TITLE

Agreement in Student Performance in Assessment Between Disciplines

AUTHORS

David James
Scott Fleming

AFFILIATION

School of Sport & Leisure
University of Gloucestershire

Agreement in Student Performance in Assessment between Disciplines

Introduction

Since 1990 there has been increasing attention to assessment in higher education (HE), and there have been various attempts to inform the professional discourse of assessment in HE. Some have been concerned with the philosophy of assessment and of assessment practice (e.g., Miller *et al*. 1998, Swann & Ecclestone 1999), others have focused on more general advice (e.g., Baume & Baume 1992, Brown 2001) and the application of specific examples (e.g., Habeshaw *et al*. 1993, Race 1995, 1996). It has also been claimed that student learning is assessment driven (Habeshaw *et al*. 1993), and even that assessment is of *singular* importance to the student experience (Rust 2002).

The rationale underpinning effective assessment in HE, as well as its importance have both been widely explored (e.g., Race 1995). Broadly, the key features include: diagnosis (of different kinds), evaluation of progress, providing feedback (to learners, tutors and external agencies), motivation, demonstration of the acquisition of skills and/or competencies, measuring achievement. It is now a widespread view that multiple methods of assessment should be used for multiple assessment expectations (Brown & Knight 1994), and that students should experience a wide and varied 'assessment diet' within a programme of study[1]. Brown *et al*. (1996: 14) explain: "Assessment that is 'fit for purpose' uses the best method of assessment appropriate to the context, the students, the level, the subject and the institution".

Innovation in assessment practice has been endorsed by different agencies (e.g., Institute for Learning and Teaching in Higher Education, Learning and Teaching Support Network), but there is still a culture of traditionalism in many Universities. As recently as 1996, Brown *et al*. reported that over 80% of assessment in Universities is based on essays, reports and traditional, timed, unseen

[1] For an exhaustive annotated list of assessment modes and methods, see Brown 2001.

examinations (Brown *et al*., 1996); and as Buswell (2002) has noted, traditional unseen examinations (as well as coursework essays) may be thought to stifle some principles of innovative assessment. Some of the impetus for innovation in assessment has also been developed through concerns about the prevalence of plagiarism and other forms of 'dishonesty in assessment' (Yorke *et al*. 2000, Larkham & Manns 2002)[2]. Though there are other examples of more educationally progressive forms of innovation in assessment (e.g., Fullerton 1995), and these are often connected to good practice in feedback to students (e.g., Cantwell 2002). Assessment, as Race (1995: 82) observes, "at best, is a very inexact science". Inevitably, and quite properly, validity and reliability in assessment continue to be emphasised, although evaluation of the degree of validity and reliability is rarely undertaken. In those Universities where a wide range of assessment methods are practised, any suggestions of differential levels of performance often raise questions of comparability.

At the module level, evidence for different performance across assessment points is superficial. Yorke *et al*. (2000) note the general perception that "…coursework marks tend to be higher than marks awarded in examinations" (p.14) and they point to some preliminary evidence to that effect. These matters need to be considered with some care and rather more attention to detail than has often been the case hitherto. Leaving aside some of the technical debates about whether students' assessment data are actually interval or ordinal data (cf. Yorke 2001), there are some important implications for modular programmes in particular. For instance, the diversity of assessment practice across different disciplines and subject areas raises profound questions about equity – especially, Yorke *et al*. (2000) claim, in modular schemes.

Whilst distinctions are often made between the natural and social science subjects, it is perhaps useful to first consider the range of assessment tasks employed within a discipline, and their effect on performance. For example, within a discipline such as exercise physiology, which forms part of most

[2] There is already a sophisticated network of websites providing students with the opportunity to buy and download written essays, e.g., www.termpapers4u.com and www.papersheaven.com.

sport and exercise sciences programmes, a range of assessment tasks are often employed, drawing on essay type, mathematically-based, and practical skill assessments. Therefore, an initial question, when exploring the broad area of performance in assessment, might usefully be, how does student performance vary across assessment tasks within the same module? Secondly, how does student performance vary in the same assessment task across modules? Such variation in student performance is, perhaps, best thought of as the level of 'agreement' in performance.

Previous (traditional) attempts to investigate 'agreement' in performance (of any type) have involved significance difference tests and intraclass correlation coefficient (Bland & Altman 1990), but neither of these approaches is suitable, and both limit the extent to which the findings are meaningful - see Technical Note. There have, however, been recent advances in statistical techniques suitable for examining 'agreement' in student performance (Bland & Altman 1986). Specifically, a 'limits of agreement' approach, widely used in medicine and sport science (Webber *et al*. 1994, Atkinson & Nevill 1998), is suggested as a 'user-friendly' and robust way to undertake this analysis.

Assessment of a student's performance on a particular module may often be thought of as a single evaluation of the extent to which the student has met some or all of the module's learning outcomes. More helpfully, however, when there is more than one assessment task in a particular module, it may be thought of as the combination of different assessment tasks (whatever the weighting attached to each of them). In this sense, the level of agreement between performances on the different tasks may elucidate the nature of overall student performance further still. Typically, assessment tasks within a module tend not to be of the same kind; often they are complementary, sometimes through the use of different media. The primary aim of the present study was to investigate agreement in student performance between assessment tasks *within* two modules.

Additionally, however, many modules adopt conventional combinations of assessment tasks. Previously, in the social sciences and the humanities for example, this might have been a written

essay and unseen examination (containing essay questions); or more recently, perhaps a group poster presentation and individual seminar (accompanied by written paper). There has been little substantive exploration of the level of agreement in student performance in similar tasks across different modules. A secondary aim, therefore, was to investigate agreement in student performance in the same assessment task *between* modules from similar disciplines (i.e., Anatomy and Physiology).

As a final but important contextualising note, the nomenclature adopted for the statistical techniques that underpin this study is, of course, value-laden. 'Agreement' should not necessarily be interpreted as a virtue in this regard, anymore than 'failure to establish agreement' (or the even more pejorative term 'disagreement') should be regarded as a deficiency or shortcoming in assessment protocols. There are important reasons why, for example, within module assessments tasks might not demonstrate agreement – they might be examining different skills, competencies and knowledge through different media. There are also reasons why similar tasks from different modules might evidence differential patterns of student performance – they may involve conceptually different material requiring different kinds of cognitive competencies. Examination of the extent to which agreement exists within a module's assessment protocol, or between similar tasks in different modules, however, may signal some important characteristics about the diet of student assessment experiences, and of performance on them. The levels of agreement may, therefore, provide a basis for more nuanced and context-sensitive examination of student assessment. This is a theme to which the discussion will return in the conclusion.

Method

Study Design

The sample for this study was drawn from two modules, both of which form part of the (introductory) level curriculum for students undertaking one of the three 'science' programmes of

study in sport and exercise[3]. The two modules were Anatomy and Assessment of Structure (hereafter referred to as Anatomy) and Introduction to Physiology of Sport, Exercise & Health (hereafter referred to as Physiology). The basis for selection of this sample reflected the need to assess agreement across assessment points within a module, and across modules. When looking at agreement across modules, it was possible to assess student performance in the same type of assessment. To ensure potential confounding variables were minimised, the modules were taken from the same level of study, and ran in the same academic year (2000-2001). There were 267 students registered for the Anatomy module, and 196 for the Physiology. A total of 180 were registered for both.

Student performance was assessed on each module through three assessment points. Agreement of performance within each module was assessed by comparing performance in each assessment point against each other assessment point in turn. This resulted in three comparisons within each module. Additionally, two of these assessment points were similar when comparing the two modules, which also allowed cross module comparisons of student performance. Specifically, the common assessment points were a Skills Test and an Examination. In the case of the Physiology module, the other assessment point was a Laboratory Report; and in the case of the Anatomy module, the other assessment point was a Practical File. In both modules, the Examination was multiple-choice, and of one-hour duration; the only difference being that the Anatomy Examination was computer based, whereas the Physiology Examination was a traditional paper based examination.

The Skills Test was a practical test that was designed to assess a student's ability to undertake skills developed through the module. There were four different skills testing stations, and the test required that each student spend a maximum of ten minutes at one of them. Students had prior knowledge of the skills upon which they would be assessed, but were randomly assigned to one of the stations on arrival for the test. The Anatomy Skills test required the students to identify an anatomical landmark

[3] Validated in 2000, the University of Gloucestershire's portfolio of sport and exercise related provision includes three

and to measure a length, breadth, girth or skin-fold. The Physiology Skills Test required the students to undertake assessment of lung function, blood pressure, minute ventilation or a progressive exercise protocol whilst complying with health and safety guidelines.

Data Analysis

Student performance data (i.e., percentage marks) for each assessment point were acquired from central student records of electronic module results. Data were initially cleaned by removing student marks when no attempt was made at an assessment point. Data were then sorted by student identification number in order to match students across modules. This process allowed deletion of marks if a student was not registered on both modules. Clearly this was only necessary when student performance was compared across modules. Once paired data were available after the initial cleaning, it was no longer necessary to store students' identification numbers.

The cleaned data were then used to assess agreement between assessment points following the procedure described by Bland and Altman (1986) – see Technical Note. The first part of this process involved calculating the arithmetic mean mark for each student, and the difference between the two marks for each student. The arithmetic mean of the differences was then calculated, and used to represent the accuracy or 'bias'. The standard deviation (SD) of the differences was also calculated, and used to represent the precision or 'agreement'. Normally the extent of agreement is represented as 95% confidence intervals (i.e., 1.96 x SD), and the findings are presented through a 'limits of agreement plot' for each comparison[4].

However, in the case of many comparisons in the present study, the limits of agreement plot showed a clear trend in the data, such that the differences (plotted on the y-axis) increased or decreased as the arithmetic mean performance (plotted on the x-axis) increased. This is a common finding when

named B.Sc. (Honours) awards in Sport and Exercise Sciences, Sport Science and Exercise and Health Sciences.

[4] A 95% confidence interval is derived from a sample of normally distributed data points, and defines the interval within which 95% of data points are contained.

examining agreement data (Bland & Altman, 1999), so an approach was adopted to account for the trend. Accounting for the trend is necessary, since failure to do so results in a meaningless value for bias and an exaggerated value for agreement. The approach for accounting for the trend involved fitting a least squares' regression line to the limits of agreement plot. The equation of the regression was used to remove the trend from the data, allowing revised differences to be calculated. These differences were then used to determine agreement (i.e., 1.96 SD) around the regression line, and plotted on the original limits of agreement plot. The bias then being reflected by the regression line.

Ethics statement

The University's principles and procedures on research ethics were adhered to throughout the study. In particular, data on student performance were presented such that identification of individual student performance was impossible, thereby complying with the requirements of the Data Protection Act. Restricted access to the data is permitted only to those who have administrative (e.g., data collation and processing) and academic functions (e.g., management roles with teaching, learning & assessment [TLA] responsibilities; roles overseeing pastoral responsibility, and course leaders). In this instance one of the authors (DJ) had joint responsibility for TLA within the School of Sport & Leisure.

Results

The findings are considered by first examining agreement of assessment within a module (Anatomy followed by Physiology module), followed by agreement of assessment between modules (Skills Test followed by Examination). In all cases, the findings are presented as figures (limits of agreement plots) and in the form of summary tables. Throughout, the application of legends to figures, and headings to tables, shows which assessment point is subtracted from another to give the bias. For example, Practical File – Exam, identifies that the Examination score is subtracted from the Practical File score to give the bias.

The second summary table for each module considers students' performance across the assessment points being compared. For example, a 'high' level of performance is indicated by an arithmetic mean score of greater than 70% in the two assessment points being compared. A 'low' level of performance represents a score of less than 40%, and a medium level of performance represent a score of ~55%.

INSERT FIG 1 ABOUT HERE

Anatomy Module

Agreement between the student performance in the Practical File and Examination is shown in Figure 1 (top panel) and summarised in tables one and two respectively. With the exception of the students with the high level of performance, the performance in the Examination was stronger than performance in the Practical File. The general trend is that as the students' overall performance deteriorates (moving from high to low levels of performance), the performance in the Practical File gets relatively weaker, and the performance in the examination gets relatively stronger. The limits of agreement plot shows an agreement of +/-32.9 % between the Practical File and the Examination.

Table 1: Bias when comparing assessment performance within the Anatomy module

Performance level	40%	50%	60%	70%
Practical File – Examination	-10.0%	-6.1%	-1.6%	2.9%
Practical File – Skills Test	-9.4%	-12.0%	-14.7%	-17.4%
Skills Test – Exam	-4.0%	2.7%	9.4%	16.1%

Agreement between the student performance in the Practical File and Skills Test is shown in Figure 1 (middle panel) and summarised in tables one and two respectively. In general, the performance in the Skills Test was stronger than performance in the Practical File. The general slight trend is that as the students' overall performance deteriorates (moving from high to low levels of performance), the

performance in the Practical File ceases to be so relatively weak, and the performance in the Skills Test ceases to be so relatively strong. The limits of agreement plot shows an agreement of +/-43.1 % between the Practical File and the Skills Test. In this module, the agreement between the two non-examination assessment points demonstrated greater bias at the good performance extreme, but perhaps more importantly, considerably greater lack of agreement across the entire performance range.

Table 2: Summary of assessment performance in the Anatomy module according to performance category

Low performance	Medium performance	High performance
Examination > Practical File	Examination > Practical File	Practical File > Examination
Skills Test > Practical File	Skills Test > Practical File	Skills Test > Practical File
Examination > Skills Test	Skills Test > Examination	Skills Test > Examination

Agreement between the student performance in the Skills Test and Examination is shown in Figure 1 (bottom panel) and summarised in tables one and two respectively. The general trend is that as the students' overall performance deteriorates (moving from high to low levels of performance), the performance in the Skills Test gets relatively weaker, and the performance in the Examination gets relatively stronger. The limits of agreement plot shows an agreement of +/-34.1 % between the Skills Test and the Examination.

An overall rank order of relative performance in assessment tasks therefore indicates that, in general, students performed better in the Skills Test than in the Examination, and better in the Examination than in the Practical File. However, it is interesting to note that students with a low level of performance tend to do better in the Examination relative to other points of assessment (see table 2).

INSERT FIG 2 ABOUT HERE

Physiology Module

Agreement between the student performance in the Report and Examination is shown in Figure 2 (top panel) and tables three and four respectively. The students with a high level of performance tended to perform relatively better in the Examination, whereas the students with a low level of performance tended to perform relatively better in the Report. The limits of agreement plot shows an agreement of +/-33.6 % between the Report and the Examination.

Table 3: Bias when comparing assessment performance within the Physiology module

Performance level	40%	50%	60%	70%
Report – Examination	5.9%	2.3%	-1.4%	-5.0%
Report – Skills Test	-25.9%	-21.8%	-17.6%	-13.5%
Skills Test – Examination	19.0%	20.0%	20.9%	21.8%

Agreement between the student performance in the Report and Skills Test is shown in Figure 2 (middle panel) and in tables three and four respectively. Throughout the range of student performance (i.e., low to high level of performance), the bias suggests that students perform poorly in the Report relative to the Skills Test. Also, the general trend was that as students' overall performance deteriorated (moving from high to low levels of performance), students' tended to perform relatively worse in the Report. The limits of agreement plot shows an agreement of +/-38.3 % between the Report and the Skills Test. In this module, the agreement between the two non-examination assessment points demonstrated greater bias at the poor performance extreme, and interestingly, a greater lack of agreement across the performance range. The greater lack of agreement and considerable bias is a feature shared with similar assessment points in the Anatomy module.

Table 4: Summary of assessment performance in the Physiology module according to performance category

Low performance	Medium performance	High performance
Report > Examination	Report > Examination	Examination > Report
Skills Test > Report	Skills Test > Report	Skills Test > Report
Skills Test > Examination	Skills Test > Examination	Skills Test > Examination

Agreement between the student performance in the Skills Test and Examination is shown in Figure 2 (bottom panel). Throughout the range of student performance (i.e., low to high level of performance), the bias suggests that students perform poorly in the Examination relative to the Skills Test. Also, the general slight trend was that as students' overall performance deteriorated (moving from high to low levels of performance), students' tended to perform relatively worse in the Skills Test. The limits of agreement plot shows an agreement of +/-32.7 % between the Skills Test and the Examination. The positive bias, whereby students perform better in the Skills Test rather than the Examination, is a striking feature of this comparison.

An overall rank order of relative performance in assessment tasks therefore indicates that, in general, students performed better in the Skills Test than in the Report, and better in the Report than in the Examination. However, a distinction is evident between high performing students and others, in that the Examination performance is better than the Report performance (see table 4).

INSERT FIG 3 ABOUT HERE

Skills Test

Agreement between the student performance in the Anatomy and Physiology Module is shown in Figure 3 and summarised in table five. The general slight trend was that as students' overall performance deteriorated, students' tended to perform relatively worse in Anatomy. It is worth

mentioning at this point that Anatomy and Physiology took place in different semesters, and any comparison might usefully note this potential confounding variable. The limits of agreement plot shows an agreement of +/-41.5 % between the Anatomy and Physiology modules.

INSERT FIG 4 ABOUT HERE

Examination

Agreement between the student performance in the Anatomy and Physiology Module is shown in Figure 4. The general slight trend was that as students' overall performance deteriorated, students' tended to perform relatively worse in Physiology. Through the range of student performance, however, performance tended to be relatively better in the Anatomy Examination (i.e., positive bias). The limits of agreement plot shows an agreement of +/-24.2 % between the Anatomy and Physiology modules. Interestingly, this agreement is considerably better than that for the Skills Test.

Table 5: Bias when comparing assessment performance between the Anatomy and Physiology modules

Performance level	40%	50%	60%	70%
Skills Test (Anatomy – Physiology)	-10.4%	-6.1%	-1.7%	2.7%
Examination (Anatomy – Physiology)	15.0%	12.4%	9.7%	7.1%

Discussion

In most Universities, students may be exposed to a range of assessment tasks within a programme of study, including examinations of various types, report writing, essay writing, poster presentations and oral presentations. Anecdotally, it is often claimed that, regardless of knowledge and understanding, performance of an individual student may vary according to the particular type of assessment task (Yorke *et al*. 2000). Also that certain types of assessment are more difficult for all students, and even that students may select modules on the basis of the assessment tasks involved. If claims about lack

of agreement in student performance between assessment tasks are true, students might be supported differently depending on the assessment task they struggle with. Alternatively, the assessment tasks themselves might require revision. Even the performance of the assessors might require investigation. Before any action may be recommended, such claims need to be investigated systematically.

The present study examined the agreement in performance in different assessment tasks within a module, and the same assessment tasks between modules. In order to control potential confounding variables, in making comparisons across modules, similar discipline modules were selected, assessment tasks were well matched (e.g., multiple choice examination in both cases), modules took place at the same level of study, but within different semesters. In making comparisons within modules, the same assessors were involved in different assessment points. A particularly useful feature of the present study was the large data set involved in each analysis, resulting in meaningful findings for the population under consideration.

Contrary to the view that students do consistently better in one form of assessment compared with another (cf. Yorke *et al*. 2000), the findings from the present study suggest that this is not the case. When comparing performance in two assessment points within each module, relative student performance varies as a function of the average mark from the two assessment points. For example, a student with a low level of performance in the Anatomy module performed relatively better in the Examination than in the Practical File. The converse is true for a student with a high level of overall performance in the module. Within a module, the only comparison of two assessment points that yielded a consistent bias across the assessment range was when the performance in the Skills Test and Examination was compared in the Physiology module. In this case, students consistently scored better in the Skills Test. The notion that examinations yield lower levels of performance than other forms of assessment (Yorke *et al*. 2000) is not evident from the present study. The relative performance in the examination appears, in general, to be a function of student level of performance.

A further common claim, that strong students score relatively better in an examination (cf. Elton & Johnston 2002), also appears not to be the case. Students with a high level of performance score relatively worse in the Examination in the Anatomy module, regardless of which other assessment point the Examination is compared with, whereas students with a high level of performance score relatively better in the Examination in the Physiology module when compared with the Report.

A claim that students generally perform consistently in the same types of assessment may be challenged based on the findings of the present study. For example, when comparing the performance in the Skills Test, it was clear that students with a lower level of performance scored better in the Physiology Skills Test, whereas the students with a higher level of performance scored better in the Anatomy Skills Test. Whilst student performance in the Anatomy Examination tended to be better than performance in the Physiology Examination, relative performance still varied as a function of average student performance. For example, a student scoring an average of 70% in the two assessment points would score 7.1% higher in the Anatomy Examination, whereas a student scoring an average of 40% would score 15.0% higher in the Anatomy Examination.

It is not possible to claim that one module was more challenging than another in the present study. It is not even possible to claim that students with lower levels of performance found one module more difficult than another, since performance varies differently as a function of level of student performance, depending on the form of assessment examined.

Conclusion

The results of the present study, which have so far been discussed in terms of the bias in performance, challenge several commonly held beliefs. First, students do not consistently perform better in one form of assessment compared with another. We have shown that relative performance

in assessment is generally a function of student level of performance. Second, students with a higher level of performance do not tend to do better in examinations. We have shown that this was the case in one module (Anatomy) but not in another module (Physiology). Third, whether students are low or high level performers, performance in a form of assessment is not consistent, even within the same broad discipline. In other words, the performance of the same student is neither always (consistently) good in examinations, nor consistently bad in examinations. Examination performance appears to be a function of the discipline, as well as the student level of performance.

When examining the degree of performance agreement between assessment points, we found that agreement between the three assessment points in the two modules examined was broadly similar. So within a module of the discipline examined in the present study, it may be claimed that student performance between assessment points agreed by about +/- 33 %. The only clear exception to this was the lower level of agreement between the Practical File and Skills Test in the Anatomy module (+/- 43.1 %) The level of agreement is not a function of the student level of performance, so no claims about students with a higher level of performance showing greater levels of agreement may be advanced.

When examining the degree of performance agreement between similar assessment points in different modules, we found that agreement between the assessment points varied according to the type of assessment. Agreement was better for the Examination (+/- 24.2%) than for the Skills Test (+/- 41.5%). It is interesting to note that there is generally no less agreement when comparisons are made between assessment points within a module, compared with similar assessment points between modules.

In summary, despite some of the prevalent beliefs about assessment in HE, in the modules examined in the present study, students did not perform consistently better in one particular form of assessment. Students who showed different levels of performance (e.g., high versus low) did not appear

consistently to do better in a particular form of assessment. Finally, performance was extremely variable, with agreement in most comparisons not being better than +/- 30%. Further research is required to examine agreement in performance in different disciplines, and between different levels of study. Once a comprehensive examination of agreement in student performance has been conducted, researchers and practitioners will be better placed to ask informed questions. Such questions might include:

- Is performance agreement a useful indicator within and between modules?
- Are interventions necessary to influence performance agreement?
- Should the variety of assessment modes be determined by student choice?
- Should assessment of performance agreement be part of routine evaluation of modules and courses?

References

Atkinson, G. & Nevill, A.M. (1998) Statistical methods in assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine* 26: 217-238

Baume, D. & Baume, C. (1992) *Assessing students' work.* Oxford: Oxford Brooks University Centre for Staff Development.

Bland, J.M. & Altman, D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* i: 307-310

Bland, J.M. & Altman, D.G. (1990) A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers Biol Med* 20 (5): 337-340

Bland, J.M. & Altman, D.G. (1999) Measuring agreement in method comparison studies. *Statistical Methods Med Research* 8:135-160

Brown, G. (2001) *Assessment: a guide for lecturers*. York: Learning and Teaching Support Network

Brown, S., Race, P. & Smith, B. (1996) *500 tips on assessment*. London: Kegan Paul.

Buswell, J. (2002) Examinations assessed! Diversity in assessment. *Link* 5: 17-19.

Cantwell, J. (2002) Formative feedback. *Link* 5: 15-16.

Elton, L. & Johnston, B. (2002) *Assessment in universities: a critical review of research*. York: Learning and Teaching Support Network.

Fullerton, H. (1995) Embedding alternative approaches in assessment, in Knight, P. (ed.) *Assessment for learning in higher education*. London: Kogan Page, pp.111-123.

Habeshaw, S., Gibbs, G. & Habeshaw, T. (1993) *53 interesting ways to assess your students*. 3rd edition. Bristol: TES.

Knight, P. (ed.) (1995) *Assessment for learning in higher education*. London: Kogan Page.

Larkham, P. & Manns, S. (2002) Plagiarism and its treatment in higher education. *Journal of Further and Higher Education* 26 (4): 339-349.

Race, P. (1995) The art of assessing 1. *New Academic* 4 (2): 3-6.

Race, P. (1996) The art of assessing 2. *New Academic* 5 (1): 3-6.

Race, P. & Brown, S. (2001) *The lecturer's toolkit,* 2nd edition. London: Kogan Page.

Rust, C. (2002) The impact of assessment on student learning: how can research literature practically help to inform the development of departmental assessment strategies and learner-centred assessment practices. *Active Learning in Higher Education* 3 (2): 145-158.

Swann, J. & Eccelstone, K. (1999) Improving lecturers' assessment practice in higher education: a problem-based approach. *Educational Action Research* 7 (1): 63-87.

University of Gloucestershire (2002) *University policy on teaching, learning and assessment*. Centre for Learning and Teaching.

Webber, J., Donaldson, M., Allison, S.P., MacDonald, I.A. (1994) A comparison of skinfold thickness, body mass index, bioelectrical impedance analysis and dual X-ray absorptiometry in assessing body composition in obese subjects before and after weight loss. *Clinical Nutrition* 13: 177-182

Yorke, M. (2001) *Assessment: a guide for senior managers*. York: Learning and Teaching Support Network.

Yorke, M., Bridges, P., Woolf, H. *et al*. (2000) Mark distributions and marking practices in UK higher education: some challenging issues. *Active Learning in Higher Education* 1 (1): 7-27.

Acknowledgements

This research was funded by a University of Gloucestershire Scholarship of Learning & Teaching (SoLT) Project Grant. We are grateful to Dr Dan Wood, Prof. Clare Morris & Dr Stephen Cooper for advice on statistical methods employed in this study.

Technical note

Traditionally, agreement between methods of measuring something (in this case student knowledge or competence) has been assessed inappropriately by using product moment correlation coefficient (*r*) and significance tests. Correlation is appropriately used to assess the strength of a relationship between two variables. However, such a relationship provides little useful information about agreement. Correlation is inappropriate for assessment of agreement between methods for the following reasons (adapted from Bland & Altman, 1986):

1. A perfect relationship, as indicated by an *r*-value of 1.00 may be attained with extremely poor agreement. For example, when viewing a scatter plot of one method of measurement plotted against another, it is only the extent to which the data points fall close to the line of identity that indicates agreement. A high *r*-value may be achieved with data points far away from the line of identity.
2. The strength of a relationship is influenced by the range of numerical values in a sample. For example, if student marks in a sample ranged between 40% and 70%, the strength of the relationship would be very different from a sample with a mark range of 0% - 100%, regardless of the degree of agreement.
3. The statistical significance of a relationship indicates little about agreement. It is highly likely that two methods of measurement of the same thing (in this case student knowledge or competence) will be related, as demonstrated through a statistical significance test.

An appropriate approach for the assessment of agreement between methods is to plot the difference between the methods (*y*-axis) against the mean value of the two methods (*x*-axis) (see, for example, figure 1). For example, if one student scored a mark of 65% in a physiology report, and 71% in a physiology exam within the same module, the difference is reported as 6% and the mean is reported as 68%. A data point is then plotted for this student. Once data points have been plotted for all

students in the sample (i.e., on the module), the mean and standard deviation of the differences is calculated. The mean of the differences represents the 'bias', and the standard deviation of the differences represents the 'agreement'.

It is suggested that the degree of agreement is expressed as a 95% confidence interval, and illustrated on the plot. The 95% confidence interval is calculated by multiplying the standard deviation by 1.96 providing the data are normally distributed. However, should the data not be normally distributed, a multiplication by 2.00 is recommended (Bland & Altman, 1986). The 95% confidence intervals illustrate that one can be 95% confident that in the population from which the sample was drawn, agreement will be contained within these limits. Having undertaken this procedure, the researcher or practitioner should normally then ask the question, is this level of agreement appropriate?

Figures

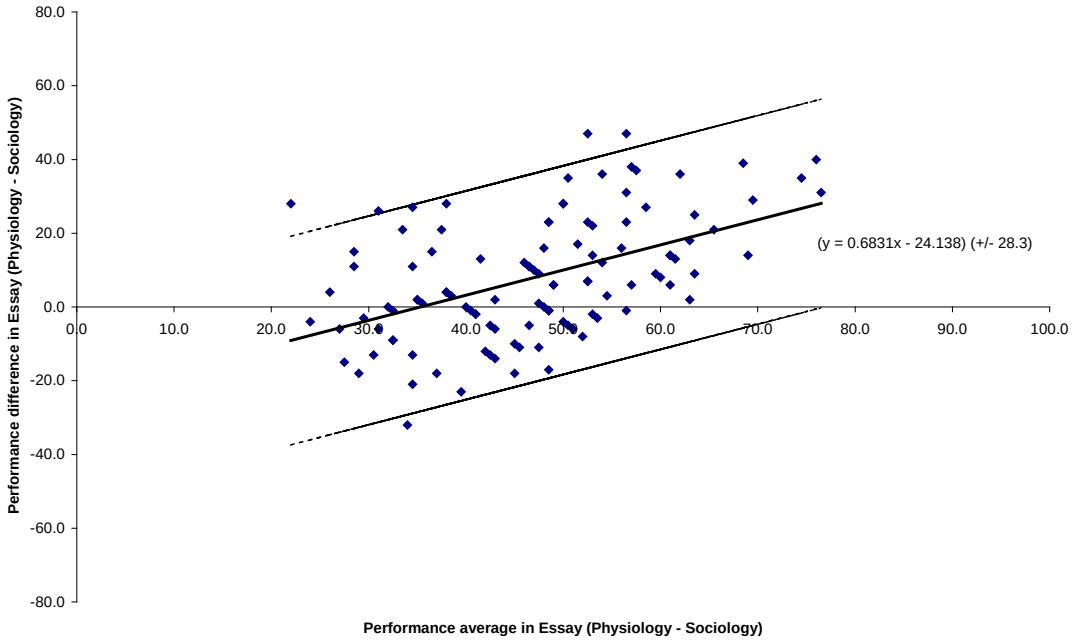

Figure 1: Agreement in student performance in an essay in Physiology and Sociology disciplines

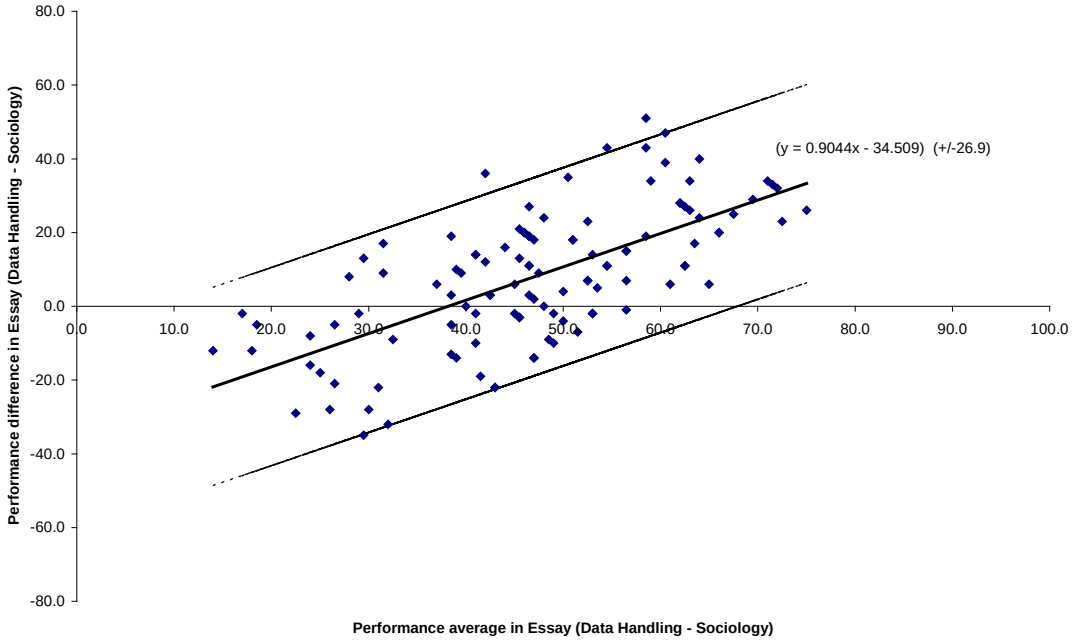

Figure 2: Agreement in student performance in an essay in Data Handling and Sociology disciplines

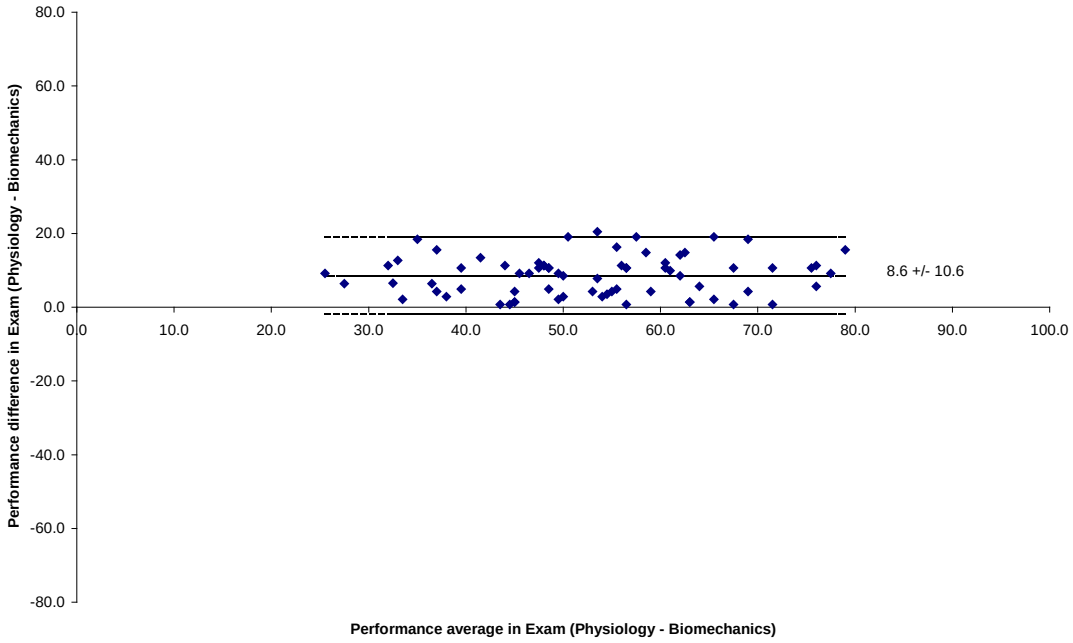

Figure 3: Agreement in student performance in an examination in Physiology and Biomechanics disciplines

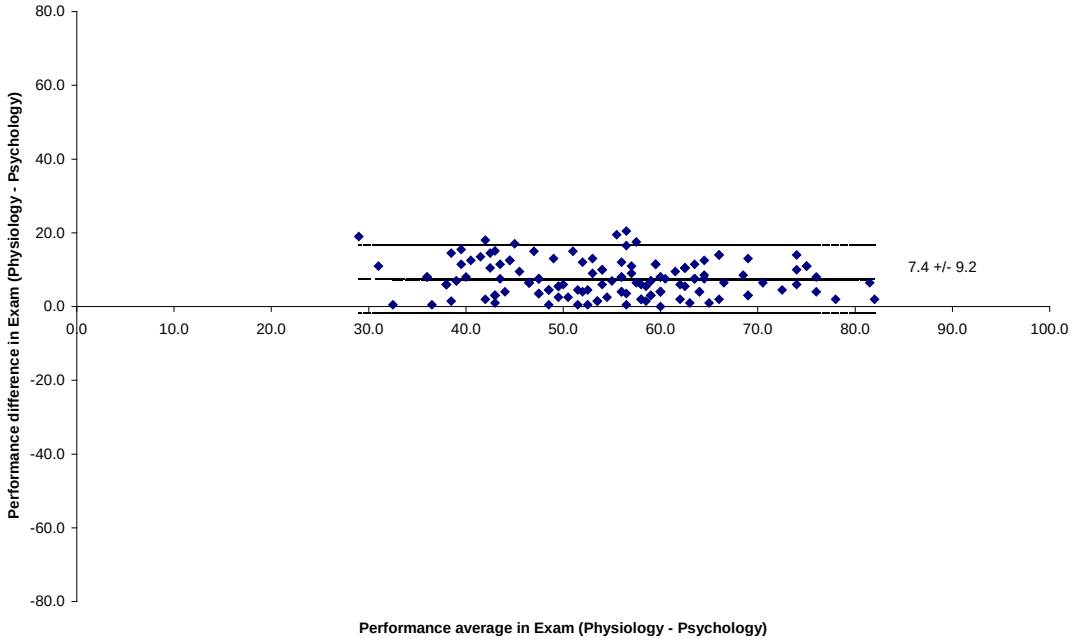

Figure 4: Agreement in student performance in an examination in Physiology and Psychology disciplines