Studies in Educational Evaluation 47 (2015) 93–101

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Peer assessment using comparative and absolute judgement

Ian Jones [a,*], Chris Wheadon [b]

[a] Mathematics Education Centre, Loughborough University, Loughborough, UK
[b] No More Marking Ltd., Guildford, UK

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 23 July 2015
Accepted 28 September 2015
Available online 5 November 2015

Keywords:
Peer assessment
Comparative judgement
Mathematics education
Validity
Reliability

ABSTRACT

Peer assessment exercises yield varied reliability and validity. To maximise reliability and validity, the literature recommends adopting various design principles including the use of explicit assessment criteria. Counter to this literature, we report a peer assessment exercise in which criteria were deliberately avoided yet acceptable reliability and validity were achieved. Based on this finding, we make two arguments. First, the comparative judgement approach adopted can be applied successfully in different contexts, including higher education and secondary school. Second, the success was due to this approach; an alternative technique based on absolute judgement yielded poor reliability and validity. We conclude that sound outcomes are achievable without assessment criteria, but success depends on how the peer assessment activity is designed.

Introduction

Assessment involves judging a student's achievement within a subject domain on the basis of a piece of evidence such as a test response. Peer assessment is an arrangement in which students are required to make this judgement about other students (Falchikov & Goldfinch, 2000; Topping, 2010). There exist a broad range of motivations for implementing peer assessment, as well as purposes to which peer assessment outcomes are applied. Gielen, Dochy, Onghena, Struyven, and Smeets (2011) listed five common goals of peer assessment: as a social control tool; as an assessment tool; as a learning tool; as a 'learn-how-to-assess-tool'; as an active participation tool. In this paper our focus is on peer assessment as an assessment tool. Gielen et al. state that this goal usually involves a focus on validity and reliability. Moreover, Kane (2013) argues that investigating validity should take account of the purposes of an assessment. In the peer assessment literature, investigating the validity of an assessment tool typically involves comparing peer outcomes to those of teachers or other experts, and to existing achievement data. Reliability is typically measured by comparing the outcomes of two or more groups of peers undertaking the assessment activity independently. These approaches to evaluation were adopted here. Longer-term goals for this programme of research are peer assessment as a learning tool and as an active participation tool. However, these were not explicit goals for the study reported here, although we consider their implications in the discussion.

There are published design principles recommending how best to ensure particular goals are realised and evaluated (Dochy, Segers, & Sluijsmans, 1999; Falchikov & Goldfinch, 2000; Topping, 2003; van Zundert, Sluijsmans, & van Merriënboer, 2010). These principles include clarifying goals, training students on how to assess, and familiarising students with explicit and detailed assessment criteria. The focus here is on the latter: the role of assessment criteria for securing valid and reliable outcomes of a peer assessment activity. Our results suggest that there are contexts in which this recommendation does not apply, but only if the assessment procedure adopts a carefully-designed *comparative* approach to peer assessment.

The role of criteria

The literature provides numerous examples of criteria that might be used in peer assessment, including those generated by students. Sadler and Good (2006), for instance, reported seventh-grade student-generated criteria for peer marking of a biology test. They provided example criteria for one of the test items as follows, with each bullet worth two marks.

Compare and contrast the classification systems of Aristotle and Linnaeus

Similarity:

- Used only observable characteristics of organisms.

* Corresponding author. Tel.: +44 1509217228.
E-mail address: I.Jones@lboro.ac.uk (I. Jones).

http://dx.doi.org/10.1016/j.stueduc.2015.09.004
0191-491X/

Differences:

- Aristotle used where animals live (air, land, water) or plant size and structure;
- Linnaeus used body structure, color, ways of getting food;
- Linnaeus named using binomial nomenclature: genus-species in Latin;
- Linnaeus used many taxonomic levels: Kingdom, phylum or division, class, order, family, genus, species. (p. 12)

The students in the Sadler and Good study were experienced at generating and applying marking criteria, and the criteria were displayed on classroom walls during the peer marking exercise. The student scores were used to allocate a grade (A to E) to each test, and the tests were independently graded by a teacher. Sadler and Good found a high correlation between grades awarded by students and the teacher, $r = .905$. In summary, the authors provided evidence that high agreement between students and teachers is possible, and argued that detailed criteria generated by students who were experienced at assessing peers contributed to the success of the exercise.

More generally, the wider literature makes clear that explicit and well-understood assessment criteria are important for ensuring that peer assessment outcomes are reliable and valid (Chang, Tseng, Chou, & Chen, 2011; Dochy et al., 1999; Falchikov & Goldfinch, 2000; Orsmond, Merry, & Reiling, 1996; Topping, 2003). This is often stated in no uncertain terms. Dochy, Segers and Sluijsmans (1999, p. 342), for example, wrote that the "development of criteria through active cooperation between teacher and students seems to be a critical success factor for self- and peer-assessment". Orsmond et al. (1996) entitled a paper "The importance of marking criteria in the use of peer assessment". Falchikov and Goldfinch (2000, p. 292) considered study designs to be faulty where "students [were] not provided with criteria or structure", and were expected instead to provide a "global rating". Similarly, Topping (2009) emphasised the need to "involve participants in developing and clarifying assessment criteria" (p. 25).

We argue here that the importance of explicit criteria for producing sound peer assessment outcomes is overstated. There are two grounds to this argument. First, the data reported in a widely cited meta-analysis by Falchikov and Goldfinch (2000) are, on closer inspection, equivocal on the role of criteria for achieving sound peer assessment outcomes. The authors identified three approaches in the literature: aggregated scores based on individually-marked criteria as exemplified above; global judgements informed by detailed criteria; and global judgements without criteria. Falchikov and Goldfinch compared the means of the reported correlations between peer and tutor assessment outcomes for each approach. They found a high mean correlation for studies that used global judgements with criteria ($N = 18$, $r = .77$)[1] or global judgements without criteria ($N = 17$, $r = .72$), and a lower mean correlation for studies that used aggregated scores of individually marked criteria ($N = 18$, $r = .53$). They also compared the mean effect sizes (Cohen's d) of the three approaches, based on the means of marks produced by peers and tutors, where the smaller the effect size the better the agreement between the assessment outcomes of peers and tutors. Peers assessed more harshly than tutors (negative effect size) when using global judgements without criteria ($N = 2$, $d = -.32$), and more generously when using global judgements with criteria ($N = 10$, $d = .17$); peers and tutors were in close agreement when using aggregated judgements across discrete criteria ($N = 12$, $d = .03$). The effect size analysis does support the use of discrete criteria, but Falchikov and Goldfinch acknowledged the small number of studies involved, notably only two studies for global judgements without criteria. Moreover, they excluded a problematic study (Butcher, Stefani, & Tariq, 1995) from the effect size analysis and noted that aggregated judgements resulted in the largest mean effect size when it was included ($N = 13$, $d = .34$). In sum then, a case can be made based on the correlational analysis that aggregated judgements across discrete criteria are inferior to global judgements with or without criteria. Conversely, a case can be made based on effect size analysis that discrete criteria are markedly superior and global judgements without criteria are markedly inferior. As such, the evidence provided by Falchikov and Goldfinch is equivocal regarding the role and nature of criteria in peer assessment studies.

Our second reason for questioning the importance of explicit criteria is a study by Jones and Alcock (2014) that investigated a novel approach to using global judgements without criteria. 193 mathematics undergraduates peer-assessed a conceptual calculus test using a technology-enabled comparative judgement technique, described later. The peer assessment outcomes were compared with those of 20 expert mathematicians who assessed the same test responses using the same technique. The correlation ($r = .77$) was higher than the overall mean reported in the meta-analysis of Falchikov and Goldfinch ($N = 56$, $r_m = .69$)[2], supporting the validity of the outcomes. The inter-rater reliability of the peer assessment outcomes were estimated and also found to be acceptable ($r = .72$). Jones and Alcock (2014) argued that if assessment arrangements are devised appropriately and carefully, good outcomes can be achieved without criteria. More generally, there may be contexts in which the aims of a peer assessment exercise are best served using global judgements without criteria.

Research aims

In this article we set out to replicate and extend the findings of Jones and Alcock (2014). We report a study in which secondary school students undertook a computer-based peer assessment exercise in comparative and absolute judgement conditions, and the inter-rater reliability and validity of the outcomes were estimated. There were empirical and theoretical motivations to the research.

The empirical motivation was to explore whether the assessment measures reported for the case of undergraduates' understanding of calculus (Jones & Alcock, 2014) were replicable for lower secondary students' understanding of fractions. This motivation is consistent with Topping's (2010) call for further research into how the arrangement of peer assessment interventions interacts with outcomes; little is known about how the "age and nature of institution of participants" (p.342) might impact on peer assessment activities. The successful replication of Jones and Alcock's main findings for secondary school students would provide support as to the generality of the approach. Given the successful use of comparative judgement in a variety of educational contexts (Bramley, 2007; Heldsinger & Humphry, 2013; Kimbell, 2012; Seery, Canty, & Phelan, 2012) we predicted that Jones and Alcock's results would be replicated.

The theoretical motivation was to investigate the extent to which the sound assessment measures reported by Jones and Alcock can be attributed to the particular comparative judgement technique, described below, rather than to some alternative implementation of global judgements without criteria. To this end we used an experimental design in which students were allocated to a *comparative* or *absolute* judgement condition, and the

[1] The reported mean correlation for global judgements with criteria excluded a problematic study by Burnett and Cavaye (1980). When this study was included the mean correlation was higher, $r = .85$.

[2] The nature of judgement (global/dimension and criteria/no criteria) was not specified in three studies, hence this overall correlation coefficient is based on 56 rather than 53 studies.