HALOS AND HORNS IN THE ASSESSMENT OF UNDERGRADUATE MEDICAL STUDENTS: A CONSISTENCY-BASED APPROACH[a]

Margaret MACDOUGALL[b]
PhD, Medical Statistician
Community Health Sciences, Public Health Sciences Section,
College of Medicine and Veterinary Medicine, University of Edinburgh
Teviot Place, Edinburgh EH8 9AG, Scotland, UK

E-mail: Margaret.MacDougall@ed.ac.uk

Simon C. RILEY[c]
PhD, Senior Lecturer in Obstetrics and Gynaecology (Non-Clinical)
Centre for Reproductive Biology, Queen's Medical Research Institute,
University of Edinburgh
47 Little France Crescent, Edinburgh EH16 4TJ, Scotland, UK

E-mail: Simon.C.Riley@ed.ac.uk

Helen S. CAMERON[d]
BSc, MBChB, Senior Lecturer and Archie Duncan Fellow in Medical Education
Director Medical Teaching Organisation, College of Medicine and Veterinary Medicine,
University of Edinburgh
Chancellor's Building, 49 Little France Crescent, Edinburgh EH16 4SB, Scotland, UK

E-mail: Helen.Cameron@ed.ac.uk

Brian MCKINSTRY[e]
PhD, Senior Research Fellow
Community Health Sciences, General Practice Section, University of Edinburgh
20 West Richmond Street, Edinburgh EH8 9DX, Scotland, UK

E-mail: Brian.Mckinstry@ed.ac.uk

Abstract: *The authors introduce a consistency-based approach to detecting examiner bias. On comparing intra-class correlation coefficients on transformed data for supervisor continuous performance and report marks (ICC1*) with those for supervisor continuous performance and second marker report marks (ICC2*), a highly significant difference was obtained for both the entire cohort (ICC1* = .72, ICC2* = .30, F = 2.47, p < .0005 (N = 1085)) and the subgroup with high supervisor ratings for continuous performance (ICC1* = .62, ICC2* = .24, F = 1.97, p < .0005 (n = 952)). A strong halo effect was detected and preliminary evidence was obtained for the presence of a strong horn effect for students with lower scores, thus providing a basis for future research.*

Key words: *halo effect; horn effect; intra-class correlation coefficient; second marker; supervisor bias; undergraduate assessment; Zegers-ten Berge general association coefficient*

Introduction

The tendency for good or bad performers over one dimension to deliver consistently good or bad performances overall is already recognized (Dennis 2007, Fisicaro & Lance 1990, Pike 1999, Pulakos et al. 1986). Thus, in an ideal assessment setting where ratings are untainted by examiner bias, one would expect there to be a detectable level of consistency in individual student performance across various assessment dimensions. It is this particular type of consistency, representative of true consistency and hence, illusory bias, which we choose to refer to henceforth in this study as natural consistency.

The need to detect and eliminate examiner bias is clearly a critical one if marks allocated to students are to be representative of performance, particularly in contexts where students are ranked against one another for future selection purposes. Moreover, assessment procedures must be rigorously monitored if the reputational quality of academic programmes is to be maintained and justified. Our specific aim here, therefore, is to introduce new methodology for testing examiner bias where examiners have prior exposure to student performance in one dimension and are required to objectively mark students in a separate but related dimension. Through use of a case study involving undergraduate medical students, this methodology will test for supervisor bias in report marking where supervisors have prior exposure to student continuous performance. The procedure adopted will also explicitly correct for natural consistency as defined above by identifying supervisor bias as that specific contribution to consistency in supervisor ratings across continuous performance and written report performance which is explicitly over and above that of natural consistency. Where this type of bias is found to coincide with the attribution of high or low marks to student assignments, we shall refer to it as a *halo* or *horn effect*, respectively.

Two similar tendencies are apparent in the literature wherever the term 'halo effect' is adopted. The first of these tendencies is a non-prescriptive use of language (as in Wakeford et al. 1995) which suggests that the halo effect is merely the existence of evidence for the rating of one attribute influencing the rating of another. The second, and more common, tendency is to use the term 'halo effect' to refer to a phenomenon akin to *any one* of the two forms of bias considered in this study whilst, with some exceptions (for example, Brown 1965, Pulakos et al. 1986, Fisicaro & Lance 1990), leaving the problem of natural consistency unchallenged.

The latter tendency originates with the inception of the term 'halo effect' to discuss phenomena in measurement data under the auspices of Thorndike (1920); thus those who choose to assume this interpretation (see, for example, Bowden 1933, Anastasi 1988, Fairweather 1988 and Streiner & Norman 2003) may be referred to as his followers. Nevertheless, it makes a great deal of sense to keep the original everyday use of this notion, with its positive connotation, in mind when passing from the material world to the world of measurement theory (Dudycha 1942), we suggest not least because of the greater opportunity this affords to differentiate between different kinds of examiner bias.

The above two generalizing tendencies have the effect that the terms 'horn effect' and 'stigma effect' occur much more rarely in the literature than that of 'halo effect' as their interpretation is already subsumed within the intended notion of halo effect. Nevertheless, confusion can arise in this area too. For example, Marshall (2003) appears to use the terms 'stigma effect' and 'negative stigma' interchangeably to refer to negative bias in examiners where pupils are known to be repeating a grade. Moreover, he omits to provide a definition for either of these terms at the outset and the reader is left to interpret their meaning either implicitly or based on the hidden assumption that their meaning is in some sense obvious. Further, Evans (2002) appears to make a distinction by referring to 'The "halo" effect and the opposing "horns" effect,' but in the absence of any supporting definitions for either of these effects. By contrast, Rubin (1982) uses the term 'horn effect' to refer simply to the tendency to limit the overall assessment of an individual to a single negative attribute.

It is interesting to note, however, that within the context of employee appraisal, Arnold and Pulich (2003) make the interesting distinction between the 'horn' and 'halo' effects, whereby, for example, the horn effect is specifically that 'which occurs when a manager perceives one negative aspect about an employee or his or her performance and generalizes it into an overall poor appraisal rating.'

In seeking to make a similar distinction, the notions of halo and horn effect which we define in this paper (both intuitively and mathematically) are contrary to the two tendencies outlined above. Moreover, these notions make a substantial contribution to addressing Pike's 'critical [problem] for assessment research' (Pike 1999) of differentiating between supervisor bias and true 'regularities' in performance across different dimensions. Our study also benefits from there being a meaningful standard *against which* to measure examiner bias. Precisely, we utilize second marker ratings with second markers having been blinded to the student's identity (and hence their participation in the project) and to the continuous performance rating allocated by their supervisor. As such, our study avoids the potential for uncertainty in other studies (Pulakos et al. 1986, Fisicaro & Lance 1990) wherein correlations across ratings for multiple attributes assigned by expert or trained markers are assumed as surrogates for measures of natural associations (or, associations based on student abilities which are uncontaminated by examiner bias). Moreover, due to constraints on staff time, inclusion of at most a second marker (that is, 'double-marking') is by far a more common choice of assessment regime across different disciplines and places of learning than those involving further markers. Thus, we consider our approach to detecting bias a pragmatic one in so far as, realistically speaking, it may be replicated to test for bias in a wide variety of real-life assessment scenarios.

Method

Background to Participants

Within the 4th year of the undergraduate medical curriculum at the University of Edinburgh, all students are required to identify a supervisor and field of interest to enable them to participate in a 14-week research project known as the 4th year Student Selected Component (SSC4).

During the SSC4 period, the students must prepare a project report, usually in the form of a medical or scientific article of up to 3000 words, which reports on their research findings. The project supervisor allocates a total of two percentage marks to each of their students. The two marks constitute a continuous performance rating measuring overall performance throughout the duration of the project and a report rating measuring the quality of the final written report. The quality of the written report is also allocated a percentage mark by a second examiner with concurrent experience of supervising and marking SSC4 projects within the same student cohort. In their capacity as a second marker, this rater is, however, also blinded to the continuous performance rating allocated to the student concerned and to the identity of that student.

All supervisors are advised to use the same detailed list of performance indicators to assist them in allocating continuous performance ratings to their students. In the allocation of ratings for written reports, all supervisors and second markers are recommended to use a separate comprehensive but shorter list of marking criteria, this list being identical for all markers.

Each of the above three percentage marks is then converted to a grade (A – F), with grades A, B, C, D, E and F corresponding to marks 90 - 100, 80 - 89, 70 - 79, 60 - 69, 50 - 59 (marginal fail) and 0 - 49 (fail), respectively. In the majority of cases, there is no need to call in a third marker to correct for mismatch between supervisor and second marker ratings and the final grade assigned to the student is that obtained from combining the supervisor continuous performance, supervisor report and second marker report ratings.

Whilst continuous performance and report writing are intended to constitute two separate dimensions of SSC4 student performance, that is not to say that student abilities across these two dimensions should differ markedly. Thus, we assumed that there was natural consistency best assessed by the correlation between the supervisor continuous performance mark and the second marker report mark and that supervisor bias could be evaluated by looking for additional consistency between the supervisor continuous performance and report marks. We therefore used intra-class correlation coefficients (ICCs) to assess the evidence that consistency between supervisor continuous performance and written report ratings was significantly greater than that between supervisor continuous performance ratings and the corresponding second marker report ratings.

Data Preparation

All SSC4 continuous performance and report performance data corresponding to the period July 2001 to June 2006 (*N* = 1096) were extracted in an anonymized format from internal undergraduate medical student examination records at the University of Edinburgh and stored in an MS Excel database. Ethical approval to use these data for the current study was formally granted by the University of Edinburgh College of Medicine and Veterinary Medicine Committee on the Use of Student Volunteers.

Statistical Analyses and Underlying Theory

Calculations and data analyses were performed using MS Excel 2003 and the statistical packages Minitab (Version 14.12) and SPSS (Version 14.0).

The model we assumed for this study was a two-way mixed effects model (McGraw & Wong 1996) in which examiners were recognized as fixed effects and students as random effects. In calculating ICCs for consistency rather than absolute agreement, we chose to measure the extent to which corresponding sets of marks agreed according to an additive transformation rather than in absolute terms. Thus, in the notation of Fagot (1993), we used the consistency-based intra-class correlation coefficient *ICC(3,1)* for a two-way mixed model in which raters are fixed and subjects are random.[1]

In testing for a halo effect, two ICCs were calculated over the period 2001 – 2006. The first of these, *ICC1*, measured consistency between supervisor continuous performance and report marks and the second, *ICC2*, measured consistency between supervisor continuous performance and second marker report marks. In our study, these ICCs represent the proportion of the total variance in marks (inclusive of error variance) which can be explained purely in terms of variation between the students in the study. As is well known, ICCs range from -1 to 1. However, within the current context, they are understood to converge towards 1 as the association between the two corresponding sets of marks increases, with negative ICCs indicating the extreme case where on examination of ratings, error variance is greater than that across individual students.

Using the above terminology, in testing for a halo effect, our preliminary null hypothesis was as follows:

$$ICC1 = ICC2. \qquad (1)$$

The hypothesis test which we used was based on the method of Alsawalmeh and Feldt (1994). Alsawalmeh and Feldt already allow for the comparison of two ICCs based on the same sample, although in the absence of any application to educational data or any allowance for the possibility that ratings for different ICCs might violate the assumption of rater independence. Our sample size for subjects was much greater than that assumed by Alsawalmeh and Feldt. We were therefore able to apply the asymptotic properties of the mean square terms to simplify the algebra used in the calculation of the degrees of freedom whilst allowing for the non-independence of raters across *ICC1* and *ICC2*.

Nevertheless, the original requirement of Normality for the Alsawalmeh-Feldt test still required to be met. Thus, we sought an optimal transformation for ensuring that the data for each of supervisor continuous performance mark, supervisor report mark and second examiner report mark approximated to Normality. With the aid of the Box-Cox transformation procedure (Box & Cox 1964), we therefore assumed the polynomial transformation

$$\textit{transformed mark} = (\textit{original mark})^5 \qquad (2)$$

as the single choice of transformation to be applied in each case. Consequently, in practice, it was necessary for us to apply our hypothesis test to refute the null hypothesis,

$$ICC1^* = ICC2^*, \qquad (3)$$

with $ICC1^* = ICC1^5$ and $ICC2^* = ICC2^5$.

In testing the null hypothesis for the transformed data, we used the property (Alsawalmeh & Feldt 1994) that the test statistic $F = \frac{1-ICC2*}{1-ICC1*}$ approximates to a central *F*-distribution with degrees of freedom d_1 and d_2 defined as strictly positive integers in accordance with the method of Satterwaite (1941). One notable impact of our use of the asymptotic properties of the mean square in our adaptation of the hypothesis test for larger samples was that of decreasing the degrees of freedom d_1 and d_2, above for the sample sizes we assumed. This made our test more conservative (with the effect that the probability of a Type I error was reduced).

In order to differentiate between halo and horn effects, we divided the data into two cohorts according to the grades corresponding to the percentage marks for continuous performance assigned by supervisors. Thus, the high grade cohort referred to those percentage marks corresponding to grades A and B, whilst the lower grade cohort referred to those percentage marks corresponding to grades C – F.

Using the raw percentage data, we determined the ICCs and corresponding confidence intervals for both grade cohorts. On the basis of the Box-Cox transformation procedure, we found that the transformation defined under (2) was also the optimal one for Normalization of data for the high grade cohort. On application of this transformation, we tested hypothesis (3) as previously. For the lower grade cohort, on the other hand, it was not possible to find a Normalizing transformation for the data. Thus, in adherence to the assumptions of our hypothesis test, we did not test hypothesis (3) for these data.

For each application of our hypothesis test, we assumed a significance level of .05.

In interpreting our choice of ICC as a measure of examiner consistency, it is useful to consider Zegers and ten Berge's notion of a general association coefficient (Zegers & ten Berge 1985). The latter coefficient was designed to measure the level of absolute agreement between two variables in terms of the mean squared distance once each of these two variables has undergone a specific admissible transformation (ibid.) in accordance with the type of data under consideration. Later, Stine (1989) coined the useful term 'relational agreement' rather than 'association' to refer to the type of measurement represented by Zegers and ten Berge's coefficient. In adopting this term, Stine recognized absolute agreement under the *identity* transformation to be the strictest of a *family* of possible types of agreement which are meaningful in a measurement theoretic sense, the appropriate transformation being dependent on the particular measurement scale represented by the data.

Fagot (1993) has already established a useful identity between a particular case of the Zegers-ten Berge general association coefficient and *ICC(3,1)* for continuous ratings when they are understood to be representative of Normally distributed data on an additive scale. In particular, for a study involving k examiners and N subjects, let X_i denote the variable ranging over all N ratings for examiner i (i = 1, 2, …k), $\overline{X}_i$ denote the arithmetic mean of all ratings for examiner i and let V_i be defined according to the admissible transformation $V_i = X_i - \overline{X}_i$ (i = 1, 2, …k). Then the general association coefficient for the transformed variables is precisely equal to *ICC(3,1)* for the corresponding untransformed variables.

This result is particularly useful because it informs us that, within the context of our study in which two sets of ratings are being compared at any one time, *ICC(3,1)* is a measure of the extent to which the distribution of the marks about the mean for one set of data is the same as that for the other. For the case in which two sets of marks are being compared at any one time, this interpretation of relational agreement can be understood graphically in terms of the degree of scatter of the data points (V_1, V_2) about the line $V_2 = V_1$. Moreover, as any one of our *ICC1** and *ICC2** approaches 1, the two corresponding sets of marks should tend towards perfect agreement in the above sense.

On applying the above admissible transformation to supervisor and second marker Normalized ratings, we therefore used scatter plots to address the challenge of providing a visual representation of the contrasting relationships between supervisor continuous performance and report marks and supervisor continuous performance and second marker report marks which had previously come to light by means of the ICCs. We carried out this procedure separately for the data in its entirety and for the high grade cohort but not for the lower grade cohort, on account of the absence of a suitable Normalizing transformation for the corresponding data.

Results

The ICCs used to assess examiner bias together with their corresponding 95% CIs are provided in Table 1 both for the raw data and for the transformed data, where appropriate.

Table 1. ICC-Based Consistency Between a) Supervisor Continuous Performance Mark and Supervisor Report Mark (*ICC1* and *ICC1**) and b) Supervisor Continuous Performance Mark and Second Marker Report Mark (*ICC2* and *ICC2**)

Grade cohort	*ICC1* (95% CI)	*ICC1** (95% CI)	Grade cohort	*ICC2* (95% CI)	*ICC2** (95% CI)
All grades (*N* = 1085)[a]	.76 (.74, .79)	.72 (.69, .75)	All grades (*N* = 1085)[a]	.33 (.28, .38)	.30 (.25, .36)
High grades: A - B (*n* = 952)	.59 (.55, .63)	.62 (.58, .65)	High grades: A - B (*n* = 952)	.22 (.16, .28)	.24 (.18, .30)
Lower grades: C - F (*n* = 133)	.72 (.63, .80)		Lower grades: C - F (*n* = 133)	.42 (.27, .55)	

Note. ICC = intra-class correlation coefficient

[a] All ICCs were calculated only for those students for whom all three percentage marks, corresponding to supervisor continuous performance and supervisor and second marker report ratings, were available. *ICC1** and *ICC2** denote the consistency measures for the data further to the transformation defined under (2), above. Marks were incomplete for 11 out of 1096 (1.0%) of the students within the 2001 - 2006 dataset.

On testing hypothesis (3) for the data in their entirety and in particular, for the high grade cohort, a highly significant difference was found between *ICC1** and *ICC2** in each case (F= 2.47, p < .0005 (N = 1085), and F = 1.97, p < .0005 (n = 952), respectively).

The relationships between supervisor continuous performance and report marks and supervisor continuous performance and second marker report marks are represented in Figure 1 for all of the data and separately for those data corresponding only to students who received high grades for continuous performance.

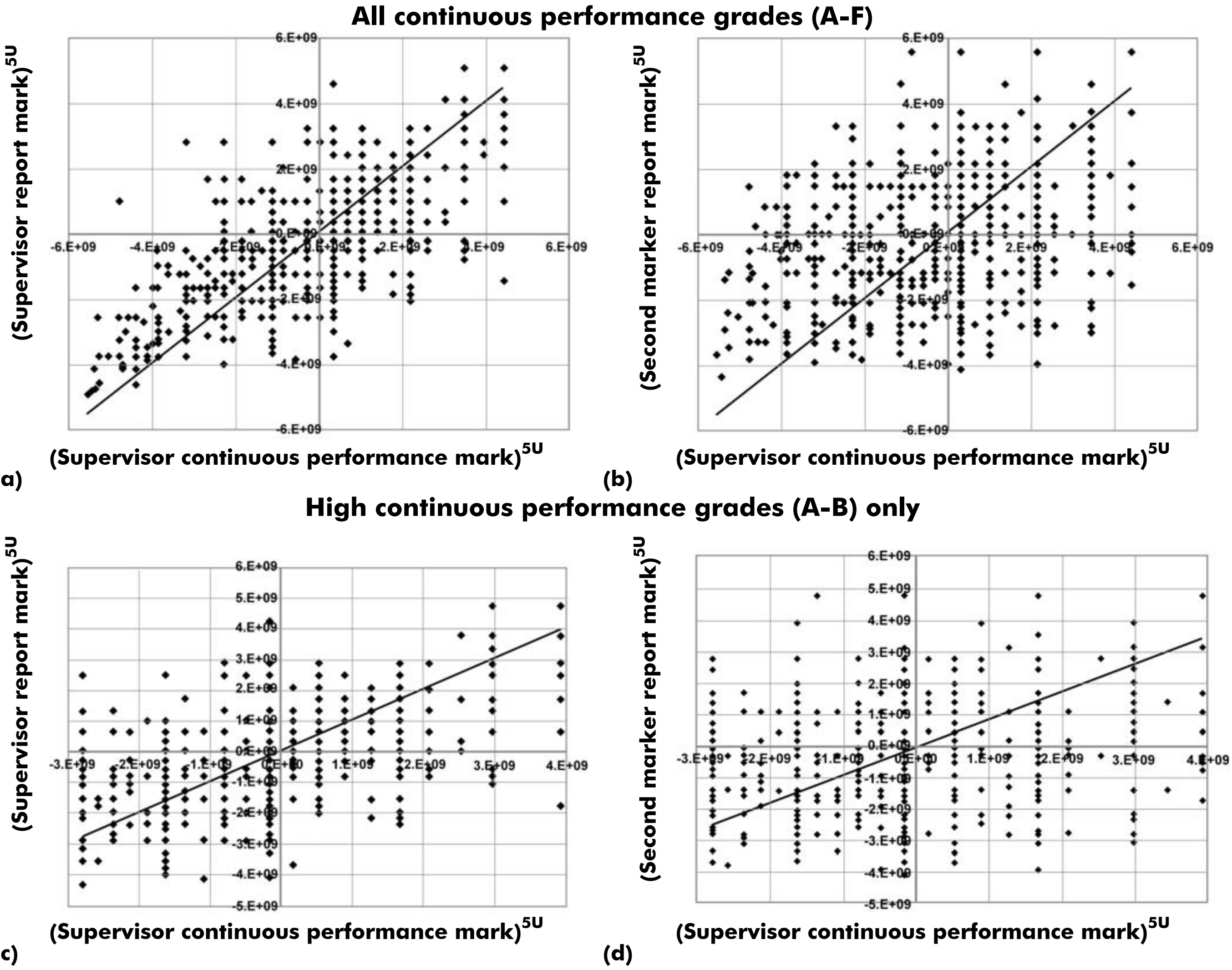

Figure 1. Relationship between continuous performance marks and report marks relative to the 45^0 line of perfect agreement through the origin following Normalization and subsequent application of the Zegers-ten Berge uniforming transformation $V_i = X_i - \overline{X}_i$, where X_i ranges over all ratings for a given type of measurement *i* and $\overline{X}_i$ denotes the arithmetic mean of all ratings for measurement type *i* (*i* = 1, 2).

Note. The notation '(Supervisor continuous performance mark)^{5U}', '(Supervisor report mark)^{5U}' and '(Second marker report mark)^{5U}' is used here to denote that the expression in brackets has been transformed, first through Normalization by exponentiation to the power 5 and subsequently through application of the above Zegers-ten Berge uniforming transformation.

Discussion

Having tested hypothesis (3) for our data, according to our definitions, there is extremely compelling evidence for the existence of supervisor bias and more specifically, for the existence of a strong halo effect in supervisor assessment of SSC4 reports.

The result (Table 1) that none of the ICCs for the grade subgroups attain or exceed the corresponding values for the entire cohort is an inevitable consequence of the increase in the ratio of error variance to true variance across students, which occurs when sample size is reduced. Nevertheless, the ICCs for the subgroups can be considered in their own right, together with their corresponding confidence intervals. The failure to Normalize the data for the lower grade cohort was undoubtedly influenced by the relatively small cohort size (n = 133). Even in the presence of a suitable transformation, it is doubtful that n would have been sufficiently large here to satisfy the underlying asymptotic assumptions of our hypothesis test.

On comparing *ICC1* and *ICC2* with *ICC1** and *ICC2**, respectively in Table 1, it is clear that the Normalizing transformation has had very little impact on the level of consistency as represented by these indices. The transformation is also appreciably conservative of the original confidence intervals. These observations support the testing of hypothesis (3) as a surrogate for (1) in satisfying the requirements of our *F*-test. Further, they are supportive of the meaningfulness of the idea of comparing the untransformed values of the ICCs for the lower grade cohort with a view to finding preliminary evidence for the existence of a horn effect. Notice in particular that for this cohort the ICC for consistency between first examiner continuous performance and report marks (.72) is a great deal higher than that for consistency between first examiner continuous performance and second examiner report marks (.42). This discrepancy in ICCs is of the same order of magnitude as that for the corresponding ICCs for each of the complete cohort and high grade cohort prior to and subsequent to transformation, suggesting the possibility of a strong horn effect.

On moving from part a) to b) and from part c) to d) of Figure 1, an increase in the visual spread of the data about the 45^0 line is recognizable in each case, indicating a tendency for greater agreement in dispersion from the mean when comparing supervisor continuous performance and report marks than when comparing supervisor continuous performance and second marker report marks. These findings are consistent with those which would be expected on examination of the corresponding raw and transformed ICCs in Table 1. However, the more rigorous analysis afforded by hypothesis testing serves to provide a more precise indication of the level of supervisor bias suggested by these discrepancies. For example, given the large sample sizes considered in each case, differences in spread between corresponding figures are masked through overlapping of multiple points and it is difficult to overcome this effect, even through jittering.

Our use of ICCs and the corresponding graphical representation of varying levels of relative agreement illustrated in Figure 1 ought to be distinguished from efforts based on the Pearson Correlation Coefficient to establish the level of conformity of data to merely any straight line. It is already recognized (Streiner & Norman 1985) that the latter coefficient has a tendency to inflate true agreement levels.

Our findings suggest that the report mark allocated by SSC4 supervisors is not purely based on written performance but that prior knowledge of continuous performance has a highly significant role to play. They are also supportive of the more general view that the provision of detailed descriptors does not suffice to remove examiner bias. As has been observed elsewhere (Eric et al. 1998), in improving the reliability of assessment procedures, there is the additional challenge of the successful training of examiners in the use of these descriptors.

Thus, whilst it is has been traditionally assumed (Evans 2002) that the provision of detailed objective descriptors counters examiner bias, much more may need to be done to make this type of intervention sufficiently effective.

We acknowledge that in being blinded to student continuous performance, second examiners may be restricted in terms of their knowledge of the subject matter of the student projects they are marking. This could have led to an attenuation of the values of the transformed and untransformed ICCs which we used to represent natural consistency in this study and an inflation of the corresponding measures of halo and horn effects. Nevertheless, given that second markers are usually selected for their expertise in the field of study covered by the project reports which they mark, we assume here that the above confounding effect on level of supervisor bias is minimal.

Future research

It is very clear from this study and from our ongoing work with assessment data that there is a tendency for continuous performance marks for SSC4 students to be heavily skewed towards those representative of high grades. Such behaviour in assessment data is not unique to SSC4 data (see, for example, Dennis 2007 and Phelps et al. 1986). Additionally, whilst we have benefited from the availability of assessment data on a continuous scale and a successful Normalizing transformation as a means of ensuring that the assumptions of our hypothesis test have been satisfied, these conditions are not guaranteed within the context of the analysis of assessment data in general. In particular, successful Normalizing transformations may not be forthcoming or studies may be limited to the consideration of Likert scale data.

In future work, therefore, we anticipate using bootstrap sampling on existing educational data to assess the robustness of our hypothesis test to Type I errors following departures from Normality, in a manner akin to Hsu and Feldt (1969). This work would prove particularly valuable where the intention is to find a reliable procedure for testing for a horn effect with smaller sample sizes. Furthermore, such testing should be extended to the consideration of Likert scale data. Investigations of these types would have applications not only within the context of the current study but wherever it is of interest to quantify agreement for non-parametric assessment data.

Bibliography

1. Alsawalmeh, Y.M. and Feldt, L.S. **Testing the Equality of Two Related Intraclass Reliability Coefficients,** Applied Psychological Measurement 18, no. 2**,** 1994, pp. 183-190
2. Anastasi, A. **Psychological Testing,** 6th edn, New York, MacMillan Publishing Company, 1998
3. Arnold, E. and Pulich, M. **Personality Conflicts and Objectivity in Appraising Performance, Health Care Manager 22,** no. 3, 2003, pp. 227-232
4. Bowden, A.O., Caldwell, F.F. and West, G.A. **Halo Prestige,** Journal of Abnormal and Social Psychology 28, no. 4, 1934, pp. 400-406
5. Brown, E.M. **Influence of Training, Method and Relationship on the Halo Effect,** Journal of Applied Psychology 52, no. 3, 1965, pp. 195-199
6. Dennis, I. **Halo Effects in Grading Student Projects,** Journal of Applied Psychology 92, no. 4, 2007, pp. 1169-1176
7. Dennis, I., Newstead, S.E. and Wright, D.E. **A New Approach to Exploring Biases in Educational Assessment,** British Journal of Psychology 87, 1996, pp. 515-534

8. Dudycha, G.J. **A Note on the "Halo Effect" in Ratings,** Journal of Social Psychology; Political, Racial and Differential Psychology 15, Short Articles and Notes, 1942, pp. 331-333
9. Evans, A.W. **Facing the Challenges of Competency-Based Assessment of Postgraduate Dental Training,** Medical Education 36, Letters to the Editor, 2002, p. 586
10. Fagot, R.F. **A Generalized Family of Coefficients of Relational Agreement for Numerical Scales,** Psychometrika 58, no. 2, 1993, pp. 357-370
11. Fairweather, J.S. **Reputational Quality of Academic Programs: The Institutional Halo,** Research in Higher Education 28, no. 4, 1988, pp. 345-355
12. Fisicaro, S.A. and Lance, C.E. **Implications of Three Causal Models for the Measurement of Halo Error,** Applied Psychological Measurement 14**,** 1990, pp. 419-429
13. Holomboe, E.S. and Hawkins, R.E. **Methods for Evaluating the Clinical Competence of Residents in Internal Medicine: A Review,** Annals of Internal Medicine 129, no. 1, 1998, pp. 42-48
14. Hsu, T.-C. and Feldt, L.S. **The Effect of Limitations on the Number of Criterion Score Values on the Significance Level of the F-test,** American Educational Research Journal 6, no. 4, 1969, pp. 515-527
15. Marshall, J.H. **Grade Repetition in Honduran Primary Schools,** International Journal of Educational Development 23, 2003, pp. 591-605
16. McGraw, K.O. and Wong, S.P. **Forming Inferences about Some Intraclass Correlation Coefficients,** Psychological Methods 1, no. 1, 1996, pp. 30-46
17. McKinstry, B.H., Cameron, H.S., Elton, R.A. and Riley, S.C. **Leniency and Halo Effects in Marking Undergraduate Short Research Projects,** BMC Medical Education 4, no. 28, 2004, pp. 1-5
18. Phelps, L.A., Schmitz, C.D. and Boatright, B. **The Effects of Halo and Leniency on Cooperating Teacher Reports Using Likert-Type Rating Scales,** Journal of Educational Research 79, no. 3, 1986, pp. 151-154
19. Pike, G.R. **The Relationship between Perceived Learning and Satisfaction with College: An Alternative View,** Research in Higher Education 34, no. 1, 1993, pp. 23-40
20. Pike, G.R. **The Constant Error of the Halo in Educational Outcomes Research,** Research in Higher Education 40, no. 1, 1999, pp. 61-86
21. Pulakos, E.D., Schmidt, N. and Ostroff, C. **A Warning about the Use of a Standard Deviation Across Dimensions Within Ratees to Measure Halo,** Journal of Applied Psychology 71, no. 1, 1986, pp. 29–32
22. Ram, P., Grol, R., Rethans, J.J., Schouten, B., van der Vleuten, G. and Kester, A. **Assessment of General Practitioners by Video Observation of Communicative and Medical Performance in Daily Practice: Issues of Validity, Reliability and Feasibility,** Medical Education 33, 1993, pp. 447-454
23. Rubin, S. **Performance Appraisal: a Guide to Better Supervisor Evaluation Processes,** Panel Resource Paper No. 7, in "National Society for Internships and Experiential Education" Washington, DC US, 1982
24. Satterwaite, F.E. **Synthesis of Variance,** Psychometrika, 6, 1941, pp. 309-316
25. Spike, N., Alexander, H., Elliot, S., Hazlett, C., Kilminster, S., Prideaux, D. and Roberts, T. **In Training Assessment – Its Potential in Enhancing Clinical Teaching,** Medical Education 34, 2000, pp. 858-861
26. Stine, W.W. **Influence of Training, Method and Relationship on the Halo Effect,** Journal of Applied Psychology 52, no. 3, 1965, pp. 195-199
27. Streiner, D.L. and Norman, G.R. **Health Measurement Scales: A Practical Guide to Their Development and Use,** 3rd edn., Oxford University Press, 2003.
28. Thorndike, E.L. **A Constant Error in Psychological Ratings,** Journal of Applied Psychology 4, 1920, pp. 25-29

29. Wakeford, R., Southgate, L. and Wass, V. **Improving Oral Examinations: Selecting, Training, and Monitoring Examiners for the MRCGP,** BMJ, 311, Education and Debate**,** 1995, pp. 931-935

30. Zegers, F.E. and ten Berge, J.M.F. **A Family of Association Coefficients for Metric Scales,** Pyschometrika 50, no. 1, 1985, pp. 17-24

[a] **Acknowledgements**
We express our gratitude to Ronnie Grant, Senior Graphic and Web Designer, University of Edinburgh for his dedication, expertise and patience in assisting with the preparation of the figures for this paper.

[b] **Dr Margaret MacDougall, Bsc (Hons), PGCE, PhD, FRSS University of Edinburgh**
Personal web page: http://www.chs.med.ed.ac.uk/people/staffProfile.php?profile=mmacdoug
Margaret MacDougall is the statistician at the University of Edinburgh responsible for providing advice on statistical design and analysis to undergraduate medical students involved in 4th year Student Selected Component (SSC4) projects. In 2004, she helped design the 3rd year EBM station for Edinburgh's first OSCA (Online System for Clinical Assessment). She is a tutor on the international Msc/Dip/Cert Course in Pain Management, which is run by the University of Edinburgh in partnership with the University of Sydney, Australia.
At an external level, she served for 2 consecutive years (2005 – 2006) as Module Organizer and Lecturer for the Statistics and Clinical Epidemiology module of the Royal College of Physicians of Edinburgh Scottish Oncology Course for SpRs. Her principal area of research is Medical Education. She has been awarded three consecutive one-year grants from the University of Edinburgh Principal's e-learning Fund.
Recently, she was awarded funding by the Maths, Stats & OR Network to lead a project entitled 'Statistics in medicine: a risky business?' and funding by the MEDEV subject centre of the Higher Education Academy to lead a project entitled 'Research-based SSCs: a pragmatic approach to advancing the research-teaching nexus' (http://www.medev.ac.uk/resources/funded_projects/show_mini_project?reference_number=549).

[c] **Dr Simon C Riley, BSc (Hons), PhD, FHEA, MRCOG, University of Edinburgh**
Personal web page:
http://www.rds.mvm.ed.ac.uk/RDS%20Team%20Pages/Simon%20Riley%20Main%20and%20Linked%20Page/Simon%20Riley%20Linked%20Page.html#current
Simon Riley is a Senior Lecturer in Obstetrics and Gynaecology. He is also Programme Director for the undergraduate self-selected student projects, known as Student Selected Components (SSCs), within the undergraduate medical curriculum at the University of Edinburgh.
He has been involved in the development of SSCs since their inception by the General Medical Council in 1993, when they were previously referred to as Special Study Modules (SSMs). He is Convenor for the SSC Directors Liaison Sub-Group of the Scottish Teaching Deans' Scottish Doctor Medical Curriculum Group (http://www.scottishdoctor.org) and External Examiner for Student SSCs for the Undergraduate programme in Medicine at Queen's University, Belfast. He is also a Fellow of the Higher Education Academy.
His research interests in Medical Education include the effective assessment of professional and generic skills and attitudes of students performing self-selected research projects across a wide range of specialties within medicine. He is also interested in how well medical students use the optional components of their curriculum to optimize their own personal professional development in terms of research and professional skills and as a means of gaining a better insight into future career options. His research activities also involve collaboration as a Project Partner in the Higher Education Academy funded project 'Research-based SSCs: a pragmatic approach to advancing the research-teaching nexus'
(http://www.medev.ac.uk/resources/funded_projects/show_mini_project?reference_number=549).

[d] **Dr Helen S Cameron BSc (Med Sci) MBChB MRCP FHEA**
Helen Cameron has a clinical background but since 1994, the focus of her work has been in medical education. Since 2003, she has been Director of the Medical Teaching Organisation (MTO), the centre for educational innovation, development and research in the Edinburgh Medical School.
Her principal academic interest is the investigation of assessments to inform the development of the assessment strategy for the undergraduate medical curriculum. Key emphases include maximising the validity and reliability of assessment and developing students' learning through good feedback and self-appraisal. Her other interests within medical education include developing a portfolio approach to learning and assessment within the undergraduate curriculum, and exploring the use of technology to develop and adapt teaching and assessment to the imperatives of today's healthcare. The latter has involved major contributions to online teaching materials and an online, media-rich system for clinical assessment.
She has published and presented papers across a range of topics in medical education, including *supervisor leniency in assessment*, *comparison of methods for standard-setting*, *development and assessment of clinical skills*, *portfolio learning and assessment*, *use of patient recordings and images in medical education*, *peer-assisted learning*, and *learning outcomes in medical education*.
She is a member of the Scottish Deans Medical Education Group, which collaborates on projects to develop medical education across Scotland, and is a Foundation member of the Academy of Medical Educators. She is also a Fellow of the Higher Education Academy. As a member of the Steering Group of the Tuning Workforce for the EU funded MEDINE Network (Medical Education in Europe) she helped develop the project which produced a consensus on the

learning outcomes for a medical graduate in Europe. She is also reviewer for the academic journals *Medical Teacher* and *Medical Education*.

[e] **Dr Brian McKinstry, MBChB, MRCPUK, MRCGP (Dist), MD, FRCGP, FRCPE, University of Edinburgh**
Personal web page: http://www.chs.med.ed.ac.uk/people/staffProfile.php?profile=bmckinst
Brian McKinstry is Reader in General Practice at the University of Edinburgh a General Practitioner and holds a Fellowship from the Chief Scientist Office of the Scottish government. His main interests are in tele-health and medical education.

[1] This notation is derived from Table 4 of McGraw and Wong 1996, where a two-way mixed effects model is specified as Case 3 and the additional '1' was introduced in order to distinguish from the case where agreement is being assessed for ratings taken as averages over c independent ratings ($c > 1$). The formula for *ICC(3,1)* is also available from the last row of this table as a special case of formulae, *ICC(C,1)*, for ICCs based on consistency between measurements.