Validating Written Feedback in Clinical Formative Assessment

Michael Page*[1], John Gardner[2], Joe Booth[3]

1. Institute for Health Sciences Education, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK
2. Faculty of Social Sciences, The University of Stirling, UK
3. Royal College of Radiologists, London, UK

* Garrod Building, Turner Street, Whitechapel, E1 2AD, London, UK
m.page@qmul.ac.uk

M Page ORCID iD 0000-0003-1932-9925
J Gardner ORCID iD 0000-0002-3844-7305

This is an Accepted Manuscript of an article published by Taylor & Francis Group in *Assessment & Evaluation in Higher Education* on 15 Nov 2019, available online: http://www.tandfonline.com/10.1080/02602938.2019.1691974.

Validating Written Feedback in Clinical Formative Assessment

Abstract

Formative assessment is widely accepted in Education circles as being crucial to promoting student learning and, since 2010, the UK General Medical Council (2010) has mandated its use in workplace-based clinical training for all new doctors. As a result, the Royal College of Radiologists (RCR) instituted a range of formative workplace-based assessments including the Radiology Direct Observation of Procedural Skills (Rad-DOPS), in which supervisors appraise trainees' performance in carrying out clinical procedures. This paper reports on the quality of the written feedback in 2,500 Rad-DOPS online feedback forms in addressing the aims of the new assessment approach. Random samples of 500 were selected from the first three years of the new assessment implementation: 2010-13, and from 2016-17. Using an appropriate coding frame, the feedback was analysed across the samples against key trainee attributes including stage of training and level of adjudged competence. Criteria for identifying high quality feedback were derived from the literature and a simplified form of Qualitative Comparative Analysis (QCA) was used to identify the conditions associated with high quality feedback. An average of 97% of the assessments contained written feedback but the number of instances of high quality feedback was found to be exceedingly small at around 5%. The paper offers suggestions for making the feedback process more purposeful in achieving the aims of formative assessment.

Keywords: formative assessment; written feedback; medical education; workplace-based assessments.

Introduction

In the late 1980s and 1990s, a series of seminal papers (e.g. Crooks 1988, Sadler 1989 and Black and Wiliam 1998) awakened education in all its disciplinary and age range contexts to the importance of assessment being used to support rather than merely

measure student learning (assessment for learning vs. assessment of learning). In higher education specifically, interest in formative assessment and feedback has attracted considerable research attention (see, for example, Evans 2013 and Carless et al. 2011). Since the 1990s, medical education has also been undergoing radical changes and in the UK, the General Medical Council's *Tomorrow's Doctors* report (GMC 1993), set in motion the transformation of postgraduate medical education from a time-based qualification process to one based on the verifiable development of competence by trainee doctors. The period of change is described by Augustine et al. (2010) as a transition to outcome-based education, with the key to establishing clinical competence being workplace-based assessments.

To support the reforms, the GMC published *Workplace Based Assessment: A Guide for Implementation* (GMC 2010) and mandated that all specialty training should have opportunities for trainees to undertake workplace-based assessments. The aim was to foster a culture of 'nurture and of professional educational support' and 'an environment where assessment for learning (along with assessment of learning) is seen as normal' (2010, 1/2).

In addition to this formal endorsement of assessment for learning, the guide also states that the workplace-based assessment records should be used summatively to compile reports on the trainee's progress as part of the Annual Review of Competence Progression (ARCP). During this review, judgements are made about trainees' suitability to progress to the next stage of training or to have successfully completed training (Committee of Postgraduate Medical Deans, 2016). The workplace-based assessment records may comprise both tick-box assessments against performance-related criteria and written feedback commentaries intended to support next steps in learning.

Workplace-based assessments are therefore cast both as on-going, low-stakes assessments to inform regular discussions, reflection and planning about *progress* (formative); and an accumulated evidence base for informing annual, high stakes decisions relating to *progression* (summative). Successful blending of these purposes may be hypothesised to depend on many factors, for example, the extent to which the trainee's performance is perceived to warrant specific types of feedback. The nature and quality of feedback from clinical supervisors is therefore the focus of this study. Based on a large sample of one type of workplace-based assessment records, namely the Radiology Direct Observation of Procedural Skills, Rad-DOPS (RCR 2016), the focus of this research was to examine the extent to which the feedback demonstrates sufficient quality to be capable of supporting learning and providing valid sources of evidence for its formative and summative purposes?

This formative and summative duality raises a potential conflict in the use of the terms: assessment for learning and formative assessment. Some researchers argue that in most cases, but not all, the terms may be used synonymously (e.g. see Gardner 2012) and there are particular grounds for distinguishing them in this medical training context. From Swaffield's (2011) wider educational perspective, for example, the distinctions are non-trivial. Citing the Assessment Reform Group's definition (ARG 2002), she argues that assessment for learning alludes to '... assessment as a process rather than an event, to planning for gathering information, to interpretation and reflection, to the agency of learners, and to the appropriate adjustment of future learning' (Swaffield 2011, 436). In contrast and arguably more aligned with its use in workplace-based assessments, she suggests (443) that formative assessment is distinguished from assessment for learning in that it is a purpose and function of certain assessments, can involve and be of use to others in different settings, concentrates on curriculum

objectives and can cast learners as passive recipients of teachers' decisions and actions.

This dual purpose of workplace-based assessments, serving progress in learning and progression, fits Bennett's characterisation of a binary role for this type of assessment activity, being primarily formative with a formal but secondary contribution to a summative function (Bennett 2011). In its most recent guidance on *Designing and Maintaining Postgraduate Assessment Programmes* the GMC now avoids much of the duality debate by defining assessment simply as being all judgements, whether they are for '… summative (determining satisfactory progression or completion of training), or formative (developmental) purposes' (GMC 2017a, 4). This primacy of the assessment as judgment, regardless of subsequent usage, satisfies Newton's (2007) distinction of being a first or 'judgement level' purpose (150). Judgement level assessment, he argues, has the technical aim of making a criterion-referenced judgement about learners and does not imply how that judgement is to be used. Newton's second level of purpose, the 'decision level', is characterised by discourse about the 'decision, action or process' (*ibid*., 150) that may be supported by the assessment judgement e.g. determining satisfactory progression or providing feedback for formative (developmental) purposes. The GMC's view is that all assessments have a judgement dimension made by assessors using their professional expertise and experience, which may then be used formatively or summatively, or both.

Feedback in Workplace-Based Assessments

Medical education endorses the widely held view (e.g. Bloom, 1971; Ramaprasad, 1983; Kluger and DeNisi, 1996; Perrenoud, 1998; Shepard, 2000; Hattie and Timperley, 2007 and Shute, 2008) that appropriate feedback from assessments has the potential to help learners move from where they are in their learning to where they want to be. In

formulating their 2010 workplace-based assessment policy, the GMC had the benefit of a flurry of mainly small-scale research studies undertaken in medical schools around the world to pilot or evaluate the role of feedback in workplace-based assessments. For example, Johnson et al's (2008) study found that the majority of a sample of medical trainees in the UK perceived workplace-based assessment events to be a valuable source of feedback. In Wilkinson et al's (2008) pilot of workplace-based assessments across the UK medical specialties, 80% of the 230 volunteers reported favourably on the perceived educational benefit of the assessments whilst in the US, Holmboe et al. (2004) reported that workplace assessments provided useful opportunities for senior doctors to give developmental feedback to junior colleagues.

However, the various findings pre- and post-2010 are not consistently positive. In the UK, Fernando et al. (2008) found that written feedback in 396 mini-clinical evaluations had no positive features in 23% of the cases, no development suggestions in 28% and no action plan in 50%. Cohen et al's (2009) UK study with dermatology trainees reported that the extent of trainees' positive comments on the educational benefits of a range of workplace-based assessments was 'striking' but some 20% expressed dissatisfaction with the quality of feedback and 55% did not identify any learning points from the assessments. Holmboe et al. (2004) also tempered their positive findings (above) by reporting that feedback resulted in an action plan being formulated in only 8% of the assessment encounters, despite 80% of them containing at least one suggestion for improvement. The Massie and Ali review of workplace-based assessment usage (2016) argues that much of the current English-medium research on workplace-based assessments is UK-based. However, the various concerns relating to their effectiveness are undoubtedly shared in medical schools around the world, for example, in the US (e.g. Canavan et al 2010), the Netherlands (Driessen and Scheele 2013),

Ireland (Barrett et al 2016) and Australia (Preston et al 2019).

Against this backdrop, and citing a variety of sources, the UK Academy of Medical Royal Colleges (AoMRC 2016) advise that feedback should be provided during or immediately after workplace-based assessments; should be descriptive, non-judgemental and focused on trainees' behaviours; should be specific and related to the learning goals; and should enable trainees and trainers to formulate action plans and future learning

These stipulations suggest that workplace-based assessment feedback is expected to be mainly verbal and therein lies a major challenge for researchers who wish to determine the quality of the feedback. To do so, they would have to observe situations that are intensely personal for everyone concerned: the patients, the trainees and the assessors. For example, the patient's consent for an external observer would be required. The trainees are also in a vulnerable situation in which their performance in undertaking a procedure with a patient is being observed and judged by a senior clinician, who in turn may also find the situation very intrusive. Given the complexities of consent and privacy in these circumstances, it is not surprising that there is little research in the literature on the quality of verbal feedback in workplace-based assessment contexts.

Written feedback by its nature is more accessible to researchers and a number of studies have focused on its use in clinical assessments. For example, in the Netherlands, Prins et al (2006) asked 46 general practitioner (GP) trainees and 12 of their trainers to view and write a feedback report on a 6-minute video consultation between a GP and a patient. They concluded, inter alia, that giving and receiving feedback did add value but both groups 'delivered feedback reports without structure and with limited stimulation for reflection … the reports contained hardly any reflective questions, suggestions for

performance improvement, or examples.' (300). In the US, Canavan et al (2010) examined 970 multi-source feedback (MSF) forms from four institutions and found that only 29% contained written feedback whilst Vivekananda-Schmidt et al. (2013) in the UK found only 42% of 11,483 MSF forms contained written comments. They concluded that written feedback in MSFs 'is unlikely to provide information … on the true strengths and weaknesses of a colleague that will facilitate that individual's personal development' (1086). Notwithstanding these views, a number of authors make important claims for the utility of written feedback. Orsmond *et al*. (2005) and Carless (2006), for example, have demonstrated that learners often review written feedback with the intention of making improvements to their work, thereby supporting reflection, consolidation and repeated attempts to comprehend and apply the advice of the tutor. Jolly and Boud (2013) also highlight the potential for written feedback to be private, allowing learners to avoid the embarrassment of public criticism or even public praise.

The GMC specifically requires that written feedback should provide trainees with a basis for action to improve their performance. In theory, then, written feedback could be superior to verbal feedback as the assessors may be expected to give more consideration to their comments than the relative immediacy of the latter might allow.

Validity versus Validation of Workplace-based Assessments

Bennett (2011) argues that for formative assessment to be considered effective it must satisfy a 'theory of action' with two types of argument: 'a *Validity Argument* to support the quality of inferences and instructional adjustments, and an *Efficacy Argument* to support the resulting impact on learning and instruction' (14, original emphases). However, a more nuanced view (e.g. Kane 2006) distinguishes between validation and validity; the former being a largely theoretical examination of the nature of the

formative assessment, and the latter being an all-encompassing concept (including Bennett's efficacy) in which empirical investigation establishes that learning is actually improved as a result of the formative assessment. For various reasons, (see, for example, Sargeant et al (2005), (2007); Archer et al, (2010) and Burford et al, (2010)) the receipt of good feedback cannot in itself guarantee improved learning or performance but the ideal situation is that feedback should be reflected upon and should then 'feed forward' into a plan of action to improve performance.

The Validity-Validation distinction is particularly important in the current study, which focuses on validation per se by taking the position that the provision of appropriate formative assessment feedback is an important precursor to enabling learners to improve their learning. This is underlined in the GMC's curriculum standards, which unequivocally require training providers to ensure that their programmes offer '… opportunities for formative assessment and feedback to support learning, linked to learning outcomes' (GMC 2017b, 21).

In combination with the workplace-based assessment guidelines for medical colleges (GMC 2017a), the standards signal several potential features of high quality feedback, namely: links to learning goals, timeliness, the capacity to provide evidence and guidance relating to performance, a stimulus for reflection and a basis for planning follow-up action. The Royal College of Radiologists therefore focuses on feedback as an important dimension of workplace-based assessment, viz: '… engaging in constructive conversations about learning, successes, difficulties and progress are all part of an effective professional learning environment' (RCR, 2016, 173).

Given the pressured, time-bound environment of the clinical settings in which workplace-based assessments often take place, it is optimistic to expect comprehensive

dialogic conversations always to occur between trainees and supervisors. Accordingly, there may be a lack of continuity in the educational relationship, and limited scope for follow-up on feedback. Indeed, when the GMC (2013) asked trainee doctors (n=52,484) specifically about the quality of formal progress meetings with supervisors and formal assessment of performance in the workplace, 32% reported that they rarely or never had informal feedback from a senior clinician on their performance. In 2012, the corresponding figure was 33% (unfortunately the area has not been explored in surveys since 2013). Clearly, more than 65% did have what they described as educationally useful progress meetings, but the scale of the minority who report not receiving adequate formative engagement, if it persists today, must be a concern.

There are other vehicles for providing informal and formal support for trainees. Bloxham and Campbell (2010), for example, argue that professional learning also occurs by immersion in the particular community of practice, with extensive opportunities for 'observation, imitation, participation and dialogue' (292). In the absence of appropriate verbal feedback or community of practice support, however, the formal failsafe is the written feedback that should be recorded on the completed workplace-based assessment forms. Although unlikely to encompass the dialogic richness of a feedback conversation (see Crisp, 2007; Bloxham and Campbell, 2010), written feedback, if appropriately constructed, has the potential (but cannot guarantee) to facilitate the trainee's reflection and any necessary improvement in their performance in the observed procedure.

Drawing on Kane's (2006) distinction above, the *validation* of the quality of the written feedback must therefore be established before adjudging the formative assessment procedure to be *valid* in all its contexts including the improvement of learning. *Validation* that is at the centre of this study. Put simply, the study sought to address the

question: is the written feedback in the RCR's Rad-DOPS assessments of sufficient quality to be able to achieve the purposes designated by the GMC and the RCR, i.e. to promote the trainees' formative reflection, action planning and learning, and provide evidence for judgments on progress and progression?

The Rad-DOPS Assessment

The Rad-DOPS assessment is designed specifically to assess trainee doctors as they undertake clinical procedures with a patient. Trainees are required to undertake a minimum of six Rad-DOPS in each training year, and on each occasion the assessor uses an online feedback form to record the following assessments:

- Performance against a series of 11 specific criteria, rated on a 6-point nominal scale from 'Well below expectation for stage of training' (1) to 'Well above expectation for stage of training' (6). 'Communication with patients/staff' is an example of an assessment criterion.
- 'Overall Competence', judged against four narrative ratings ranging from 'Trainee requires additional support and supervision' to 'Trainee requires little/no senior input and [is] able to practise independently'.

Written feedback from both the assessor and the trainee is mandatory in two free text fields at the end of the form; designed respectively to offer constructive feedback and capture reflections on performance and any actions.

Methods

The research was facilitated by the Royal College of Radiologists, who approved access to the anonymised database of Rad-DOPS forms for the first three years of the workplace-based assessment policy implementation: 2010-11 to 2012-13; and for the

year 2016-17. The analysis of data from the year 2016-17 was designed to test the resilience, four years on, of the characteristics of the first three years' samples. In any given year, there are radiology trainees at each of six stages of specialty training (ST): ST1 to ST3, known as the 'core' training years, and ST4 to ST6: the 'higher' training stages with ST6 being the final year. A minimum of six Rad-DOPS assessments was required to be undertaken by all new trainees from 2010-11 onwards but trainees already in the system pre-2010 could opt voluntarily to undertake the assessments year-on-year. Rad-DOPS assessments are conducted by senior clinicians when appropriate opportunities arise and each assessment is logged on the trainees' e-portfolios. Ethical approval for the research was granted by the first author's institution.

In seeking to validate the quality of written feedback in the Rad-DOPS assessments, the research design involved a series of interrelated steps:

- Constructing a coding frame to analyse feedback comment;
- Establishing an adequate sample size;
- Examining the occurrence of comment types across trainee profiles;
- Establishing a rigorous definition of high quality written feedback;
- Investigating the conditions associated with high quality written feedback.

Constructing a Coding Frame to Analyse Feedback Comments

The study initially drew on Canavan et al's (2010) coding frame, which included references to behaviour (general or specific), the valency of the feedback (positive or negative) and suggestions for development (general or specific). This basic set of codes was then expanded using the literature and insights gained from a data immersion process in which we carried out iterative 'sweeps' of a random sample of 500 Rad-DOPS assessments from the training year 2010-11. This process ultimately yielded the

coding frame presented in Table 1.

Table 1: Coding framework with criteria for assigning the codes to comments

Comment Code	Explanatory Criteria: *The comment …*
Positive Valency	... is clearly intended to be positive in nature
Negative Valency	... is negative in nature and includes any suggestion that improvement is necessary
General Performance	... refers in general terms to how the trainee has performed the procedure
Specific Performance	... is clear in indicating how the trainee has performed the procedure
Link to Assessment Criteria	... clearly invokes one or more of the assessment criteria on the Rad-DOPS form
Describes Procedure	... is limited to a description of the procedure only
General Development	... makes a suggestion for improvement that is unclear or ambiguous
Specific Development	... makes a suggestion for improvement that is clear and unambiguous
Personal	... refers to an aspect of the trainee's personality or personal qualities
Global Assessment	... refers to the trainee's overall progress within the clinical placement
Assumed Future Improvement	... suggests that time or continued practice would bring about improved performance
Absent	- there is no written comment -

Establishing an Adequate Sample Size

In the first year of implementation (2010-11), workplace-based assessments were compulsory for trainees in their first year of training, ST1, and optional for ST2-ST6. Table 2 shows the numbers of discrete trainees in each stage of training, the number of Rad-DOPS assessments they collectively undertook and the number of discrete assessors involved for each year.

Table 2: Number of trainees (N_T), assessors (N_A) and completed Rad-DOPS forms for each stage of training for the years: 2010-11 to 2011-13 and 2016-17

Year	**10-11**		**11-12**		**12-13**		**16-17**		**Total**	
Trainees	**N_T**	**RDOPS**	**N_T**	**RDOPS**	**N_T**	**RDOPS**	**N_T**	**RDOPS**	**N_T**	**RDOPS**
ST1	223	1934	307	2383	218	1663	349	3537	1097	9517
ST2	136	1056	275	2131	236	1624	358	2964	1005	7775
ST3	118	877	172	1346	215	1345	319	2450	824	6018
ST4	113	912	170	1312	151	1180	306	2258	740	5662
ST5	5	17	104	835	137	897	252	1788	498	3537
ST6	0	0	1	6	17	96	55	501	73	603
Totals	595	4798	1029	8013	974	6805	1584	12997	4182	32613
Assessors (N_A)	1691		2395		2260		3670			
Mean Rad-DOPS per Assessor	2.8		3.3		3.0		3.5			

In order to establish an adequate sample size, two approximately 10% samples (S1 and S2) of 500 forms were randomly drawn from the first year data set. The assessor's

written feedback statement was analysed using the coding frame in Table 1 and the coding unit was any 'basic unit of text that consisted of a complete idea' (Brann and Mattson 2004, 156). The feedback statements were analysed for what Graneheim and Lundman (2004) term the manifest content (what the words actually mean) and latent content (contextualised by the researchers) with adaptations, as necessary, to the framework and the criteria attaching to the codes. The coding decisions were corroborated by inter-rater checking between the authors; and an example of the coding process is set out in Table 3:

Table 3: Illustration of the application of the coding framework

Statement: *A very competent examination of the abdomen as well as the soft tissues and muscle. More scanning of patients with complicated clinical pictures would help to adapt scanning technique.*	
Coding Unit	**Attributed Codes**
[(A very competent examination)$_{a,c}$ of the abdomen as well as the soft tissues and muscle]$_b$	a. Positive Valency b. Link to Assessment Criteria c. General Performance
[More scanning of patients with complicated clinical pictures (would help)$_c$ to adapt scanning technique]$_{a,b}$	a. Negative Valency b. General Development c. Assumed Future Improvement

Chi-squared tests of independence were carried out on the frequencies of the comment codes found in each sample and the results showed that aside from Assumed Future Improvement (Chi-squared=7.22, df=1, N=1000, p=0.007), the two samples showed no significant variation. Re-examination of the data confirmed that the Assumed Future Improvement code had been applied consistently and no clear reason for the difference could be identified. On the basis of the good agreement, a sample size of 500 assessment forms was therefore considered to be sufficiently representative of the range of feedback to be found in any year group.

Examining the Occurrence of Comment Types across Trainee Profiles

The Rad-DOPS forms yielded several features of trainee profiles for analysis and these included their stage of training: the 'core' phase (ST1-ST3, the first three years of

training) and the ‘higher’ phase (ST4-ST6), the nominal ratings for the Rad-DOPS assessment items, the overall judgement of their competence, the time of year they undertook the assessments and the number of Rad-DOPS assessments they completed. The forms also enabled the number of words in the written feedback to be captured. Using these data, it was possible to examine any associations with the various feedback comment codes; for example, do trainees in the ‘core’ phase of training receive the same types of feedback as those in the ‘higher’ phase?

Establishing a Rigorous Definition of High Quality Written Feedback

Owing to the overlapping and sometimes contradictory comments that, for a variety of reasons, can occur in a written feedback statement, the instrument in Table 1 may not be exhaustive in defining comment types. However, in order to identify whether the feedback could fulfil the GMC/RCR purposes of supporting learning through reflection and planning, we needed to identify a rigorous set of feedback characteristics. The outcome in which we were interested – high quality written feedback – is clearly a composite concept. The previously-cited ‘canonical’ literature of formative assessment could arguably summarise high quality feedback as requiring the *presence* of comments that are linked to the relevant assessment criteria, are constructively critical, and are specific to the development and/or performance. The literature also strongly endorses the *absence* of comments about the person. However, the evidence is largely ambivalent on whether general comments (e.g. on skill performance or development) or broadly-based assessment feedback (e.g. a global assessment of overall progress) constitute ‘good’ or ‘bad’ types of feedback as there may be circumstances when general observations stimulate reflection that leads to improved performance. On this basis it was decided to include such comments. For the purposes of this study, therefore, a theoretical model of high quality feedback was established as having the following

combination of features:

The *presence* of

- positive or negative comments on the observed performance, and
- specific or general comments on the observed performance, and
- comments linked to the assessment criteria, and
- specific or general suggestions for further development, and

the *absence* of

- personal comments.

Establishing the Conditions Associated with High Quality Written Feedback

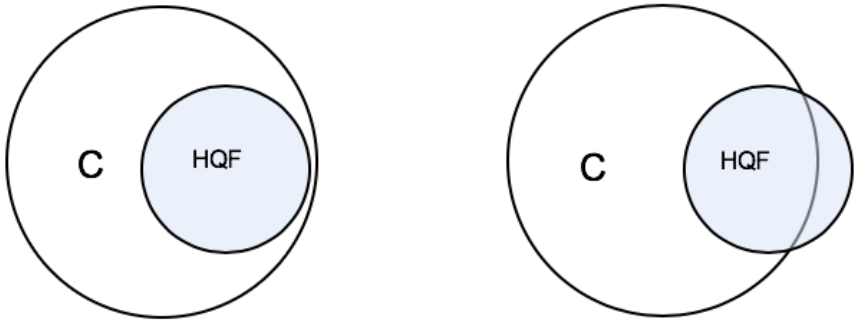
It is possible to hypothesise that assessors are more likely to give feedback, designed to promote improved performance, to those trainees in most need of it, for example, trainees in the early stages of training or trainees struggling to achieve competence. Conversely, it may be reasonable to predict that such feedback is less likely to be given to trainees whom the assessors deem to be already competent, e.g. those nearing the end of their training. In order to establish which of the various conditions are most likely to give rise to the provision of high quality written feedback we chose to use a simplified form of Ragin's Qualitative Comparative Analysis, 'QCA', (Ragin 2008, Legewie 2013).

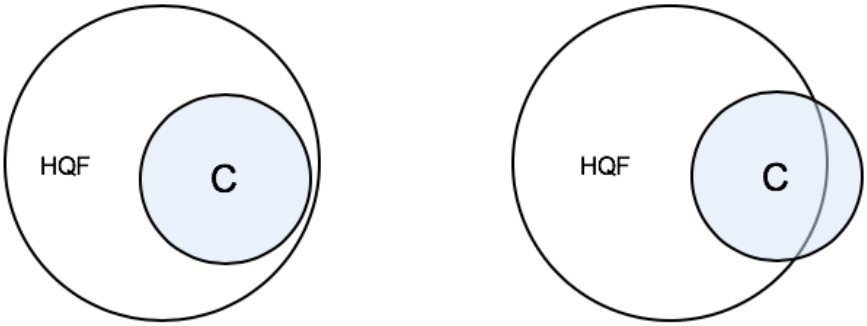
According to Glaesser and Cooper (2012), the QCA approach recognises the causal complexity of interrelated and interdependent factors in social contexts rather than attempting to regard them as independent, which is often a requirement for conventional statistical analysis. They argue that particular conditions may function together or in

isolation to bring about an outcome. These relationships between variables in social settings are recognised to be imperfect. Thus, instead of a condition always being true for a particular outcome to be achieved, a relationship can be designated partly or mostly true. These 'quasi' relationships are usually illustrated by means of Venn diagrams (cf. Glaesser and Cooper, 2012) as in Figure 1, with high quality written feedback as the target outcome and C representing a condition relating to the trainees (e.g. stage of training) or to the feedback statements (e.g. the number of words in the written feedback statement).

In Figure 1(a), therefore, condition C is deemed necessary for the outcome high quality feedback to occur because C is present for every instance of high quality feedback. In (b), however, we can only say that for the outcome high quality feedback to occur, condition C will *almost always* be present– the presence of condition C is therefore quasi-necessary for high quality feedback to occur. Conversely, in (c), for all instances of condition C, the outcome high quality feedback will *always* occur, i.e. the presence of C is sufficient for high quality feedback to occur. In (d), however, we can only say that for *almost all* instances of condition C, high quality feedback will occur – the presence of condition C is quasi-sufficient for high quality feedback to occur. It is prudent to note Legewie's (2013) caution at this point that QCA can reveal associations that signal causal relationships but do not prove them.

Figure 1: Venn diagrams for relationships between high quality feedback (HQF) and the identifie





The diagrams also illustrate the extent of association between the occurrence of high quality feedback and the condition under scrutiny. For example, in (a) 100% of the instances of high quality feedback are subsumed by the set of instances of condition C, implying that C is *always* necessary to give rise to high quality feedback. In (b), however, approximately 70% of the instances of high quality feedback are associated with the presence of condition C and 30% are not. In this case, C is *almost always* necessary to give rise to high quality feedback at the 70% level. Analogous to variance limits in conventional statistics, this 70% threshold is conventionally attributed to the state of quasi-necessity (and likewise for quasi-sufficiency). The written feedback in the samples was therefore examined to establish the extent of any QCA-based associations between the identified conditions (e.g. stage of training) and the provision of high quality feedback.

Results

Data Characteristics

Table 4 shows that the mean number of assessments recorded by trainees across the samples (including those opting in, indicated in parentheses) was in keeping with the curriculum requirement for at least six Rad-DOPS assessments to be completed within

each training year. However, there was a pronounced mode of 1 (i.e. just one completed assessment) in 2016-17 for ST1, ST2 and ST4 trainees (though multiple modes were apparent and 6 was the next most frequent number of assessments). In contrast, ST3 trainees had a mode of 7 and ST5 and ST6 trainees (for whom the workplace-based assessment policy was now compulsory) had 6 and 8 respectively. It is not clear how or why large proportions of relatively new ST1 and ST2 trainees accumulated less than the compulsory six assessments, thereby opening themselves up to potential sanction, but difficulties in organising assessments (see McKavanagh et al. 2012) and personal circumstances may account for some of them.

Table 4: Mean data for trainees taking Rad-DOPS assessments across four samples

Trainees	Mean Number of Assessments Taken by Trainees			
	Sample by Year			
	10-11 (S2)	11-12	12-13	16-17
ST1	8.7	7.8	7.6	10.1
ST2	(7.8)	7.7	6.9	8.3
ST3	(7.4)	(7.8)	6.3	7.7
ST4	(8.1)	(7.7)	(7.8)	7.4
ST5	(3.8)	(8.0)	(6.5)	7.1
ST6	-	-	(5.6)	9.1

Predictably (and in accord with other researchers' findings, e.g. Barrett et al 2016), the peaks of assessment activity were evident around 50% and 90% through the training years, aligning with the points when formal progress discussions would normally be scheduled by supervisors. Worryingly however, from a formative assessment perspective, there was evidence of a large number of assessments, representing 9-12% across the four samples, being recorded at the very end of placements and a substantial number being completed retrospectively (7-13%). Clearly, these very late or retrospective assessments have relatively poor utility as formative assessments as they present little or no opportunity for trainees to improve their performance during their placements.

Written feedback statements were found in approximately 97% of the Rad-DOPS forms but these were of widely different word counts with approximately 35% having 10 or less and 65% having 11 or more words (Table 5). On this basis, the arbitrary categorisation of 'brief' (≤10) and 'extended' (≥11) word counts was adopted for subsequent analyses.

Table 5: Frequency of statement word counts for Rad-DOPS with written feedback (WFB)

	Feedback Statement Word Count % (n)				
No. of Rad-DOPS with WFB	**1-5**	**6-10**	**11-20**	**21-50**	**51+**
2010-11 S1 (486)	16.9 (82)	17.7 (86)	32.3 (157)	27.0 (131)	6.2 (30)
2010-11 S2 (493)	15.4 (76)	21.1 (104)	28.2 (139)	28.0 (138)	7.3 (36)
2011-12 (489)	17.4 (85)	20.9 (102)	33.3 (163)	21.9 (107)	6.5 (32)
2012-13 (473)	15.9 (75)	17.5 (83)	27.9 (132)	29.0 (137)	9.7 (46)
2016-17 (489)	13.9 (68)	17.8 (87)	24.3 (119)	32.9 (161)	10.6 (52)
Average Proportion of Each Word Count	16%	19%	29%	28%	8%

Comparison of Feedback Comment Types across the Samples

The content of the written feedback statements was analysed according to the coding framework in Table 1 using the process illustrated in Table 3. Only the proportions of the Global Assessment and Absent comment types showed a significant difference across the five samples (Chi-squared =18.94, df=4, $p<0.001$ and Chi-squared = 16.86, df=4, $p<0.005$ respectively) but when the outlier values for these comments were removed (52 for Global Assessment in 2011-12 and 27 for Absent in 2012-13), the differences among the remaining four samples were not significant. A review of the data found no particular explanation for the outlier values.

Association of Comment Code Types with Rad-DOPS Judgements, Stage of Training and Feedback Statement Word Count

The assessors' judgments on the Rad-DOPS forms are presented in two main formats: a summary statement of overall competence selected from four options, and a series of ratings from '(1) Well below expectation' to '(6) Well above expectation' against 11 competence criteria. The latter are nominal judgments, enabling the modal rating on each form to be examined, the extremes being trainees with all 1s and 6s respectively (a 'mean score' was inappropriate because some criteria could not be addressed in particular procedures). Using these two types of data, and the stage of training and statement word count, it was possible to investigate associations between them and the comment codes.

Overall competence: The samples revealed that very few trainees (0-7 across the samples) received the lowest level of overall competence judgement: 'Trainee requires additional support and supervision'. However, for all trainees there was no significant difference between overall competence judgements and the frequency of most of the feedback comment types. The two exceptions: General Development and Negative Valency, were significantly more likely to occur for trainees judged to need Direct Supervision than those judged either to need Indirect Supervision or to be sufficiently competent for Independent Practice (see Table 6).

Table 6: Significant Chi-squared results for comment codes and modal ratings

Chi-Squared Values (df=1, p<0.01, n=500)				
Sample	**2010-11 S2**	**2011-12**	**2012-13**	**2016-17**
General Development				
Direct vs Indirect Supervision	12.84	18.15	19.19	26.79
Direct vs Independent	12.45	25.90	20.40	24.32
Negative Valency				
Direct vs Indirect Supervision	22.7	16.54	25.90	17.34
Direct vs Independent	19.27	25.90	25.90	27.62
Modal Rating				
≤4 vs 6	17.83	23.02	12.11	-NS-

It is arguably predictable that developmental or negative commentary might be associated with a judgement that a trainee needs Direct Supervision but the data revealed that the more common outcome for trainees of any level of performance, including those deemed to require Direct Supervision, was that they received no negative feedback and no suggestions for improvement. Furthermore, the general comments were relatively unhelpful, for example: ‘technique is currently a bit unrefined’, 'get more practice', 'see more patients' and 'learn tips and tricks'. It was only in the 2012-13 sample that more useful, specific feedback such as: ‘watch collimation is not too tight’ and ‘refine [communication] further e.g. by indicating beforehand how long the procedure will take’ was more associated with trainees judged to need Direct Supervision than trainees rated as needing Indirect Supervision (Chi-squared=11.83, df=1, p<0.01).

Modal rating: Very few reports (2 to 35 across the samples) recorded a modal rating of 3: ‘borderline for stage of training’ and the number of reports with modal rating <3 was even fewer (0-7 across the samples). The only significant difference for comment types was found for the incidence of negative comments for the combined modal ratings ≤4 versus 6 (Table 5). Again, it is perhaps predictable that this should be the case, but receiving no negative comments was actually a more frequent feature of low modal rating reports. For example, of the 262 reports in 2010-11 with modal rating ≤4, 171 did not have negative feedback comments compared with 91 which did.

Stage of training: All but one of the comment types, Link to Assessment Criteria in 2010-11 S2, gave rise to non-significant differences between the Rad-DOPS reports for core phase trainees (ST1 to ST3) and higher phase trainees (ST4 to ST6); suggesting that the feedback patterns remained more or less static across the sample years and stages of training.

Feedback statement word count: Perhaps unsurprisingly, there were significant differences for all comment types in the extent of their association with brief vs extended statement word counts (≤10 or ≥11+), largely due, we are sure, to the difficulty in making any kind of meaningful feedback comment in 10 words or less.

Conditions Associated with the Provision of High Quality Written Feedback

The adopted definition of high quality written feedback was only identified in very small numbers of the Rad-DOPS forms (average 5%). Less stringent criteria did not make sufficient difference to persuade us to compromise the definition; for example, removing Link to Assessment Criteria only marginally improved the frequencies to an average of 7.5%.

These very small proportions of high quality feedback in the samples imply that none of the conditions (overall competence, modal rating, stage of training and statement word count) could satisfy the quasi-sufficiency requirement for high quality feedback to be associated with them (the threshold being *almost always*, i.e. ≥70%). Perhaps disappointingly, it must be concluded that a large proportion of the assessments with extended passages of feedback failed to deliver the rigorous requirements for high quality feedback (though some aspects of high quality could be present).

In contrast, and as illustrated in Table 7, one condition, extended word count (≥11), was

present for all (100%) instances of high quality feedback and therefore in QCA terms can be deemed necessary. Requiring some level of supervision (Direct or Indirect), a modal rating of 4 or less, and being in core stage of training (ST1-ST3) all surpassed the 70% threshold and can be deemed quasi-necessary for high quality feedback to be given.

Table 7 Extent of association of selected conditions with high quality feedback

Conditions	**Statements with High Quality Feedback**			
	2010-11 S2 N=23	**2011-12** N=19	**2012-13** N=22	**2016-17** N=32
	% (n)	% (n)	% (n)	% (n)
Overall Competence				
Direct or Indirect Supervision	96 (22)	90 (17)	96 (21)	93 (30)
Independent Practice	4 (1)	11 (2)	5 (1)	9 (2)
Modal Rating				
Less than/equal to 4	70 (16)	74 (14)	91 (20)	81 (26)
Greater than/equal to 5	30 (7)	26 (5)	9 (2)	19 (6)
Stage of Training				
Core ST1-3	87 (20)	90 (17)	77 (17)	75 (24)
Higher ST4-6	13 (3)	11 (2)	23 (5)	25 (8)
Feedback Word Count				
Brief ≤10	0 (0)	0 (0)	0 (0)	0 (0)
Extended ≥11	100 (23)	100 (19)	100 (22)	100 (32)

The analysis also suggested that several conditions could be deemed logical ‘NOT’ conditions i.e. their *absence* is required in order for high quality feedback to be provided. These include instances of Independent Practice judgments for overall competence, a modal rating of 6 or final stage training (ST6). Similarly, high quality feedback is unlikely when it is practically difficult, i.e. with statements of 10 words or less.

Discussion

The 2018 Consensus Framework for Good Assessment in medical education (Norcini et al 2018) proposes that for single assessments (e.g. workplace-based assessments) a

range of elements should be met depending on whether they are intended to be primarily formative or summative. Of these, the consensus framework proposes that formative assessment works best when it is embedded in the clinical workflow, provides specific and actionable feedback, is ongoing, and is timely. As argued above, the emphasis in our research was on validation of the written feedback in workplace-based assessments, where validation considers the quality of the feedback per se (i.e. its *potential* to promote improved trainee learning) and is a step removed from looking specifically at the actual effectiveness of the feedback in promoting learning. To address this validation goal, therefore, this research analysed the samples of Rad-DOPS forms using a putative profile of high quality characteristics of supervisors' written feedback.

Specific and Actionable Feedback

Only 5% of the feedback statements in 2,500 Rad-DOPS forms over four years achieved the quality standards adopted for the research. The results showed that these instances of high quality feedback were associated with certain trainee characteristics such as having low modal performance ratings or being in the earliest (i.e. core) stage of training. All of the high quality feedback statements were also associated with word counts of 11 or more words. However, it must be noted that the majority of trainees with low modal ratings did not receive suggestions for improvement or negative feedback, i.e. they did not receive appropriate prompts for reflection and planning ahead. Moreover, the large majority of extended feedback statements (word count ≥11) and all of the brief statements (comprising 35% of the total) also did not meet the standards for high quality.

Ongoing and Timely Assessments

With up to 75% of the assessments across the samples being recorded during the

training placements, albeit with peaks around halfway and 90% through, the assessments could be considered to be continuous in a manner appropriate to a formative purpose. However, there were clear signs of fatigue in the system as the 2016-17 sample revealed significant numbers of ST1, 2 and 4 trainees completing only one Rad-DOPS assessment. Further research would be needed to determine the extent to which workplace-based assessment is meeting the Acceptability element of the Consensus Framework (Norcini et al., 2018), in which the key stakeholders find the process and results to be credible.

The finding that up to 25% of the assessments were either at the very end of the placements or actually undertaken retrospectively, is a worrying aspect in relation to any claim of timeliness in serving a formative purpose. As reported in similar research by Rees et al. (2014), the intended 'real time' formative assessment design is thwarted by late or retrospective completions. This may suggest that many trainees prioritise the fulfilment of training obligations rather than the pursuit of useful learning experiences. It may also reflect what Dannefer (2013) sees as some trainees struggling to adapt to a culture of assessment *for* learning, as opposed to assessment *of* learning; a view not too distant from Torrance's (2007) argument that post-compulsory education is actually missing out on assessment *for* learning, having gone straight from assessment *of* learning to assessment *as* learning.

In writing about the conditions required for productive formative assessment and feedback, Carless (2013) considers frequent formative activity to be important for the development of trust, which he argues is central to the subsequent development of a 'transformative, dialogic learning environment' (91). However, this idea of frequent assessment contrasts with the findings that most of the trainees undertook the minimum required numbers of Rad-DOPS assessments. It is likely that the existence of a high-

stakes, summative review of progress (ARCP) at the end of each training year, based partly on the workplace-based assessments, undermines the environment for building trust; encouraging trainees instead to present a carefully curated portfolio of assessment evidence that demonstrates high performance throughout the preceding period (Viney et al. 2017).

In summary, for high quality feedback to be given, the conditions of being judged to need Direct or Indirect Supervision, having a modal score on the criteria of 4 or less or being in the early, core stage of training (ST1-ST3) were found to be almost always present (quasi-necessary); and statement word counts greater than 10 were found to be necessary. Also for this small proportion of statements, high quality feedback was clearly not associated with observed or implied competence (Independent Practice), a modal score of 6, final stage training (ST6) or statement word counts less than 11. However, so few in number (5%) were these high quality feedback statements that the results overall suggest that Rad-DOPS written feedback cannot be convincingly validated as an appropriate precursor for trainees to reflect upon and plan to improve their learning.

Preston et al (2019) have highlighted research showing that trainees are unhappy with the poor quality and tardiness of feedback in clinical training, and that they appreciate feedback that is less tick-boxing and more oriented to immediate suggestions for improvement in clinical assessment tasks. In considering the results of the current study, it is therefore tempting to reiterate the often-repeated call for more training for assessors in the delivery of quality formative feedback. However, the data (Table 2) suggest that individual assessors undertake around three assessments per year and any skills developed in training could reasonably be expected to deteriorate through infrequent application. It may therefore be worth experimenting with the format of the assessment

judgments for example by replacing the tick box ratings (currently behaving as proxy 'scores') with narrative comments designed to improve the meaningfulness of the feedback. Similarly, continuous messaging on the key elements of high quality feedback, and their potential impact on learning, especially perhaps for improving trainees' feedback literacy and understanding of the role of workplace-based assessments, may also prove valuable.

Concluding Remarks

Medical education is arguably dominated by high stakes assessment and may remain so for years to come unless progress is made on fostering a culture of 'nurture and of professional educational support' (GMC, 2010, p. 1). There is an urgent need to reduce the impacts of competition, high stakes examinations and the duality of workplace-based assessment purposes on the quality of learning and its outcomes. We feel that adopting such formative processes as co-construction of learning activities, dialogic feedback in communities of practice, and trainee peer and self-assessment, could move workplace-based assessments to a position in which the 'quality of engagement [with learning] that it helps to secure and to shape is personally, institutionally and/or socially valuable' (Newton 2017, 5). Curriculum designers in the UK have attempted to address this challenge by introducing a form of workplace-based assessment known as a supervised learning event (SLE) which is intended to have formative impact on progress but no direct impact on decisions about progression (Cho et al., 2014). It remains to be seen whether assessors and trainees embrace them accordingly. For the moment, though, this research has highlighted important concerns that need to be addressed in relation to the current models of formative assessment and, specifically, written feedback in clinical workplace-based assessments.

Acknowledgements:

The authors wish to thank the Royal College of Radiologists for their permission to use the anonymised Rad-DOPS data and also the anonymous reviewers for their very helpful comments and suggestions.

References

AoMRC. 2016. “Improving Assessment: Further Guidance and Recommendations”. London, Academy of Medical Royal Colleges.

Archer, J., M. McGraw, and H. Davies. 2010. "Assuring validity of multisource feedback in a national programme." *Arch Dis Child* 95 (5):330-5. doi: 10.1136/adc.2008.146209.

ARG. 2002. Assessment for Learning: 10 principles. Research-based principles to guide classroom practice. London, Assessment Reform Group. https://www.researchgate.net/publication/271849158

Augustine, K., P. McCoubrie, J. R. Wilkinson, and L. McKnight. 2010. "Workplace-based assessment in radiology-where to now?" *Clin Radiol* 65 (4):325-32. doi: 10.1016/j.crad.2009.12.004.

Barrett, A., R. Galvin, Y. Steinert, A. Scherpbier, A. O’Shaughnessy, G. Walsh and M. Horgan 2016 Profiling postgraduate workplace-based assessment implementation in Ireland: a retrospective chort study. SpringerPlus 5:133 doi: 10.1186/s40064-016-1748-x

Bennett, R. E. 2011. "Formative assessment: a critical review." *Assessment in Education: Principles, Policy & Practice* 18 (1):5-25. doi: 10.1080/0969594X.2010.513678.

Black, P. and D. Wiliam. 1998. “Assessment and Classroom Learning,” *Assessment in Education: Principles, Policy & Practice*, 5(1), pp. 7-74.

Bloxham, S. and L. Campbell. 2010. "Generating dialogue in assessment feedback: exploring the use of interactive cover sheets." *Assessment & Evaluation in Higher Education* 35 (3):291-300. doi: 10.1080/02602931003650045.

Bloom, B. S., J.T. Hastings and G. Madaus. 1971. *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Brann, M., and M. Mattson. 2004. "Reframing Communication during Gynaecological Exams: A Feminist Virtue Ethic of Care Perspective." In *Gender in Applied Communication Contexts*, edited by H. Sterk P.M. Buzzanell, L.H. Turner, 147-168. London: Sage Publications Limited.

Burford, B., J. Illing, C. Kergon, G. Morrow, and M. Livingston. 2010. "User perceptions of multi-source feedback tools for junior doctors." *Medical Education* 44 (2):165-76. doi: 10.1111/j.1365-2923.2009.03565.x.

Canavan, C., M.C. Holtman, M. Richmond, and P. Katsufrakis. 2010. “The Quality of Written Comments on Professional Behaviors in a Developmental Multisource Feedback Program.” *Academic Medicine*, 85(10 Suppl), pp. S106-S109.

Carless, D. 2013. "Trust and its role in facilitating dialogic feedback." In *Feedack in Higher and Professional Education: Understanding and doing it well.*, edited by D. and Molloy Boud, E, 90-103. London: Routledge.

Carless, D. 2006. “Differing Perceptions in the Feedback Process.” *Studies in Higher Education*, 31(2), pp. 219–233.

Carless, D., D. Salter, M. Yang and J. Lam. (2011) Developing sustainable feedback practices. *Studies in Higher Education,* 36:4, 395–407 doi: 10.1080/03075071003642449

Cho, S.P., D. Parry and W. Wade. 2014. “Lessons learnt from a pilot of assessment for learning.” *Clinical Medicine* 14: 577-584. doi: 10.7861/clinmedicine.14-6-577.

Cohen, S. N., P. B. Farrant, and S. M. Taibjee. 2009. "Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools." *British Journal of Dermatology* 161 (1):34-9. doi: 10.1111/j.1365-2133.2009.09097.x.

Committee of Postgraduate Medical Deans. 2016. *A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide*. 6th Edition ed.: Conference of Postgraduate Medical Deans of the United Kingdom.

Crisp, B. R. 2007. "Is it worth the effort? How feedback influences students’ subsequent submission of assessable work." *Assessment & Evaluation in Higher Education* 32 (5):571-581. doi: 10.1080/02602930601116912.

Crooks, T. J. 1988. “The Impact of Classroom Evaluation Practices on Students.” *Review of Educational Research*, 58(4), 438–481. https://doi.org/10.3102/00346543058004438

Dannefer, E. F. 2013. "Beyond assessment of learning toward assessment for learning: educating tomorrow's physicians." *Medical Teacher* 35 (7):560-3. doi: 10.3109/0142159x.2013.787141.

Driessen, E. and F. Scheele 2013 What is wrong with assessment in postgraduate training. Lessons from clinical practice and educational research. *Medical Teacher*, 35:7, 569-574, doi:10.3109/0142159X.2013.798403

Evans, C. 2011. Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research*, 83:1, 70–120 doi: 10.3102/0034654312474350

Fernando, N., J. Cleland, H. McKenzie, and K. Cassar. 2008. "Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments." *Medical Education* 42 (1):89-95. doi: 10.1111/j.1365-2923.2007.02939.x.

Gardner, J. 2012. "Assessment and Learning: Introduction." In *Assessment and Learning*, edited by John Gardner, 1-8. London: Sage Publications Limited.

Glaesser, J., and B. Cooper. 2012. "Educational achievement in selective and comprehensive local education authorities : a configurational analysis." *British Journal of Sociology of Education.* 33 (2):223-244.

GMC. 1993. *Tomorrow's Doctors*. London: The General Medical Council.

GMC. 2010. *Workplace Based Assessment: A Guide for Implementation*. Manchester.

GMC. 2013. *National training survey 2013: key findings*. Manchester.

GMC. 2017a. *Designing and maintaining postgraduate assessment programmes*. Manchester, General Medical Council.

GMC. 2017b. *Excellence by design: Standards for postgraduate curricula*. Manchester, General Medical Council.

Graneheim, U. H., and B. Lundman. 2004. "Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness." *Nurse Education Today* 24 (2):105-12. doi: 10.1016/j.nedt.2003.10.001.

Hattie, J. and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1):81-112. doi: 10.3102/003465430298487.

Holmboe, E. S., M. Yepes, F. Williams, and S. J. Huot. 2004. "Feedback and the mini clinical evaluation exercise." *Journal of General Internal Medicine* 19 (5 Pt 2) 558-61. doi: 10.1111/j.1525-1497.2004.30134.x.

Johnson, G., J. Barrett, M. Jones, D. Parry, and W. Wade. 2008. "Feedback from educational supervisors and trainees on the implementation of curricula and the assessment system for core medical training." *Clinical Medicine (Lond)* 8 (5):484-9.

Jolly, B., and D.B. Boud. 2013. "Written Feedback. What it is Good for and How Can We Do it Well?" In *Feedback in Higher and Professional Education: Understanding it and Doing it Well*, edited by and E. Molloy and D. Boud, 104-124. London: Routledge.

Kane, M. 2006. "Validation." In *Educational Measurement*, edited by R.L. Brennan, 17-64. Washington D.C.: American Council on Education/Praeger.

Kluger, A. N., and A. DeNisi. 1996. "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory." *Psychological Bulletin* 119 (2):254-284. doi: 10.1037/0033-2909.119.2.254.

Legewie, N. 2013. An introduction to applied data analysis with qualitative comparative analysis (QCA) Forum: *Qualitative Social Research* 14 (3) Art 16

Massie, J. and M.J. Ali. 2016. "Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings." *Advances in Health Sciences Education Theory and Practice* 21(2):455-473. doi: 10.1007/s10459-015-9614-0

McKavanagh, P, A Smyth, and A Carragher. 2012. "Hospital consultants and workplace based assessments: how foundation doctors view these educational interactions?" *Postgraduate Medical Journal* 88 (1037):119-124. doi: 10.1136/postgradmedj-2011-130121.

Newton, P. E. 2007. "Clarifying the purposes of educational assessment." *Assessment in Education: Principles, Policy & Practice* 14 (2):149-170. doi: 10.1080/09695940701478321.

Newton, P. E. 2017. "There is more to educational measurement than measuring: the importance of embracing purpose pluralism." *Educational Measurement: Issues and Practice*, 36 (2) 5-15 doi.org/10.1111/emip.12146

Norcini, J., M. B. Anderson, V. Bollela, V. Burch, M.J. Costa, R. Duvivier, R. Hays, M.F.P. Mackay, T. Roberts and D. Swanson. 2018. "Consensus framework for good assessment." *Medical Teacher*. doi.org/10.1080/0142159X.2018.1500016

Orsmond, P., S. Merry, and K. Reiling. 2005. "Biology students' utilization of tutors' formative feedback: a qualitative interview study." *Assessment & Evaluation in Higher Education* 30 (4):369-386. doi: 10.1080/02602930500099177.

Perrenoud, P. 1998. "From Formative Evaluation to a Controlled Regulation of Learning Processes. Towards a wider conceptual field." *Assessment in Education: Principles, Policy & Practice* 5 (1):85-102. doi: 10.1080/0969595980050105.

Preston, R., M. Gratani, K. Owens, P. Roche, M. Zimanyi and B. Malau-Aduli. 2019. "Exploring the Impact of Assessment on Medical Students' Learning." *Assessment & Evaluation in Higher Education* doi: 10.1080/02602938.2019.1614145

Prins, F. J., D. M. Sluijsmans, and P. A. Kirschner. 2006. "Feedback for general practitioners in training: quality, styles, and preferences." *Advances in Health Sciences Education Theory and Practice* 11 (3):289-303. doi: 10.1007/s10459-005-3250-z.

Ragin, C.C. 2008. *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ramaprasad, A. 1983. "On the definition of feedback." *Behavioral Science* 28 (1):4-13. doi: 10.1002/bs.3830280103.

RCR. 2016. *Specialty training curriculum for clinical radiology*. The Faculty of Clinical Radiology, The Royal College of Radiologists.

Rees, C. E., J. A. Cleland, A. Dennis, N. Kelly, K. Mattick, and L. V. Monrouxe. 2014. "Supervised learning events in the Foundation Programme: a UK-wide narrative interview study." *BMJ Open* 4 (10):e005980. doi: 10.1136/bmjopen-2014-005980.

Sadler, D.R. 1989. "Formative assessment and the design of instructional systems." *Instructional Science*, 18(2):119-144.

Sargeant, J., K. Mann, and S. Ferrier. 2005. "Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness." *Medical Education* 39 (5):497-504. doi: 10.1111/j.1365-2929.2005.02124.x.

Sargeant, J., K. Mann, D. Sinclair, C. van der Vleuten, and J. Metsemakers. 2007. "Challenges in multisource feedback: intended and unintended outcomes." *Medical Education* 41 (6):583-91. doi: 10.1111/j.1365-2923.2007.02769.x.

Shepard, L.A. 2000. "The Role of Assessment in a Learning Culture." *Educational Researcher*, 29(7):4-14.

Shute, V. J. 2008. "Focus on Formative Feedback." *Review of Educational Research* 78 (1):153-189. doi: 10.3102/0034654307313795.

Swaffield, S. 2011. "Getting to the heart of authentic Assessment for Learning." *Assessment in Education: Principles, Policy & Practice* 18 (4):433-449. doi: 10.1080/0969594X.2011.582838.

Torrance, H. (2007) "Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning." *Assessment in Education*, 14(3) pp. 281-294.

Viney, R., A. Rich, S. Needleman, A. Griffin, and K. Woolf. 2017. “The validity of the Annual Review of Competence Progression: a qualitative interview study of the perceptions of junior doctors and their trainers.” *Journal of the Royal Society of Medicine*, 110(3), 110–117. doi:10.1177/0141076817690713

Vivekananda-Schmidt, P., L. MacKillop, J. Crossley, and W. Wade. 2013. "Do assessor comments on a multi-source feedback instrument provide learner-centred feedback?" *Medical Education* 47 (11):1080-8. doi: 10.1111/medu.12249.

Wilkinson, J. R., J. G. Crossley, A. Wragg, P. Mills, G. Cowan, and W. Wade. 2008. "Implementing workplace-based assessment across the medical specialties in the United Kingdom." *Medical Education* 42 (4):364-73. doi: 10.1111/j.1365-2923.2008.03010.x.