A&R

AUTONOMY AND RESPONSIBILITY

JOURNAL OF EDUCATIONAL SCIENCES

EST 2015

JOURNAL HOMEPAGE I https://journals.lib.pte.hu/index.php/arjes/index

ELABORATED PEER ASSESSMENT OF ACADEMIC WRITING BETWEEN POSTGRADUATE STUDENTS

Keith J. Topping

Professor of Education and Society, University of Dundee, Dundee DD1 4HN, Scotland

ORCID: 0000-0002-0589-6796

To cite this article:

Keith, T. J. (2023). Elaborated Peer Assessment of Academic Writing Between Postgraduate Students. *Autonomy and Responsibility Journal of Educational Sciences 8*(2). 5-26. DOI: 10.15170/AR.2023.8.2.1.

To link to this article: https://doi.org/10.15170/AR.2023.8.2.1.

Autonomy and Responsibility Journal of Educational Sciences. Vol. VIII. No. 2. 5–26.
https://doi.org/10.15170/AR.2023.8.2.1.

Elaborated Peer Assessment of Academic Writing Between Postgraduate Students

Keith J. Topping
Professor of Education and Society, University of Dundee, Dundee DD1 4HN, Scotland
ORCID: 0000-0002-0589-6796

Peer assessment in higher education has grown enormously in the last decade but is more commonly used with undergraduates. In this study, reciprocal paired peer assessment of academic writing was undertaken by twelve postgraduate students of educational psychology, who gave elaborated formative feedback on each other's work, as did staff. Overall, staff and peer assessments showed a very similar balance between positive and negative statements, but this varied according to assessment criterion. However, only half of the content of detailed formative assessment statements made showed correspondence between staff and peers. Nevertheless, there was very little evidence of conflict between the views of staff and peers - rather, they focused on different details. Subjective feedback from students indicated that most found the process time consuming, intellectually challenging and socially uncomfortable, but effective in improving the quality of their own subsequent written work and developing other transferable skills. The reliability and validity of this type of peer assessment thus appeared adequate, and the partiality of overlap in detail between staff and peer assessments suggested that the triangulation peer assessment offers is likely to add value. However, caution is indicated regarding the generalisation of this finding. Implications for action are outlined.

Keywords: *peer assessment, writing, elaborated, formative, postgraduate*

Introduction

Much assessment in higher education has been purely summative. By contrast, formative assessment aims to improve learning while it is happening in order to maximise success, rather than aiming to determine success or failure only after the event. Thus, formative assessment seems likely to be most helpful if it yields rich and detailed qualitative feedback information about strengths and weaknesses, not merely a quantitative mark or grade.

Peer Assessment

Peer assessment can be defined as an arrangement for peers to consider the level, value, worth, quality or successfulness of the products or outcomes of learning of others of similar status. Early studies asked students to grade, score or mark the work of other students, but this was found to be variously reliable. More recently interest has grown in having students provide elaborated qualitative feedback, sometimes in addition to grades. A review of 145 studies of peer assessment between students in college and university indicated that such activities were very various in type (Topping, 1998). A typology derived from this literature offers a conceptual framework for the reader (elaborated in Topping, 2018) (see Table 1). Different types of peer assessment might generate positive effects through different mechanisms.

	VARIABLE	RANGE OF VARIATION
1	Curriculum Area/ Subject	All
2	Objectives	Of staff and/or students Time saving or cognitive/affective gains
3	Focus	Quantitative/summative or Qualitative/formative or both
4	Product/Output	Tests/marks/grades or writing or oral presentations or other skilled behaviours
5	Relation to Staff Assessment	Substitutional or supplementary
6	Official Weight	Contributing to assessee final official grade or not
7	Directionality	One-way/reciprocal/mutual
8	Privacy	Anonymous/confidential/public
9	Contact	Distance or face to face
10	Year	Same or cross year of study
11	Ability	Same or cross ability
12	Constellation Assessors	Individuals or pairs or groups
13	Constellation Assessed	Individuals or pairs or groups
14	Place	In/out of class
15	Time	Class time/free time/informally
16	Requirement	Compulsory or voluntary for assessors/ees
17	Reward	Course credit or other incentives or reinforcement for participation?

Table 1: *A Typology of Peer Assessment in Higher Education*

Theoretical Underpinnings of Peer Assessment

Cognitively, peer assessment might create its effects by increasing a number of variables, for assessors, assessees, or both. Depending upon the type of peer assessment, how it is organised and in what contexts it operates, these variables could include levels of time on task, engagement, and practice, coupled with a greater sense of accountability and responsibility. Formative peer assessment is likely to involve questioning - intelligently and adaptively, together with increased self-disclosure and thereby assessment of understanding. It could enable earlier diagnosis of misconception and earlier error identification and analysis. Both of these could lead to the identification of gaps and engineering their closure, through explaining, simplification, clarification, summarising, reorganisation, and cognitive restructuring (Topping & Ehly, 2001).

Increased levels of feedback (corrective, confirmatory, or suggestive) could be coupled with greater immediacy, timeliness, and individualisation of feedback. This might increase post hoc reflection and improve generalisation to new situations, promoting self-assessment and greater meta-cognitive self-awareness. Indeed, cognitive and meta-cognitive benefits might accrue before, during or after the peer assessment actually takes place. Also, there might be meta-cognitive benefits for staff as well as students. Peer assessment might initiate scrutiny and clarification of the objectives and purposes, criteria and marking scales of assessment, and indeed the objectives of the course itself.

Peer assessment might also have an impact on affect, increasing motivation through an enhanced sense of ownership and personal responsibility, greater variety and interest, activity and inter-activity, and also improving self-confidence, identification and bonding, and empathy with others - for assessors, assessees, or both. It has also been proposed that peer assessment might increase a range of social and communication skills, including negotiation skills and diplomacy, verbal communication skills, giving and accepting criticism, justifying one's position and assessing suggestions objectively.

Effects of Peer Assessment

Research in peer assessment is now voluminous. Summarising his review, Topping (1998) concluded that peer assessment of writing appeared capable of yielding outcomes as least as good as teacher assessment, and sometimes better. Formative feedback was variously oral, written, and both combined. Since then, Li et al. (2020) has meta-analyzed 58 studies on peer assessment, finding an effect size of 0.29. The most critical moderating factor was training. When students received rater training, the effect size of peer assessment was substantially larger than when students did not receive such training. Computer-mediated peer assessment was also associated with greater learning gains than paper-based peer assessment. A meta-analysis of 54 experimental and quasi-experimental studies by Double et al. (2020) found an overall small to medium effect of peer assessment on academic performance (effect size = 0.31), but again peer assessment was found more effective than teacher assessment (effect size = 0.28). The effectiveness of peer assessment was remarkably robust across a wide range of contexts. Peer assessment of writing is found in a wide range of subjects, for example: composition, technical and business writing, psychology, education, social science, engineering, geography and computing.

Reliability and Validity of Peer Assessment

Many studies of the reliability and validity of peer assessment utilise comparison of marks, grades or scores, rather than of more open-ended, qualitative, formative feedback. This doubtless reflects the greater ease of comparing quantitative indices. The majority of these studies suggest peer assessment is of adequate reliability and validity in a wide variety of applications (e.g., Topping, 1998), although this seems likely to vary with type and organisational differences. However, a substantial minority of studies question the reliability and validity of peer assessment as they operated it, which of course raises questions about implementation integrity. Acceptability to students is various and does not seem to be a function of actual reliability. There is an evident need for more reliability and validity studies of purely qualitative peer assessment.

Aim, Type and Context of The Present Study

Aim

The present study sought to explore the reliability and validity of pairwise reciprocal elaborated formative peer assessment in the area of academic writing, using given assessment criteria and not coupled with peer marking. The participants in the present study were mature postgraduates with substantial experience of the "real world". However, they were a closely knit group and the peer assessment was one to one. None of them had experienced peer assessment before. It was expected that they would find the experience socially and emotionally as well as cognitively challenging. The acceptability of the procedure before and after involvement in it was to be explored, and subjective views regarding the formative impact of participation as both assessor and assessee gathered, together with information about practical disadvantages and cost-effectiveness.

Type of Peer Assessment

In terms of the typology of peer assessment (see Table 1), this project was an example of a: same year, purely formative and qualitative, out of class, compulsory, supplementary, paired, reciprocal, randomly matched within topic, distance and face to face, confidential peer assessment system in academic writing in postgraduate psychology, targeted on cognitive gains, not contributing to official grade and without extrinsic reinforcement.

Context of the Present Study

The study involved a cohort of 12 students undertaking a two-year Master's level postgraduate course of professional training leading to qualification as a chartered educational psychologist. Entrants already had a good first degree in psychology and at least the equivalent of two years' practical experience with children, parents, schools and/or welfare agencies. In this cohort, 10 were female and two were male, and the average age was 31.

The aim of the course was the acquisition and development of information, strategies, skills, products and services relevant to co-operative work with children, parents, teachers and other carers and professionals, and particular emphasis was placed upon the prevention, assessment, management and resolution of learning and behaviour problems with clients of all ages. The importance of transferable interpersonal and professional skills was explicit, and they were specifically taught in a 40-contact-hour module as well as integrally developed and practised in many other course activities. There was also an emphasis on trainee self-assessment.

All assessment for the course was continuous, and amongst other assessed outputs were written "Academic Reports", one in each of the three ten-week terms per academic year, minimally of 5,000 words. Students chose their own specific topics, in any order, under the general headings of: Normal Child Development, A Case Study of an Individual Child, Organisational Analysis of a Psychological Service, Exceptional Child Development, Intervention Analysis, and In-service Project (with presentation materials).

Reports were to be based on a critical analysis of existing relevant research literature, new data gathered by the trainee where appropriate, and had to relate to professional practice, particularly as experienced during the practical placements which were continuous throughout the course. Students were advised that faults they should seek to avoid were: lack of structure, over-inclusion, irrelevance, repetition, shallow generality, regurgitation, unsupported claims, excess speculation, excess of personal experience, fragmentation, and lack of practical implications.

Course staff normally assessed the reports and graded them Pass or Fail, with double or triple marking for possible Fails and the usual moderation by external examiners. They also gave trainees detailed qualitative formative feedback in relation to the 14 assessment criteria developed by course staff, on a proforma designed for this purpose (see Appendix 1) and available electronically for ease of individual adaptation. This was sometimes supplemented with face-to-face discussion at the request of the member of staff or of the trainee. The course staff assessing the reports were well practised in the use of the assessment criteria. However, it should be noted that given the breadth of student choice of topic, staff often assessed reports on topics about which they themselves had little specialised knowledge.

Methodology

Procedure for Data Gathering

The peer assessment exercise was targeted on the second Academic Report required of the trainees, to be submitted at the end of the second term of the first year. It was thought that at this point the anxiety possibly connected with starting the course and passing the first academic report would have subsided, while much time remained for any formative impact of the procedure to have its effects. Trainees were advised of the upcoming exercise and its practical purposes toward the end of the experimental term, assured that staff marking would be conducted in parallel and be paramount, advised that participation was not optional, and given the opportunity to ask questions (in a class meeting - no subsequent individual enquiries were forthcoming). Trainees submitted their Academic Reports in the usual way at the end of the term, which were allocated for staff marking in the usual rotation. Staff completed the usual feedback sheets (see Appendix 1) but did not give these to the trainees at this point. At the start of the third term, the trainees were advised that all their reports had “passed”.

Trainees were then allocated to pairings for the reciprocal peer assessment exercise. Seven trainees had chosen to do their “Case Study” that term, while five had chosen to do their “Organisational Analysis”. It was decided to pair trainees undertaking different topic areas so far as possible (on the assumption that this might maximise formative impact, although in a less mature group perhaps risking facilitating plagiarism). Names were thus drawn randomly from the topic area groups of seven and five until only two (who had done the same topic) remained, and these were perforce paired together.

Participants were then asked to assess their partner’s report and complete the same assessment feedback proforma used by the staff (see Appendix 1), within four weeks. Copies of the completed proformas were to be exchanged between partners and also given to the course director. Trainees were told that they might want to discuss with their partner the feedback they wished to give before and/or after handing them the proforma, but it was accepted that geographical and time constraints might prevent this. Trainees again had the opportunity to ask questions and voice concerns, and concern was expressed about their ability to assess the work of their peers with reference to the “Originality of Thought” criterion, and to a lesser extent the “Critical Awareness” criterion. This seemed to stem from their awareness of their overall apprentice status and their assignation to cross-topic pairings, in which the assessor would usually be quite new to the topic. Trainees were reassured that they were not required to make a positive and/or negative comment under every category if they did not feel they could validly do so.

During the period allocated for completion of the peer assessment, the trainees also engaged in a two-hour session in the course Research and Evaluation module on “Critical Analysis of Research Reports”, which included an exercise in criticising one of the course director’s own peer reviewed journal publications. Twenty-eight defects were identified by the group.

When all the completed peer assessment proformas had been gathered in by the course director, each trainee was given the staff assessment feedback proforma on their own report. Trainees were then presented with the draft of a follow-up questionnaire designed to solicit their views on the process and outcomes of the exercise, and were asked to critically analyse it and suggest improvements (but not actually answer any of the questions). It was expected that this piloting of the questionnaire with respect to face validity would also serve to promote further thinking about the peer assessment exercise, while the concomitant passage of time brought the need to prepare for the next academic report nearer, and thereby possibly heightened the salience of the task. The follow-up questionnaire was revised in response to the suggestions of the trainees (see Appendix 2), who were then asked to complete the revised version immediately after handing in their academic report at the end of the third term. All 12 were subsequently returned.

Procedure for Data Analysis

Any analysis of the comparability of qualitative feedback from parallel assessors is bound to involve some subjectivity, and the establishment of inter-rater reliability is important in any such process. Analyses were therefore conducted in parallel by the course director (who had not been involved in assessing the reports, but knew the assessment procedure well) and a research assistant who had no familiarity with the course or its procedures.

Initial scrutiny of the peer feedback forms indicated that some statements had no flag (+, -, O; see Appendix 1) attached, while others had flags attached which appeared to be inappropriate (usually O where - was appropriate; suggesting a reluctance to be seen to be negative). Additionally, a few statements appeared to be located under inappropriate categories. Given differing response styles (terse and segregated versus verbose and integrated), there was also some difficulty in isolating what constituted a single statement or unit of meaning. Also problematic were statements made more than once (not necessarily in exactly the same words or in the same category on each occasion), since double counting would confound the analysis. It was decided to count each statement (in whatever equivalent form) only once. Examples given to support an evaluative comment could also prove a problem, since staff and peer assessors might make the same general point, but support it with different examples from the text. It was decided to disregard examples and analyse only general evaluative comments. As had been expected, peer feedback in the “Originality” (and to some extent “Criticality”) categories was relatively sparse.

Given these initial observations, the two raters first independently reviewed the peer feedback forms, sectioning feedback into statements, adding flags where absent, changing flags where the original seemed inappropriate, re-categorising inappropriately located statements, and discarding examples and repetitions. Descriptive statistics from this process are given in Table 2.

RATER 1	Positive	Negative	Neutral	Total
Total	208 (68.8%)	72 (23.8%)	22 (7.3%)	302 (100%)
Signs added	84	33	14	131
Signs changed	1	9	0	10
Statements re-categorised	5	1	0	6
RATER 2				
Total	195 (69.6%)	69 (24.6%)	16 (5.7%)	280 (100%)
Signs added	79	34	10	123
Signs changed	1	8	0	9
Statements re-categorised	5	1	1	7

Table 2: *Comparison of Rater Restructuring of Responses*

A high degree of correspondence between the judgements of the two raters is indicated in Table 2, but of course simple quantitative correspondence could mask qualitative divergence. The differences between raters were largely attributable to one rater's tendency to identify more separate statements than the other, the majority of the "extra" statements being coded either positive or neutral. Considering each statement which was the subject of disagreement individually and qualitatively, the degree of inter-rater agreement is outlined in Table 3.

Positive		Negative		Neutral	
agree	disagree	agree	disagree	agree	disagree
193	17	61	8	15	11
(91.9%)		(88.4%)		(57.7%)	

Table 3: *Inter-rater Reliability*

This indicates an inter-rater reliability of 88.2% overall. However, neutral codings were of little significance (many were due to one rater coding "no opinion" as a neutral comment, while the other rater merely ignored such statements). Consequently, the inter-rater reliability of + and - combined is more important. This was 91.0% - satisfactorily high. In ensuing negotiation between the raters, it was agreed to retain nine of one rater's additional positives and drop six. Four of this rater's additional negatives were retained and three dropped, and six of this rater's additional neutrals retained and five dropped. A final master version of the coding was agreed for the next stage of the analysis. The assessment forms completed by the staff were similarly rationalised, where necessary.

Comparison of peer and staff feedback then proceeded, firstly by comparing the number of +, -, and O flags for each report from the two sources (Table 4 in the Results section below). The raters then independently rated the similarity of the semantic content of statements within categories for each report from the two sources, on a five-point scale in which 0 = no relationship to any statement made by the parallel assessor, 1 = virtually no similarity, 2 = a little similarity, 3 = quite a lot of similarity, and 4 = almost identical (see Table 5 in Results section below). Statements coded 0 were divided into those made by peer assessor only and those made by staff assessor only. Finally, the follow-up process and outcomes questionnaires completed by the trainees were analysed.

Results

Comparison of Flagging Between Peers and Staff

Peer and staff flagging was compared across assessees, between staff assessors, and across assessment criteria. For each assessee, the difference in overall positivity (number of positive statements minus number of negative statements) of staff and peer flagging was calculated. In every single assessment, by either staff or peer, positive statements outnumbered negative statements.

An overall positivity difference of more than 4 between the staff and the peer assessment was considered substantial enough to be worthy of note (somewhat arbitrarily, although there were indications that this point was a trough in a bimodal distribution). On this basis, peers were more positive than staff in three cases, staff more positive than peers in 1 case, and peer and staff positivity was approximately equal in eight cases. In total, staff made 191 positive statements (71.8%) and 75 negative (28.2%), while peers made slightly more positive statements (206 - 74.1%) and a very similar number of negative statements (72 - 25.9%). The summary statistics in Table 4 show that the variance in peer positive statements was greater than in staff positive statements, although this was not true of negative statements. This is unsurprising, given there were 12 peer assessors but only two staff assessors. Thus, there was evidence of a tendency, albeit not a strong one, for the peer assessments to be more positive than staff assessments.

	Positive Statements		Negative Statements		Positivity (+ve - -ve)	
	staff	peer	staff	peer	staff	peer
Total	191	206	75	72	116	134
Mean	15.92	17.17	6.25	6.00	9.67	11.20
Standard Deviation	1.93	3.31	2.24	2.04	3.04	4.08

Table 4: *Comparison of Flagging in Peer and Staff Feedback*

Considering the reports assessed by the two staff assessors separately, one staff assessor recorded 96 positive statements and the other 95 positive statements - almost identical. One staff assessor recorded 42 negative statements and the other 33 negative - a more substantial difference.

Considering the data by assessment criterion rather than assessee, peers were substantially more positive than staff on six criteria, staff more positive than peers on three criteria, and peer and staff positivity was approximately equal in five cases. Assessment criterion #7 was anomalous in that many peer assessors felt unable to comment competently on "originality of thought", whereas staff commented freely on this. Peer assessors were more positive (i.e. less critical) than staff in the areas of: structure (including headings and paragraph), critical awareness, and spelling/punctuation/syntax. Staff were more positive (i.e. less critical) in the areas of: advance organisers (abstract and contents) and conclusion/ synthesis. Peer and staff positivity was approximately equal in the areas of: conceptualisation of main ideas, literature review, new data, psychological content, precision of language, economy of language, action orientation, and references.

Inter-rater Reliability of Similarity of Semantic Content

Only four of the twelve peer assessors had felt able to comment at all on the criterion of "Originality". Of the four that did, a fair degree of agreement with staff assessors was evident (mean rating 2.6). However, given the incompleteness of the data, this criterion was disregarded in the ensuing analysis.

The question of what constituted a single statement for the purposes of comparison was even more problematic in this stage of the analysis, and considerable variation between the raters was evident in their segmenting of the material. Accordingly, rather than comparing the raw ratings of assessor agreement for each assignment and each criterion directly, mean ratings for these were compared.

Statements coded 0 (no relationship to any statement made by the parallel assessor) were divided into those made by peer assessor only and those made by staff assessor only. For these statements, inter-rater agreement ranged from high to low for different assessees and different criteria. The raters showed very similar total numbers of one to four ratings (some degree of similarity between peer and staff assessment) (166 and 159), and very similar total numbers of peer only zero ratings (97 and 103). However, total numbers of staff only zero ratings were considerably different between raters (57 and 93).

Considering "shared" statements coded 1 to 4, some disagreement between the independent raters was evident, even using mean rating per cell and summing the ratings across assignments or criteria (see Table 5). Of course, this simple counting does not consider any relative weighting of the comments, intended by the assessor or inferred by the assessee.

By Assignment			By Criterion		
Assignment	Rater A	Rater B	Criterion	Rater A	Rater B
A	2.88	2.67	1	3.56	3.60
B	3.00	2.47	2	2.67	2.54
C	3.38	2.38	3	2.88	2.36
D	3.08	2.64	4	2.67	2.60
E	3.15	2.43	5	3.00	2.13
F	3.13	3.07	6	2.27	2.33
G	2.93	2.07	7		
H	2.92	2.92	8	2.62	1.91
I	2.50	2.83	9	2.83	2.67
J	2.43	2.43	10	2.82	2.82
K	2.38	3.09	11	3.00	2.20
L	2.46	2.58	12	2.73	2.67
			13	3.15	3.00
			14	2.64	2.50

Table 5: *Similarity of Shared Semantic Content in Peer and Staff Feedback: Mean Ratings by Independent Raters*

There was evidence of an overall tendency for one rater to give higher ratings than the other. Considering the inter-rater agreement by assignment (A - L), a fairly high level of agreement is indicated for seven of the 12, but a lower level on the other five. Inter-rater agreement by assessment criterion appears higher overall. High agreement is indicated for criterion 1 (advance organisers), 2 (structure), 4 (literature review), 6 (critical awareness), 9 (precision of language), 10 (economy of language), 12 (conclusion, synthesis), 13 (spelling, punctuation, syntax), and 14 (references). A fair degree of agreement is indicated for criterion 3 (conceptualisation of main issues). Low agreement is indicated for criterion 5 (new data), 8 (psychology content) and 11 (action orientation). The ratings for these latter were not characterised by high within-rater variance. High inter-rater agreement appears more likely in relation to criteria which focus on structural features of the text, and less likely on criteria which focus on the quality of thought within the assignment.

Staff/Peer Similarity of Semantic Content

On average, 52% of statements were zero rated, and 48% rated as having some shared semantic content (1 to 4). However, very few major clashes of opinion between peer and staff assessors were evident - only three out of 156 possible (12 assignments x 13 criteria in the analysis). Thus, the modest proportion of shared content reflected staff and peers focusing on different specific aspects or exemplars of the assignment, rather than disagreement about aspects on which both had focused. The data on zero rated items were not readily amenable to further analysis and interpretation.

Caution is needed in concluding that the degree of correspondence between staff and peer assessment varied according to the assignment assessed (and peer assessor associated with it) and the assessment criterion addressed, since the variation in the data in Table 5 might be partially attributable to variation between raters. However, it is worth noting that the staff assessors did not differ from each other in overall degree of agreement with the peer assessment - both staff assessors showed a range from high to low agreement across their six assessed assignments (staff assessor A: mean = 2.65, s.d. = 0.24; staff assessor B: mean = 2.61, s.d. = 0.33).

Aggregating ratings from both raters on assessment statements with semantic content common to both staff and peer assessors, the overall mean rating of similarity lies between "a little similarity" and "quite a lot of similarity", tending to the latter. Perhaps it is unsurprising that this mean should lie more or less in the middle of the four-point scale of similarity used.

There was some evidence that on average, the peer assessors gave more feedback statements than did the staff assessors. Staff comments showed a relative tendency to be global, while peer comments could be more particular and detailed, mentioning more specific examples. Whether this could still be expected if the peer assessor had more than one assignment to assess, or if peer assessment was a more regular and routine commitment, is another question. Presumably staff comments are likely to set the assessed assignment in the context of the overall development of the student during the course and the standard all students are expected to eventually reach, while this would be less likely for peer assessors. Interestingly, peer assessors tended to be more critical of completeness and layout of references than staff assessors.

Follow-up Process and Outcomes Questionnaire

Given the small numbers, responses to the Peer Assessment Follow-up Questionnaire will be reported discursively rather than in tabular form. Proportionality should be self-evident. Frequencies are given in brackets. Some participants did not respond to every question, and this should be evident from the text and frequencies.

Process Behaviours

Assessors reported reading their partner's report between three and four times on average (mean 3.46, range 2-6). This was felt necessary to achieve adequate familiarity, several assessors reading once for overall impressions, a second time for more detailed scrutiny and a third or fourth time for conscious and consistent application of the assessment criteria. Two assessors also reported a final reading to check their draft written assessment.

Half of the assessors read their own report again as well, before the peer assessment to practise using the assessment criteria and to give a calibrated baseline (1), or after to check it against the peer assessment (2), to apply the criteria used on the peer report to their own work (1), or to compare their own work with that of the peer (1). Those who did not do this stated that their own report was of a different type and thus of doubtful relevance (2), that they did not think this necessary (2), that they could remember their own report (1), that they did not have the time (1), and that this was not possible as their peer assessor had the only copy (1). However, there is evidence here of peer assessment spontaneously stimulating self assessment.

All assessors reported reading their peer's report while looking at the assessment criteria, and the half who read their own report again as well all also did this while looking at the assessment criteria. All assessors reported discussing their peer's report face to face with them, mostly both before and after completing the written assessment form (7), or only before (4), but rarely only after (1). All assessors reported drafting their written assessment comments before finalising them, either before discussion with their partner (5), after (2) or both (4). Most of the trainees felt the time spent in the peer assessment exercise was "about right" (9), while three felt it was too much (although how "about right" was construed in this context is not certain).

Process Feelings

Five of the trainees reported finding the exercise unequivocally intellectually challenging, while four said they found it a little challenging and three not at all. However, all trainees reported a degree of socio-emotional discomfort, either unequivocally (5), or "a little" (7). The majority (9) reported feeling better after completion ("same" = 3), but the implication of "feeling better" is uncertain, and this might merely have reflected relief rather than adaptation.

Other reported feelings were that the content assessed was useful and interesting (2), that the exercise focused the assessor on their own next report (1), that it focused the assessor on searching for positives (1), that it was very constructive and actually brought people closer together (1), that it was useful to look closely at another's work (1), and that the discussion was enjoyed (1). Less positively, individuals said that a lot more time was needed to do it effectively (2), that the assessor felt pressured to accord the work value it deserved (1), that the assessor was busy and wanted to get it over with (1), and that the group had a positive ethos which made criticism difficult (1).

Comments about ways of reducing discomfort included several variants (4) of a request for graduated experience and/or training prior to such an exercise, perhaps involving anonymous reports initially, although it was acknowledged that might prevent face to face discussion which was of great value (1), and would take more time (1). Another assessor proposed focusing only on positive aspects. Although five trainees felt they would experience less discomfort carrying out peer assessment for a second time, another five felt it would be just as bad, perhaps improved by the prospect of having the same partner (1), but worsened if their own or their partner's report proved particularly poor (1).

Process Evaluation

Eight trainees did not think the pair matching could be done better, while two were uncertain. Two felt choosing your own partner might be better, (although the logistical difficulties of this were acknowledged). One trainee felt same topic area pairing would be better, while another felt cross topic area pairing would be better. Eight trainees reported using the +/-/0 flagging convention, while four did not, (three of these feeling it added nothing and one omitting to do so owing to failure to read the instructions properly). One trainee felt the flagging helped by forcing the assessor to be critical. In fact, the flagging had been introduced largely for research purposes, but one trainee noted that it was dangerous to assume the flags were of equal weight.

Difficulties with the layout of the assessment form were reported by three trainees, no difficulties by five. Some felt there was not enough space for general overall comments, although the form had been provided electronically and was spatially adaptable. Several difficulties with particular assessment criteria were reported, especially originality of thought (8), and to a lesser extent critical awareness (3), literature review (3), discrimination between precision and economy of language (3), psychology content (2), action orientation (1), and conclusion/synthesis (1). Suggestions for additional criteria were not requested, but in retrospect this might well have proved interesting.

The main factors considered potentially to have impaired the reliability and validity of the peer assessment were inexperience of the process (8) and lack of topic knowledge (7). Regarding the latter, one trainee accordingly proposed same topic area pair matching, but acknowledged that formative impact on that topic for the assessor would then be impossible. Three trainees mentioned the possibility of bias stemming from knowing the assessee personally, and three their lack of precision and clarity on terminology and criteria. In cross topic area pairs, knowing you were shortly to produce your own work on same topic could have a biasing effect (1), as could lack of time (1).

Outcomes

Ten trainees felt the exercise was an effective way of helping them reflect upon and improve their own upcoming academic report, while two did not. Ten trainees felt that acting as an assessor was an effective way of learning content which was new and important to them, while one did not. Nine felt that acting as an assessor had helped develop transferable skills which would generalise to their own future writing, while one did not and two were uncertain (one of the latter wisely commenting that this was an empirical question). Two trainees reported help in developing a more critical stance, and four different ideas about structure and organisation. Greater awareness of the reader's perspective and other writing styles were also mentioned (1 each). Nine trainees similarly felt they had gained from acting as an assessee, while one did not and two did not reply. In some cases, the opportunity for more focused discussion & reflection (3) was said to have led to an increased understanding of strengths and weaknesses of their own report (1).

Few trainees (2) could think of other, perhaps less time-consuming or more comfortable, methods which would have had the same effect. Again, graduated training and/or experience was proposed, perhaps involving several steps (perhaps from peer assessment as a group exercise on a neutral report, to individual assessment of a neutral report, to reciprocal peer assessment in private by discussion only, to the present form). Group and individual discussion were considered valuable (2), as was practice on neutral reports of other origin (1), but more comfortable ways would be more time consuming (1).

Opinions were divided on the useful of conducting a similar peer assessment exercise again during the course, five saying no, six saying yes, and one saying yes but less formally. Early in the second year was the favoured time for a second similar peer assessment exercise. Peer assessment of writing could focus on academic reports (1), research dissertations (1), or psychological reports written in practical placements (1). Two trainees felt peer assessment would be much more useful when academic reports were in draft, although it was acknowledged that time constraints and meeting deadlines would then be a problem (1). The abandonment of the flagging convention (1) and keeping feedback private from staff tutors (1) were also suggested.

Six trainees expressed interest in trying peer assessment in other aspects of the course (e.g. presentation skills), while four did not. Video recording presentations to facilitate feedback was suggested (2), as was small group discussion (2), a stepwise introductory training experience (1), feedback in private (1), and the application of assessment procedures to visiting speakers (1). The questionnaire responses of the pair who wrote on the same topic were very little different from the responses of the other pairs who did not.

Discussion and Conclusions

This study explored the reliability and validity of pairwise and reciprocal qualitative elaborated formative peer assessment in the area of academic writing, using given assessment criteria. The subjective views of the students regarding the acceptability of the procedure and formative impact of participation as both assessor and assessee were also gathered.

Unsurprisingly, reliability and validity were found to depend somewhat on the level of analysis. Previous studies had found high agreement between peers and staff when simply awarding overall quantitative marks to written work. In this study, high inter-rater reliability was found in judging whether written qualitative feedback from peers or staff was positive, negative, or neutral.

Overall, staff and peer assessments showed a very similar balance between positive and negative statements. Although peer assessment feedback tended to be slightly more positive than that from staff, this varied on different assessment criteria. Peers were less likely to be critical of the critical awareness shown by the writer, textual structure, and spelling, punctuation and syntax, and tended to avoid commenting on originality. The two staff assessors showed a similar level of agreement with peer assessments, and made equal numbers of positive comments, but one made more negative comments than the other.

However, at the level of analysis of detailed semantic content, inter-rater reliability was relatively high for assessment criteria concerned with structural features of the text, but lower for others (such as "quality of new data", "psychological content", and "action orientation. Inter-rater reliability was adequate for comments made by both peers and staff and by peers alone, but not for those made by staff alone.

Only half of all formative assessment statements made showed some degree of correspondence between staff and peers. However, there was very little evidence of conflict between the statements made by staff only or peers only - rather, they focused on different details.

Subjective feedback from the students indicated that a substantial majority found the peer assessment process time consuming, intellectually challenging and socially uncomfortable, but effective in improving the quality of their own subsequent written work and developing other transferable skills. Gains accrued from acting as assessor and from acting as assessee, but given that the peer assessment was reciprocal and all participants operated in both roles, making this distinction was probably difficult. Peer assessment had spontaneously prompted self assessment in half of the trainees. This feedback suggested that the key mechanisms were increased time on task, engagement and practice, together with the inherent pressure to scrutinise, clarify and functionally apply the assessment criteria, coupled with the deployment of interpersonal communication and negotiation skills.

Although affected by the level of analysis, the reliability and validity of qualitative formative elaborated peer assessment in academic writing appeared adequate in this study. The partiality of overlap between the semantic detail of staff and peer assessments suggests that the triangulation peer assessment offers (together with staff and self-assessment) was likely to add value. However, extreme caution was indicated regarding the generalisation of this finding to other types of peer assessment and other types of student group and course.

Action Implications

The trainees themselves pointed out that replicability and generalisation of these findings were problematic, since they were a small and highly cohesive group confident that all had passed the Academic Report under assessment and would pass the whole course, virtually free of competition and sophisticated in positive interaction. They also noted the crudity of the quantitative aspects of the procedure for comparing assessments, in particular the subtractive measure of overall positivity. Many methodological flaws were evident, but identifying viable alternatives was difficult.

The difficulty of conducting the qualitative analysis of similarity of semantic content raises questions about what students are likely to read into written feedback, even when of relatively high quality, well structured, and substantial in quantity. The assessed student might be less likely to extract the sense intended by the writer than researchers striving for objectivity. In the course which was the basis for this study, students have the opportunity to discuss written feedback on academic assignments, but tend not to take it up very often.

However, the trainees felt that traditional quantitative marking would be greatly inferior, and in this context, some questioned the reliability, validity and usefulness of the quasi-quantitative flagging convention for onward practical purposes. Generally, the trainees felt that the peer assessment exercise was worthwhile, and led to a heightened awareness of the assessment criteria. They also remarked positively on the finding that the written peer assessment feedback tended to be more detailed than that from staff. Given the uncertain reliability and validity of a qualitative assessment process, triangulation was important, and peer assessment coupled with rotation of staff assessors could provide this.

The trainees felt it was difficult to explore the acceptability of the exercise when it was presented as compulsory, which might have shaped the nature of trainee input. Preparation for “live” peer assessment by practising on anonymous academic reports from previous cohorts of students could be useful desensitisation and training. This could yield early clarification of assessment criteria which were particularly unclear or problematic.

The staff contended that the peer assessment exercise also gave the trainees live practice of transferable interpersonal and professional skills in relation to the collaborative process, which is rarely without its difficulties. This could and should transfer into subsequent professional employment, and field supervisors of practical placements could also engage in this process. Briefing regarding academic report assessment criteria was built into the induction process at the start of the course for these trainees, but clearly continuing interactive discussion in relation to subsequent experience was also necessary. Staff also felt that peer assessment of written work could lead into peer assessment of other outputs, such as portfolios and presentations (both of which are also major components of the course under study).

The extent to which this compulsory exercise led to informal peer assessment of subsequent reports in draft form was not explored, but this would clearly be desirable - and less threatening than peer assessment of final drafts by peer assessors allocated at random by staff. The problem of finding time to undertake such developmental work in a crowded curriculum and busy timetable is of course a perennial one - the usual conflict between breadth and depth.

Nevertheless, a hierarchy of activities for peer assessment (PA) of academic writing might include:

- Induction briefing from staff reassessment criteria
- First written qualitative staff assessment feedback
- Compulsory one to one discussion with staff of this assessment
- Option to discuss all subsequent written feedback with staff
- Small group discussion of assessment criteria
- Group oral PA on anonymous written work of previous students
- Individual written PA on work of previous students
- Compulsory paired PA of current drafts by peers selected by staff
- Same-topic peer matching before cross-topic matching
- Focus on positives only, or positives and negatives
- PA feedback oral, written, or both, by student preference
- Compulsory paired PA of final versions by peers selected by staff
- Focus on positives and negatives compulsory
- PA feedback both oral and written compulsory
- Rotate staff and peer assessors
- Monitoring of reliability/validity of staff/peer assessments
- Feedback re monitoring to students
- Further discussion of monitoring feedback
- Consider substitutional PA only after supplementary PA proven
- Informal self-selected PA of drafts of subsequent reports
- Consider PA of other outputs, e.g. portfolios, presentations
- Discussion of generalisation of PA to professional employment.

Peer assessment in higher education is becoming a mainstream idea, but needs further development and evaluation, together with dissemination of results and methodologies widely to practitioners. For this latter, it is important that durable, cost-effective methods are identified requiring low innovation thresholds, which have the potential to be implemented on a large scale. However, the trainees in this study strongly suggested that small pilot projects be undertaken first, careful consideration be given to potential social and time allocation difficulties, and that subsequently the effectiveness of organisational arrangements is carefully monitored.

References

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review,* 32, 481–509. https://doi.org/10.1007/s10648-019-09510-3

Li, H. L., Xiong, Y., Hunter, C. V., Guo, X. Y. & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45(*2), 193-211. doi: 10.1080/02602938.2019.1620679

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249-267. https://doi.org/10.3102/0034654306800

Topping, K. J. & Ehly, S. W. (2001). Peer assisted learning: A framework for consultation. *Journal of Educational and Psychological Consultation, 12*(2), 113-132. https://doi.org/10.1207/S1532768XJEPC1202_03

Topping, K. J. (2018). Using peer assessment to inspire reflection and learning. MacMillan, J. H. (Ed.). *Student assessment for educators series.* Routledge. Retrieved from www.routledge.com/9780815367659 (also in translation in Chinese by Zhejiang University Press).

Appendix 1

ACADEMIC REPORT ASSESSMENT FORM

Author: Year: Term:

Date Received: Date Assessed: Assessed By:

Title:

CRITERIA:

(Make at least one qualitative comment under each criteria heading. Prefix comments with + (indicating comment on aspect adding value to the work), - (aspect detracting value from the work), 0 (neutral comment). Avoid soggy blandness stemming from trying to be nice - vacuous feedback bunched around the median helps no-one.)

1 Advance Organisers (Abstract, Contents)

2 Structure (Headings, Paragraphs)

3 Clear Conceptualisation of Main Issues

4 Literature Review

5 New Data (Type, Range, Quality)

6 Critical Awareness

7 Originality of Thought

8 Psychology Content

9 Precision of Language

10 Economy of Language

11 Action Orientation

12 Conclusion/Synthesis

13 Spelling, Punctuation, Syntax,

14 References

15 Conclusion & Pass/Fail

Signed: ____ (Assessor)

(Copy to Course Director)

Appendix 2

PEER ASSESSMENT OF ACADEMIC REPORT: FOLLOW-UP QUESTIONNAIRE

Your Name __________________________ Please add longer comments on another sheet

PROCESS BEHAVIOURS

1 How many times did you read your peer's report? ____

1a Why this number? ____

2 Did you read your own again as well? Yes / No

2a Why/why not? ____

3 While looking at criteria on the Evaluation Form? 1 / 2 / Both

4 Did you discuss your peer's report face to face with them? Y / N

4a Before or after completing the Evaluation Form? B / A / Both / NA

5 Did you draft your written comments before finalising? Y / N

5a Before or after discussing with your partner? B / A / Both

6 Was the time you spent: Too much / too little / About right?

PROCESS FEELINGS

Did you find the PA exercise:

7 Intellectually challenging? Y / N / A little

8 Socio-emotionally uncomfortable? Y / N / A little

9 After completion, did you feel worse, better, same? W / B / S

9a Any other feelings you had about it?

10 Can you think of any ways to reduce the discomfort?

11 Would you feel less discomfort doing it for a second time? Y / N

PROCESS EVALUATION

12 Could the pair matching be done better? Y / N

12a How? ____

13 Did you use the +/-/0 flagging convention? Y / N

13a If not, why not? ____

14 Did you have any difficulties with the layout of the evaluation form? Yes / No?

If Yes, please list them, and indicate how to improve the form:

a

b

c

15 Were there any assessment criteria with which you had particular difficulty? Please list them and note the nature of the difficulty:

a

b

c

16 What three main factors do you think might have impaired the reliability and validity of your assessment? Please list them:

a

b

c

OUTCOMES

17 As an assessor, was the PA exercise an effective way of learning content which was new and important to you?   Y / N

18 As an assessor, was the PA exercise an effective way of helping you reflect upon and improve your own upcoming academic report?   Y / N

19 As an assessor, do you think you have developed transferable skills from the PA exercise which will generalise to other future writing?   Y / N

19a If yes, what were they? ____

20 Did you gain from being an assessee?   Y / N

20a What? ____

21 Can you think of other, perhaps less time-consuming or more comfortable methods which would have had the same effect? Y / N

21a If Yes, please name them:

22 Would it be useful for you to do this again during the course?   Y / N

22a If yes, when?

____

22b on any particular topic(s)?

____

22c If yes, with any changes? Please specify them or refer above:

23 Would you wish to try PA in other aspects of the course, e.g. presentation skills?

Y / N

23a Please specify aspect/s:

____

24 Any other comments or suggestions, please: