Investigations
in university teaching and learning vol. 4 (1) autumn 2006 ISSN 1740-5106

Reducing variability, increasing reliability: exploring the psychology of intra- and inter-rater reliability.

Dr Helen J Aslett
Department of Psychology
London Metropolitan University

Keywords: *examiner reliability, assessment, moderation, second-marking*

Introduction

Reliability relates to the fairness and consistency of assessment. Section 7 of the Quality Assurance Agency (QAA) Code of Practice (2000) requires that: *"Institutions have transparent and fair mechanisms for marking and moderating marks"*. At an institutional level the London Metropolitan Assessment Framework states: *"There should be consistency among assessors in the marking of student work against relevant criteria"* (Section A2:2). Therefore it is vital that methods of assessment have strong reliability. However, the reliability of the assessment process should never be assumed.

There are two main forms of reliability: intra- and inter- rater reliability. Intra-rater reliability is the internal consistency of an individual marker. Inter-rater reliability is the consistency between two or more markers. The former should perhaps be considered the more important of the two as without internal consistency over a series of scripts the marks assigned will be haphazard and unjustifiable and no form of moderation or second marking will be able to resolve this. In this paper some of the key psychological variables that can potential impinge on examiner reliability will be examined.

Psychological variables affecting examiner reliability

From a psychological perspective reliability is underpinned by cognitive (fatigue and concentration), emotional (knowledge of student) and behavioural (marker stringency) influences, whilst these influences can act on their own they are also heavily inter-related.

Cognitive factors

Fatigue, either mental (lack of interest/repetition) or physical (lack of sleep), has been found to significantly affect the reliability of the marks assigned by an individual assessor. Mental fatigue due to monotony and lack of interest in a task can have severe implications with regards to task performance and accuracy. From a physical

perspective, lack of sleep, whether sleep deprivation or fractal sleep disturbance can lead to lassitude affecting vigilance, attention, logical reasoning, and rational thinking (Akerstedt, 1988; Akerstedt & Gillberg, 1990; Akerstedt, Torsvall & Gillberg, 1989; Durmer & Dinges, 2005; Torsvall & Akerstedt, 1987; 1988a). Wolfe et al. (1999) coined the term DRIFT (Differential Rater Functioning over Time) to describe the process of how marking accuracy of a single assessor decreases over time due to fatigue and lack of attentional control. As a consequence of DRIFT earlier marked answers have been found to receive significantly different marks to later marked answers. Klein & El (2003) found that earlier marked papers were assigned significantly lower marks than later marked papers.

Whilst sleep deprivation has been found to have a significant effect on motor and cognitive task performance, it is subjective mood (whether the individual perceives themselves to be tired or not) which has the greatest influence on task performance (Pilcher & Huffcutt 1996).

Emotional factors

The subjectivity of the above observation highlights the degree of interplay between emotion and cognition and the implicit influences that can affect how a task is carried out and indeed how subjective variables may affect examiner reliability. This is most apparent in situations where assessors know the identity of the student whose work they are marking. Whilst an assessor would hope to remain as objective as possible throughout the assessment process, where a marker is aware of a student's identity, their marking can potentially be profoundly affected.

The examiner may not be consciously aware of their marks being biased, but in knowing the student's identity they may be implicitly positively or negatively primed towards or against the student: the halo/horn effect (Wells, 1907; Nisbett & Wilson, 1977). The halo effect is where the marker has positive expectations about the student e.g. "X has always been an "A" grade student". Thus if the examiner then reads X's script and it is not up to their usual standard they may make allowances based on their subconscious profile and beliefs of that person (Thorndike, 1920). By contrast if the assessor is aware that Y usually produces shoddy work, then an equivalent piece of work to what X submitted is likely to be marked more harshly (the horn effect). Moreover, if Y produces good work it may not be given the full credit it deserves.

Emotional biases in relation to familiarity of students through the supervision of project work may potentially impinge on intra-rater reliability. Dennis, Newstead & Wright (1996) found that 1st markers of final year psychology projects who were also the supervisors of the projects, were biased by the amount of time and effort individual students were perceived to put into their project, rather than solely appraising the submitted piece of work on its scholarship. However, there are some instances where personal knowledge of an individual student has actually been found to reduce marking bias and increase intra-rater reliability. Bradley (1984) found that

personal knowledge of an individual student reduced the sex bias of first markers; however, second markers were found to show a marking bias towards male students placing them at the extreme ends of the marking continuum (1sts/fails) whereas females were placed more tightly in the centre of the distribution band of the first markers.

Behavioural Factors

Behavioural Factors are undeniably linked to both cognitive and emotional variables surrounding examiner reliability. The most common expression of behavioural factors impinging upon examiner reliability is in terms of the relative stringency/leniency of assessors. Spear (1997) found that examiners over-inflate the grades of good work when it follows a poor quality submission and are unreasonably harsh on grading a poor piece of work following a poor submission, thus leading to potential intra-rater reliability bias. Weigle (1998) and Ruth & Murphy (1988) both observed that inexperienced markers were more stringent than experienced assessors (see also Greatorex & Bell, 2004) thus creating inter-rater reliability bias. The reasons for this disparity are unclear; however, possible factors may include novice markers being more "rule –based", more deliberative, more observant of the assessment criteria and taking more time in their marking (Ecclestone, 2001).

Ecclestone (2001) found novice markers to be more accurate compared to experienced markers who placed greater importance on their intuition. Ecclestone (2001) suggests the attitudes of experienced markers are imbedded so deeply within the experienced assessor that they are not able to articulate their reasons for assigning a particular mark as their reasoning moves from concrete to abstract over time with increased experience (Tulving, 1972). Where assessors have been re-trained in their marking processes whilst their consistency of marking is increased, and the use of marks in the extreme bands of grades is reduced both stringency and leniency remain constant over time (Weigle, 1998; Lunz & O'Neill, 1997).

Discrepancies between markers however may not simply be a result of some markers being more stringent than others, they may be a consequence of poorly designed assessment criteria. Elander & Hardman (2004) observed that first markers in Psychology address more aspects of the assessment criteria such as understanding, conveyance of information, development of argument, structure and clarity, when assessing coursework, whilst second markers use a more limited frame of reference. They suggest that this is because first markers tend to be the course organisers and are thus more familiar with the demands and expectations of the course whereas second markers only have general knowledge of the subject area which they are marking.

Can the psychological biases affecting reliability can be minimized?

It is clearly desirable that as markers we are aware of and are able to minimize the effects that psychological biases may bring to bear upon the marking process. Issues surrounding the management of cognitive biases such as fatigue are the easiest to address. It goes without saying that taking regular breaks and not marking when already tired are vitally important points to bear in mind. Revisiting earlier marked scripts and reviewing scripts marked at the end of any marking session is also essential. Marking question by question rather than script by script may also reduce some elements of fatigue as it minimizes cognitive load and enables the marker to get into the mindset of the question.

In terms of reducing examiner bias where project work is being assessed, unless assessment criteria makes specific recommendations for the effort each student has put in to producing the coursework we are left with the possibility that some students are potentially given more credit for effort than for the actual quality of work produced. Whilst we should applaud the effort students put in overcoming hurdles and obstacles, we need to ask ourselves whether at university level we should be awarding effort or academic ability. Given that projects tend to be original pieces of work supervised by the people likely to first mark them anonymous marking is going to be ineffectual in reducing possible bias. One possible solution is to have first markers who were not the supervisor and hence could be anonymous, but for the supervisor to either be a second marker or moderator. This is currently the practice in only a minority of departments at London Met.

The relative stringency of markers highlights the importance of monitoring over second marking in the assessment process. Whilst second marking is essentially double marking of scripts followed by discussions between colleagues to try and agree marks - with different levels of effectiveness - it is renowned for its inconsistency (Edgeworth, 1890; Diedrich, 1957; Laming, 1990). By contrast monitoring looks at patterns of disagreement between the first assessor and a moderator who is responsible for examining a sample of scripts from across the marking bands (White, 2001). The moderator has the power to either raise or reduce all grades if they feel that the first marker has been unduly harsh or lenient. Where marking is seen as inconsistent it is referred to an external examiner or third internal marker. From moderation it is possible to establish the behavioural patterns of individual markers, second marking does not allow for this.

Clearly a way of improving reliability between first and second markers would be to provide adequate training to staff likely to be involved in the marking of a particular module. Assessment criteria and marking schemes should be well defined and explicit to all assessors. As a matter of course the module leader should be responsible for moderating the first 5-10% exam scripts marked by any team member prior to marking proceeding.

Conclusion

This paper highlights some of the many psychological and behavioural factors affecting intra- and inter-rater reliability. These include cognitive factors such as fatigue, emotional factors such as personal knowledge of students biasing grades, and the relative stringency and leniency of markers. Ways of reducing such biases have been suggested with an emphasis on less reliance upon double marking, which in itself is an unreliable and inconsistent approach, and a move towards moderation and monitoring.

References

Akerstedt, T. (1988). Sleepiness as a consequence of shift work. *Sleep*,11: 17–34.

Akerstedt, T. & Gillberg, M. (1990)Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience* 52(1–2):29–37.

Akerstedt, T., Torsvall, L. & Gillberg, M. (1989). Shift work and napping. In: Dinges, DF, Broughton, RJ, editors. *Sleep and alertness: chronobiological, behavioral, and medical aspects of napping*. New York: RavenPress; p. 205–20.

Bradley, C. (1984). Sex bias in the evaluation of students. *British Journal of Social Psychology*, 23, 147-163.

Deidrich, P. (1957). The improvement of essay examinations. Princeton: Educational Testing Service.

Dennis, L., Newstead, S.E., & Wright, D.E. (1996). A new approach to exploring biases in educational assessment. *British Journal of Psychology*, 87, 515-534.

Durmer, J. S. & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in Neurology*, 25:117–29.

Ecclestone, K. (2001). "I know a 2:1 when I see it": understanding degree standards in programmes franchised to colleges. *Journal of Further and Higher Education*, 25, 301-313.

Edgeworth, F.Y. (1890). The elements of chance in competitive examinations. *Journal of the Royal Society*, 400-75.

Elander, J. & Hardman, D. (2002). An application of judgement analysis to examination marking in Psychology, *British Journal of Psychology*, 93, 303-328.

Greatorex, J. & Bell, J. F. (2004). Does the gender of examiners influence their marking? *Research in Education,* 71, 25-36

Huck, S.W. & Bounds, W.G. (1972). Essay grades; an interaction between graders handwriting clarity and the neatness of examination papers, *American Educational Research Journal*, 9, 279-283.

Klein, J. & El, L.P. (2003). Impairment of teacher efficiency during exended sessions of test correction, *European Journal of Teacher Education*, 26 (3) 379-392.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *Quarterly Journal of Experimental Psychology*, 42a, 239-254.

London Metropolitan University (2004) University Assessment Framework

Lunz, M.E. & O'Neill, T.R.(1997). "A longitudinal study of judge leniency and consistency". Paper presented at the annual meeting of the American Educational research Association, Chicago.

Nisbett, R.E., and T.D. Wilson. (1977). The halo effect: Evidence for the unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35, 450-456.

Pilcher, J. & Huffcutt A. (1996). Effects of sleep deprivation on performance: a metaanalysis, *Sleep*, 19, 318–26 .

Quality Assurance Agency (QAA) for Higher Education (2000) Code of practice for the assurance of academic quality and standards in higher education. Gloucester. QAAHE.

Ruth, L. & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*, Norwood NJ: Ablex

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research*, 39, 229-233.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.

Torsval,l L. & Akerstedt, T. (1987). Sleepiness on the job: continuously measured EEG changes in train drivers. *Electroenceph Clin Neurophysiology*, 66(6), 502–11.

Torsvall, L. & Akerstedt, T. (1988a). Disturbed sleep while being on call: an EEG study of ships' engineers. *Sleep*, 11, 35–8.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.) *Organisation of memory*. New York: Academic Press.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263-287.

Wells (1907) cited by Pike, G. (1999) The Constant Error of the Halo in Educational Outcomes Research, *Journal in Higher Education*, 40, 1, 61-86.

White, R. (2001). Double marking versus monitoring of examinations. *PRS-LTSN Journal* 1(1), 52-60.

Wolfe, E.W., Moulder, B.C. & Myford, C.M. (1999). Detecting differential water functioning over time (DRIFT), Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada, April (Eric document No. ED434115).

Biographical note

Dr Helen Aslett is a lecturer in Developmental Psychology within the Department of Psychology at London Metropolitan University. Helen is currently weighed down with marking, but appreciating the scientific validity of taking short regular breaks from it in order to improve her reliability! Email: h.aslett@londonmet.ac.uk