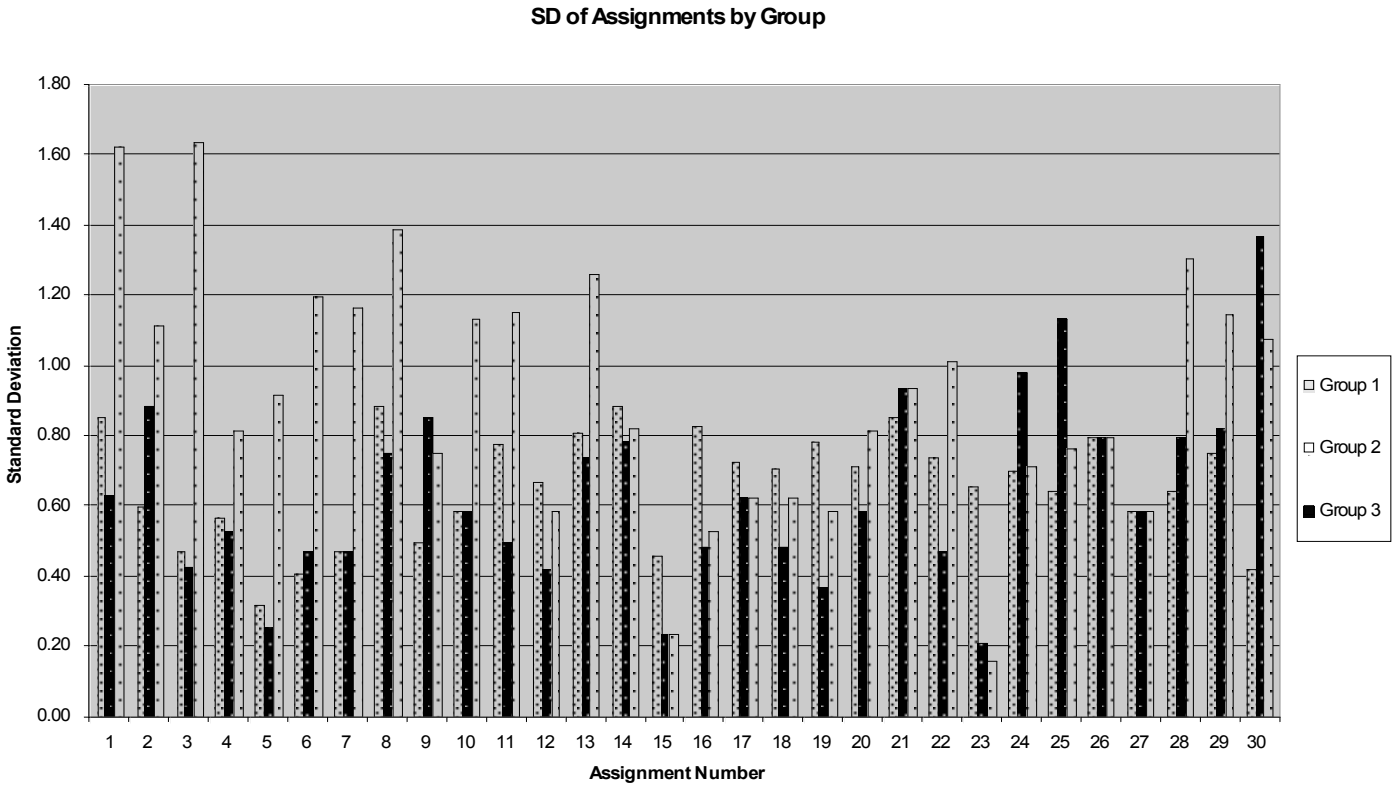


Chart 2. Standard deviation of assignments by group



Statistically, as independent samples, the t-test assesses whether the means of two groups are *statistically* different from each other.[64] This analysis is appropriate whenever you want to compare the means of two groups. In this instance, the means of all three groups will be compared with each other in pairs. That is, group 1 will be compared to group 2, group 2 to group 3 and group 1 to group 3.

The mean assignment mark of group 1 as compared to group 2 reveals a one-tailed P value equal to 0.4485 and a two-tailed P value equal to 0.8979, neither of which are considered to be statistically significant ($p>0.05$). The mean of group 2 as compared to group 3 reveals a one-tailed P value of 0.1295 and a two-tailed P value of 0.2582. By convention, these values are not considered to be statistically significant ($p>0.05$). The mean of group 1 as compared to group 3 reveals a one-tailed P value of 0.2011 and two-tailed P value of 0.2028. Again, by conventional criteria, these values are not considered to be statistically significant ($p>0.05$).

The time taken to mark each of the assignments was a recorded variable also considered in this research project. The longest average time taken to mark any of the assignments was by group 1 in connection with Assignment 22 (19.3 minutes), whilst the shortest average time taken to mark an assignment was by group number 3 with Assignment 1 (7 minutes).

The most significant difference in mean time, between the three groups, occurred with Assignment 22 for group number 1 and 2, where the mean time for group 1 was 19.3 minutes and the mean time for group 2 was 12.5 minutes. This was a difference of 6.8 minutes.

The following graph shows the average time taken to mark an assignment by each of the 3 groups.

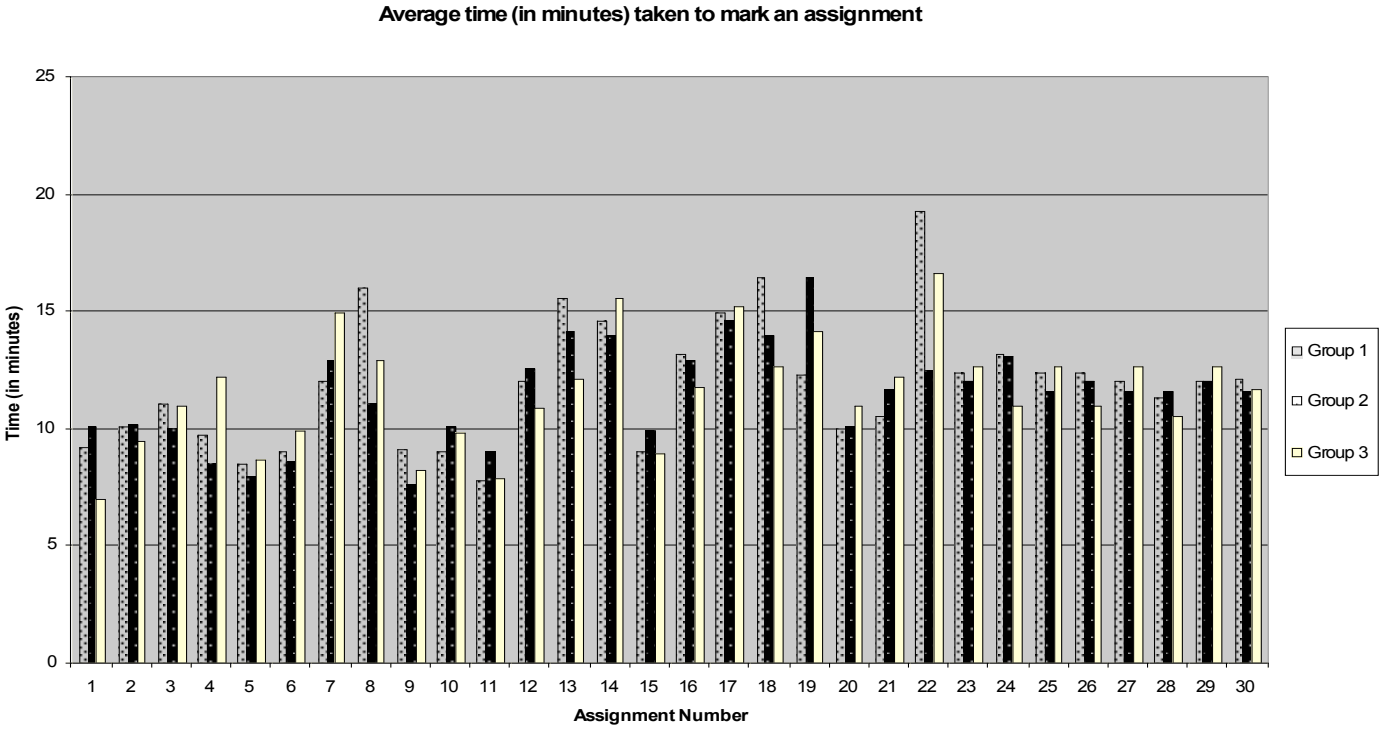


Chart 3. Average time taken to mark an assignment

64 For further information regarding t-tests and their statistical significance, see particularly Barbara G Tabachnick and Linda S Fidell, *Using Multivariate Statistics* (3rd ed, 1996) and Thomas Black, *Doing Quantitative Research in the Social Sciences: An Integrated Approach to Research Design, Measurement and Statistics* (1999). Also see Quantitative Methods in Social Sciences e-Lessons, *The T-Test* <http://www.columbia.edu/ccnmtl/projects/qmss/t_about.html> at 5 December 2008.

Again using a standard t-test statistical analysis, the mean time taken by group 1 as compared to group 2 reveals a one-tailed value of 0.249605 and a two-tailed value of 0.499210. The mean time taken by group 2 compared to group 3 shows a one-tailed value of 0.413322 and a two-tailed value of 0.826644. Finally, the average time difference between groups 1 and 3 shows a one-tailed value of 0.323618 and a two-tailed 0.647236. None of these t-test figures is considered to be statistically significant on a conventional basis. The statistical analysis via the t-test is reinforced when the average time taken by all group members is considered. For group 1, the average time taken to mark an assignment was 11.91 minutes. For group 2 it was 11.49 minutes, and for group 3 it was 11.61 minutes. Clearly there is very little difference in the overall average time taken to mark a paper.

IX. CONCLUSIONS

This research paper has reported on the critical need for ensuring the integrity of the marking process. What has become evident is that there appears to be no statistically significant difference between the three different marking methods used for marking a traditional humanities-based essay assignment for students in an interdisciplinary introductory law unit.

That is, three different methods have been used to mark the same assignments by utilising different markers to employ these different methods. One group was asked to use the criteria issued by the designer for the assessment task, another group was asked to use criteria designed by the research team responsible for this assignment, whilst the third group was provided with no specific criteria upon which to base their marking.

No statistically significant variations between the marks awarded for any of these three methods employed appears evident. In addition, there also appears to be no statistically significant variation of time taken to employ each of these methods. In particular, this has significant implications as it clearly takes significantly more time to develop an intricately based assessment criteria/marking guide, such as that provided for use by group 2, compared to doing nothing more than writing the actual question, which was the method utilised for marking by group 3.

Given all of the literature described in this paper, which highlights the need to ensure a rigorous making regime, the results obtained seem surprising and certainly need further validation. It is the intention of the authors to undertake a similar project within a tertiary environment in the near future.

It seems remarkable that, with relative ease, each of these three groups was able to obtain such similar results. It seems that intertester reliability may be more easily achieved than has been suggested by the literature. The authors are unable to directly account for the results as they have been achieved. It may be that the marking guide as presented to group 2 was so precise that it allowed for little potential deviation or that the teachers assigned to mark the assignments in group 3 (i.e. those with no guidance at all) all had a similar view of the points that they were looking for in a 'model' answer. It may also be important that this assignment was marked out of 15, allowing potentially little deviation in scores compared to, say, an assignment marked out of 40 or 50, which could very conceivably offer much greater scope for deviation of marks, potentially giving rise to statistically significant different results. We would point out that if this were the case then the parameters of this project would have been significantly altered. The experience of high school teachers, as professionals, in using standards-referenced assessment as described above could also be a consideration in accounting for the results achieved.

We do acknowledge that the marking scheme used for groups 1 and 2 has not been validated and there has been no attempt to analyse its reliability. This is clearly one of the limitations of this study and it may therefore be recognised as a contributing factor to the statistically insignificant results revealed between the means of the final marks awarded and the time taken for each of the assignments. These types of limitations could be

overcome by using both intertester and intratester reliability mechanisms on a larger scale. That is, for example, all of the 10 members of group 1 could have been asked to utilise the three different methods for marking each of the 30 assignments. Such a scheme would be extremely time consuming and well beyond the scope of this project, but would provide for more robust data to emerge.

Clearly, this was a small-scale research project and the results therefore need to be appreciated in that context. We are particularly conscious of the close contact that markers may have had as employees of the same school. We can see the clear need for this study to be repeated using additional assignments with similar size groups and with larger sized control groups. Nonetheless we are unaware of any collusion or manipulation of the final marks by any of the participants in the study and, therefore, present our conclusions accordingly.

APPENDIX 1 – FULL ASSIGNMENT ISSUED TO STUDENTS AND MARKERS

Introduction to Business Law 20084
Spring 2007
Assignment 1 Information Sheet

The assignment question is set out at the end of this information sheet. Please read the information carefully. FAILURE TO FOLLOW THE REQUIREMENTS STATED BELOW WILL RESULT IN SEVERE LOSS OF MARKS.

1. The Assignment

Marks: 15 Value: 15% of the final marks for the unit.

Discuss the processes available in alternative dispute resolution and explain its advantages and disadvantages.

In your answer you should make use of at least three secondary sources of information outside of the textbook.

2. Marking Criteria

Content & analysis
- identify relevant information
- accurately state & explain concepts
- critically analyse issues

Argument
- Structured
- Logical & coherent
- Supported by authority
- Contain supportable conclusions

Research
- Look beyond the textbook.
- Identify relevant sources accurately.
- Utilise the sources meaningfully in your discussion.
- Select sources of information which are authoritative. (An anonymous blog is not as authoritative as a report published by the Australian Law Reform Commission.)

Presentation
- Grammar, syntax, punctuation & spelling
- Layout & paragraphing
- Appropriate & adequate referencing
- Bibliography

3. Form of the assignment

The assignment should be presented as follows:
- A4 paper, stapled in the top left corner and typed
- Each page must be numbered and in black ink only (no other colouring)
- Only one side of the paper should be used

- Leave a margin of about 3cm on both sides of the page, so the marker can write comments
- Avoid eyestrain for the markers, by using:
    - 12 point font
    - one and a half line spacing
- Attach and complete the assignment cover sheet. Make sure you include the time and day of tutorial AND the tutor. A copy of the cover sheet to be used is attached
- DO NOT SUBMIT ASSIGNMENTS IN FOLDERS OR SIMILAR COVERS.

- **Word length: 1200 words**

The word limit includes reference details and any footnotes, but not the bibliography at the end. It is an indication of what the teaching staff believe to be necessary in order to provide an adequate answer on all issues. If you find you are below the word limit then you should carefully revise your work to check if you covered all relevant issues. If you are above the word limit then check to see that you are not discussing irrelevant material.

We are not interested in counting every word but you are expected to express yourself succinctly, so if you exceed the word limit by more than 10% of the total word limit, the tutor marking may not read beyond the extra 10%.

4. Referencing

Use the Harvard system, but carefully follow the modified Harvard style used by Terry and Guigni (the prescribed text) to refer to other materials.

References to texts must include page numbers in most cases (eg, p 8 of Terry and Guigni referring to Devlin). When referring to any cases or legislation *italicise* or <u>underline</u> case names and the titles of legislation, and include the standard legal citation for cases and legislation. (eg, p 33 of Terry and Guigni referring to *Walker v NSW* (1994) 69 ALJR 111 and to the *Bills of Exchange Act 1882* (UK))

A **bibliography** of all references must be attached at the end of your assignment. This is not to be counted in the word length. It should include material used by you in preparing the assignment (whether you refer to these directly in your assignment or not). Organise the references into separate sections depending upon whether the material is primary or secondary sources and written or electronic. The secondary material should be listed alphabetically by the author's surname.

5. Collusion and Plagiarism Warning

- Students will have an opportunity to discuss the assignment topic in tutorials in Tutorial 3. However, your preparation for the topic and your written answer is to be undertaken individually. No collusion is permitted.
- Copying from articles, books or other students' work without acknowledgement at each point of use, is plagiarism, as is rewording what you have read from another source without appropriate acknowledgement.
- Collusion and plagiarism are forms of academic misconduct for which severe disciplinary penalties can be imposed. For more details see the UWS policy.

APPENDIX 2 – MARKING GUIDELINES CREATED BY RESEARCH TEAM

Marking Guide

Content and analysis: /6

Accurately state and explain concepts:

Define "alternative dispute resolution"

- "any way of resolving a dispute without resorting to litigation"
- "additional" to court system, rather than "alternative"
- Negotiated agreement rather than solution imposed on parties

Does not offer a definition – 0 marks
Defines "alternative dispute resolution" from a low quality source e.g. internet research / Wikipedia or does not offer a source for definition – ½ mark
Defines "alternative dispute resolution" from an authoritative source e.g. textbook - 1 mark

States/explains four processes of ADR –

- Negotiation — Simple equitable discussions between the parties to the dispute
- Mediation — Neutral third party used to assist in resolving the dispute
- Conciliation — Third party plays an active role in assisting the parties to resolve their disputes
- Arbitration — A third party hears evidence and arguments from both parties, then imposes a decision on the parties. The arbitrator is usually an expert in the relevant field

States/explains four processes – ½ mark per process
(possible 2 marks)

Critically analyse issues:

Advantages/disadvantages of four processes of ADR

- Usually faster and less costly
- More flexible and responsive to individual needs of parties
- May provide a "second class" type of justice
- May re-enforce existing power relationships between the parties
- Works best when parties co-operative approach to problem-solving rather than insisting on maintaining a particular adversarial position with no room for flexibility
- Other advantages/disadvantages researched by student
- Comparison between processes

Provides "list-style" recitation of advantages and disadvantages of processes and ADR generally: 1 – 1½ marks
Makes some comment/critique of processes and ADR generally: 2 – 2½ marks
Critical analysis of processes, comparison of processes: 3 marks
(possible 3 marks)

Argument **/2**

Quality of argument:
Structured; logical and coherent, supported by authority, contains supportable conclusions

Low quality of argument – illogical structure, incoherent reasoning, little or no support offered for conclusions – 0 - ½ marks
Acceptable quality of argument – some attempt at structure and some authority offered in support – 1 – 1½ marks
Excellent quality of argument – logically structured and flowing, well supported by authority – 2 marks

Research **/3**

Look beyond the textbook
Identify relevant sources accurately

½ mark per secondary source identified and used by student (possible 1 ½ marks)

Utilise the sources meaningfully in discussion
Select sources of information which are authoritative

Low quality of sources used, no meaningful discussion or analysis (e.g. block quote from unreliable internet source with no critique or analysis): 0 marks
Acceptable quality of most sources used, some attempt to discuss and critically analyse sources: ½ - 1 marks
Excellent quality of sources used, authoritative sources, sources well integrated into discussion and argument – 1 ½ marks

Presentation **/2**

Grammar, syntax, punctuation and spelling:

Unacceptable – many errors, poor standard – 0 marks
Acceptable – fair standard - ½ mark
Excellent – high quality, very few errors – 1 mark

Format:
Complies with the following requirements:

- A4 paper, stapled in the top left corner and typed
- Each page numbered
- Black ink only
- Only one side of paper used
- Margin of about 3cm on page
- 12 point font
- One and a half line spacing
- Attaches completed assignment cover sheet
- Complies with word limit 1200 words, including reference details and footnotes, but not bibliography (10% leeway either way, e.g. 1080/1320)

Unacceptable – complies with very few requirements – 0 marks
Acceptable – complies with majority of requirements – ½ mark
Excellent – complies with all requirements – 1 mark

Referencing and Bibliography /2

- Referencing: Harvard system, using modified Harvard style used by Terry and Guigni (the prescribed text) to refer to other materials.

  References to texts must include page numbers in most cases (eg, p.8 of Terry and Guigni referring to Devlin). When referring to any cases or legislation *italicise* or underline case names and the titles of legislation, and include the standard legal citation for cases and legislation. (eg, p.33 of Terry and Guigni referring to *Walker v NSW* (1994) 69 ALJR 111 and to the *Bills of Exchange Act 1882* (UK))

- Bibliography: of all references attached at end of assignment, organised into separate sections: primary or secondary sources and written or electronic - secondary material should be listed alphabetically by the author's surname.

Referencing poor or non-existent – 0 marks
Referencing acceptable, correct system used, some errors – ½ mark
Referencing excellent, correct system used, minimal errors – 1 mark

Bibliography poor or non-existent – 0 marks
Bibliography acceptable, most sources listed, not separated into sections – ½ mark
Bibliography excellent – all references listed in separate sections – 1 mark

TOTAL OUT OF 15
ASSIGNMENT IDENTIFYING CODE:

MARKER:

CRITERIA			
Content & analysis			
Accurately state and explain concepts	Define "alternative dispute resolution"	/1	
	States/explains four processes of ADR: Negotiation Mediation Conciliation Arbitration	/2	
Critically analyse issues	Advantages/disadvantages of four processes of ADR	/3	
			/6
Argument			
Quality of argument	Structured Logical and coherent Supported by authority Contains supportable conclusions	/2	
			/2
Research			
	Look beyond the textbook Identify relevant sources accurately	/1.5	
	Utilise the sources meaningfully in discussion Select sources of information which are authoritative	/1.5	
			/3
Presentation			
	Grammar, syntax, punctuation, spelling	/1	
	Format	/1	
			/2
Referencing			
	Referencing	/1	
	Bibliography	/1	
			/2
TOTAL			**/15**