

# **Workplace-based assessment in clinical radiology in the UK – a validity study.**

Michael Thomas John Page

BSc(Hons), PGCE, MEd

Thesis submitted in fulfilment of the requirement for the  
Degree of Doctor of Philosophy

The University of Stirling

Faculty of Social Sciences

October 2016

## ABSTRACT

In 2010, the Royal College of Radiologists introduced workplace-based assessments to the postgraduate training pathway for clinical radiologists in the UK. Whilst the system served the purpose of contributing to high-stakes annual judgements about radiology trainees' progression into subsequent years of training, it was primarily intended to be formative. This study was prompted by an interest in whether the new system fulfilled this formative role.

Data collection and analysis spanned the first three years of the new system and followed a multi-methods approach. Descriptive statistical analysis was used to explore important parameters such as the timing and number of assessments undertaken by trainees and assessors. Using the literature and an iterative analysis of a large sample of trainee data, a coding framework for categories of feedback quality enabled assessors' written comments to be explored using deductive and inductive qualitative analysis, with inferential statistical analysis of coded assessor feedback statements. For example, Ragin's (1987, 2000, 2008) qualitative comparative analysis, QCA, was used to explore whether the assessments met necessary and/or sufficient conditions for high quality feedback. Pairs of assessor-trainee feedback comments were also analysed to establish whether any dialogic feedback interactions occurred.

The study presents evidence that despite its intentions, the new system is generally failing to meet its primary, formative aim. As a consequence, the influence of negative washback on assessment practice was reflected in a number of findings. For example, there was evidence of trainees taking an instrumental approach to the assessments, undertaking only the prescribed minimum of assessments or completing assessments in the later stages of placements. Combined with evidence of retrospective assessment, i.e. after completion of the placements, the observed patterns of assessment over the three years are consistent with a box-ticking approach. This study explores the contextual policy and practice dimensions underpinning these and related findings and discusses the implications and recommendations for future arrangements.

## ACKNOWLEDGEMENTS

I wish to thank my supervisor, Professor John Gardner, for his support, guidance and encouragement throughout this research. He has pushed me to develop as a researcher and as a writer, and has provided perceptive and helpful feedback throughout the project. His insights and advice were instrumental in guiding my thinking as the work progressed, and he has made himself available at all times to discuss all aspects of the work.

I also wish to thank my second supervisor, Professor Cate Watson, for her excellent insights and feedback. She pushed me to justify my approach and brought a fresh perspective to the work.

I am indebted to the Royal College of Radiologists, who were so generous in providing access to the data at the heart of my study. Particular thanks are due to Joe Booth, Executive Director of Specialty Training, for his patience in responding to numerous emails and phone calls over the course of the last several years.

I wish to offer sincere thanks to a number of staff at the University of Stirling. Lorna Valentine, Vicki Lawlor and Donna Caldwell have all played a vital role in scheduling meetings and facilitating the exchange of drafts and feedback, and have done so with professionalism and good humour.

Thanks are also due to a number of colleagues and friends who, in different ways, have been a vital source of support. David Parry, Deputy Director of Education at the Royal College of Physicians, has offered great advice and reassurance, and has lifted my spirits on countless occasions with his humour. Sharon Jamieson, Librarian at the Royal College of Physicians (and native of Stirling), supported me in my nascent attempts at conducting proper literature searches, and has provided regular updates on recent publications in the field of medical education. My friend, Mark Millar, played a more important role than he knows in helping me move forward with my work.

Finally, a massive thank you to Laura, Eva, Isla and Jessica. Without you, none of this would mean anything. Laura – you have been a constant encouragement, and have given me more support than I have any right to expect. I am truly blessed!

# TABLE OF CONTENTS

Abstract	2
Acknowledgements	3
Table of contents	4
Chapter 1 Introduction	9
1.1 Setting the scene	9
1.1.1 Drivers of change	9
1.1.2 Impact on clinical radiology training	11
1.1.3 The changing face of assessment in clinical radiology	12
1.2 The GMC and workplace-based assessment	13
1.3. Implementing workplace-based assessment in clinical radiology	15
1.4 Transforming assessment in clinical radiology	17
1.5 My interest in workplace-based assessment and feedback	19
1.6 Overview of the research	24
1.6.1 Research objectives	24
1.6.2 Research questions	27
1.6.3 Design of the research	27
1.7 Outline of the thesis	28
Chapter 2 Literature review Part I – Assessment in teaching and learning	30
2.1 Introduction	30
2.2 Assessment	32
2.2.1 Summative assessment	34
2.2.2 Formative assessment	36
2.3 Validity in assessment	37
2.3.1 Components of validity	39
2.3.2 Validity as a function of assessment use	40
2.3.3 Uses and purposes of assessment	41
2.3.4 Formative assessment for summative purposes	44
2.4 Washback and systemic validity	45
2.4.1 Positive and negative washback	46
2.4.2 Washback as a feature of high-stakes assessment	47
2.4.3 Working for washback	51
2.5 The influence of the system	57
2.5.1 Radiology training as professional learning	58

2.5.2 Radiology training as competency-based education	59
2.5.3 Competency-based education and instrumentalist approaches	62
2.6 Summary	64
Chapter 3 Literature review Part II – Feedback in medical education	65
3.1 Introduction	65
3.2 Why analyse feedback in workplace-based assessment?	65
3.2.1 Measuring quality in healthcare – a parallel case	67
3.2.2 Feedback as a suitable process measure for formative assessment	69
3.2.3 What is the current state of feedback in medical education?	70
3.3 What is feedback?	72
3.3.1 Feedback as information	73
3.3.2 Feedback as process	74
3.4 Written feedback in workplace-based assessment	78
3.5 Why is feedback necessary?	81
3.6 Is feedback effective?	84
3.6.1 The importance of the feedback message	85
3.6.2 The importance of the feedback source	90
3.6.3 What are the important characteristics of learners?	95
3.7 Judging the quality of feedback	98
3.8 Conclusions	103
Chapter 4 Methodology	105
4.1 Introduction	105
4.2 Summary of the research questions	105
4.3 Which assessment features?	107
4.4 Which workplace-based assessment?	109
4.5 Study design	110
4.5.1 What type of study?	110
4.5.2 Planning the research	112
4.5.3 What paradigm?	114
4.6 Data collection	115
4.6.1 Narrative review of literature	115
4.6.2 The nature of the Rad-DOPS workplace-based assessment data	117
4.6.3 Preparing for analysis of written feedback	118
4.6.4 Generating the sample	119
4.7 Data analysis	119

4.7.1	Descriptive statistical analysis	119
4.7.2	Content analysis of assessors' written feedback	124
4.7.3	The content analysis research process	126
4.7.4	Constructing an initial framework for coding assessor comments	127
4.7.5	Analysing published frameworks	128
4.7.6	Applying the coding framework	132
4.7.7	Defining the units of analysis	132
4.7.8	Defining the meaning units	133
4.7.9	Reducing the data	134
4.7.10	Problems with content analysis	136
4.8	Trustworthiness of the research	137
4.8.1	Establishing rigour in qualitative research	137
4.8.2	Establishing validity in this research	140
4.9	Scaling up the research	145
4.9.1	Expanding the coding process horizontally	146
4.9.2	Expanding the coding process vertically	146
4.10	Statistical analysis	148
4.10.1	Analysing the relationships between years	148
4.10.2	Analysing relationships between feedback and other assessment parameters	148
4.10.3	Length of feedback – brief or extended?	150
4.11	Exploring conditions for effective feedback	153
4.11.1	Ragin's approach – necessity and sufficiency	153
4.11.2	Quasi-sufficiency and quasi-necessity	154
4.11.3	Choosing an outcome	156
4.11.4	Necessity or sufficiency?	157
4.12	Analysing trainee comments	158
4.13	Summary	158
Chapter 5	Results	160
5.1	Introduction	160
5.2	Descriptive statistics	162
5.2.1	Patterns of assessment across all training grades	162
5.2.2	Patterns of assessments conducted by assessors	166
5.2.3	Timing of assessments	167
5.3	Content analysis of assessors' written comments	169

5.3.1	Establishing the representativeness of the sample	169
5.3.2	Comparing feedback characteristics across three successive years	173
5.3.3	Quality of written feedback – examples of assessors' comments	175
5.4	Inferential statistical analysis	181
5.4.1	Overall competence rating and type of feedback	181
5.4.2	Modal score and type of feedback	185
5.4.3	Stage of training and type of feedback	188
5.4.4	Length of feedback and type of feedback	188
5.5	Qualitative comparative analysis	191
5.5.1	Average assessment scores as a condition for high quality feedback	192
5.5.2	Stage of training as a condition for high quality feedback	193
5.5.3	Overall competence rating as a condition for high quality feedback	195
5.5.4	Length of feedback as a condition for high quality feedback	196
5.5.5	Summary of Ragin analysis	198
5.6	Analysing trainee comments	199
5.6.1	Themes emerging from the analysis of trainee comments	200
5.6.2	Identifying dialogical feedback exchanges	208
5.6.3	Missed opportunities	212
5.6.4	Summary of analysis of trainee comments	214
5.7	Summary of key findings	215
Chapter 6	Discussion	219
6.1	Introduction	219
6.2	Do trainees and assessors use appear to use workplace-based assessments formatively?	219
6.2.1	Timing of the assessments	220
6.2.2	Frequency of the assessments	222
6.2.3	Can written feedback in workplace-based assessments support the development of competence?	223
6.3	The broader picture of formative assessment in clinical radiology	229
6.3.1	Instrumentalism and assessment as learning	229
6.3.2	Assessment as pedagogy	231
6.3.3	Peer assessment and self-assessment	232
6.3.4	The role of the teacher and learner in formative assessment	234
6.3.5	Constitutive assessment and the long shadow of the ARCP	237
6.3.6	Summary	238

Chapter 7 Concluding remarks	239
7.1 Introduction	239
7.2 Wider implications	239
7.2.1 Fragmentation of professional competence	239
7.2.2 Distortion of formative assessment	241
7.2.3 A system in decline	241
7.3 Recommendations	242
7.3.1 Broadening the concept of formative assessment	242
7.3.2 Limiting negative washback	243
7.3.3 Conceptualising ‘learning’ in assessment for learning	244
7.3.4 Conceptualising clinical radiology	246
Chapter 8 Limitations and reflections	248
8.1 Limitations of the study	248
8.2 Personal reflections	249
References	251
Appendices	272
Appendix 1 Rad-DOPS guidance for assessors	272
Appendix 2 Rad-DOPS assessment form	273
Appendix 3 Clinical Radiology ARCP decision aid	275
Appendix 4 Evolution of the coding framework – assessor comments	276
Appendix 5 Confirmation of ethics approval	280



# CHAPTER 1

## 1. Introduction

### 1.1 Setting the scene

The last 20 years have seen a transformation in the approach to postgraduate medical education and training (PGMET) in the UK. Iobst *et al.* (2010) identified the origins of this change as having been rooted in the publication of the General Medical Council's 'Tomorrow's Doctors' document (General Medical Council, 1993) which set out a number of recommendations for the development of undergraduate medical curricula. Aimed at undergraduate rather than postgraduate medical education, this document described for the first time the expectations of the UK's medical regulator in regard to the essential outcomes of the undergraduate educational effort. The effect was that UK medical schools were compelled to shift their focus away from content and process, and instead focus on, for want of a more humanistic term, the 'product' of their courses. In short, it signalled the dawn of what is commonly referred to as competency-based education (CBE) in UK undergraduate medical education - an approach which duly made its way into postgraduate training. Thus, something of a paradigm shift occurred in postgraduate training, from an emphasis on a qualification process based on the length of time served as an 'apprentice', to an approach that was supposed to be more closely connected to the objective and verifiable development of competence<sup>1</sup> in individuals.

#### 1.1.1 Drivers of change

This transition from time-served apprenticeship to competency-based training has largely been driven by the high profile interventions of successive chief medical officers

---

<sup>1</sup> The term 'competence' is contested within medical education literature and is sometimes contrasted with terms such as 'mastery' or 'performance' in order to imbue it with a particular meaning. In this study I have at times used 'competence' in the general sense to refer to professional capability, and at other times used it to refer to the behavioural statements found within medical curricula ('competences'). Context should make the usage clear.

(CMO). The first, produced by Sir Kenneth Calman and known informally as 'The Calman Report' (Department of Health, 1993), mandated the introduction of curricula for all specialist registrars, along with a regime of assessment and feedback that, hitherto, did not exist in postgraduate medical training. Calman's reforms have been reported to have delivered an improvement in the educational experience for many trainee doctors (Paice *et al.* 2000), however the reforms only went so far. The Calman curricula largely described the experiences to which trainee doctors should be exposed, rather than the related learning outcomes or competencies they should achieve. There were no tools created for the purpose of assessment - most specialties opted for a log book in which trainees were expected to record their experiences in successive clinical attachments, with assessment occurring once a year by face-to-face interview. The model of feedback was conceived simply as 'regular informal discussions between the trainee and the supervising consultant about the trainee's progress' (Paice *et al.* 2000, p. 833).

The second report, produced by Calman's successor, Sir Liam Donaldson, set out a number of proposals for the further reform of postgraduate medical training in the NHS (Department of Health, 2002). Although it was specifically the educational lot of doctors in the Senior House Officer (SHO) grade that Donaldson initially aimed to improve, he understood that there was actually unfinished educational business to be dealt with across all training grades. The result was the introduction of a much more comprehensive strategy, known as Modernising Medical Careers (MMC) (Department of Health, 2004) which impacted significantly on the educational experience of doctors at all stages of training.

Alongside these two high profile interventions, another important factor was the introduction of the European Working Time Regulations (EWTR), and the impact of these on hospital rotas. As well as the apparently beneficial effect of limiting junior doctors' working hours, the regulations had a depleting effect on the time available for doctors to train. Consequently, adequate exposure to a sufficiently wide array of patients, conditions and interventions could no longer be assumed and it became clear that doctors in training would have to follow a much more carefully described educational process (lobst *et al.*, 2010). However, the EWTR may well have precipitated a change that was already in the making – lobst *et al.* (*ibid.*) also noted that there was a growing disquiet with a 'time served' approach to postgraduate

medical training, and that this was the case not only in the UK, but in countries, including the US, which were beyond the reach of the EWTR. Thus, various initiatives in the UK and internationally aimed to transform the nature of the training process from one in which postgraduate doctors were expected, given the correct clinical environment, to be able to spontaneously develop the necessary knowledge and skills required for expert medical practice, into a much more detailed educational process. Carraccio *et al.* (2002) describe the change as 'a paradigm shift from structure- and process-based to competency-based education' (p. 361), and observe that, as a direct result, the 'measurement of outcomes' was mandated by the approach (p. 361). Clearly, these changes had implications for assessment of doctors in training.

### *1.1.2 Impact on clinical radiology training*

In May 2010, and in response to these major policy shifts in the conception of, and approach to, postgraduate medical education, the UK Royal College of Radiologists (RCR) launched a new curriculum (RCR, 2010) for postgraduate clinical radiology training. It differed from the previous curriculum (RCR, 2007) in a number of respects. Not least, there was a significant change in the structure of the syllabus: the lists of subject-specific content that appeared in the 2007 version were, in the 2010 version, labelled as 'competences' (RCR, 2010 p.8, *ibid.*), and were categorised as being either knowledge-based competences, skills-based competences or behavioural competences (p. 27, *ibid.*). Whilst this may appear little more than a reorganisation of content, the re-structuring required the curriculum architects – doctors, for the most part – to consider, in detail, what professional capabilities trainee radiologists should be able to demonstrate, rather than specifying purely the things that they should know, or to which they should have had exposure. These extensive lists of competences – the syllabus in the 2010 curriculum spanned some 133 pages – were also mapped to relevant aspects of the GMC's regulatory framework document, known as 'Good Medical Practice' (GMP) (GMC, 2006; GMC 2013). This document outlines the main professional responsibilities of doctors practising in the UK, and articulates the standards by which doctors may be said to be fit to practise, or otherwise. The curriculum competences were also mapped to assessment methods, and it was in respect to assessment that, arguably, the greatest difference between the 2007 and 2010 curricula could be detected.

### 1.1.3 The changing face of assessment in radiology

The approach to assessment in clinical radiology training prior to 2010 consisted almost solely of the Fellowship of the Royal College of Radiologists (FRCR) exams, with assessment in the workplace being comprised of nothing more than a log-book of experience (see, for example, RCR, 2004; RCR, 2007). Judgements on trainees' competence were made at the end of each block of training by the trainee's supervisor (RCR, 2007) and the totality of the trainee's experience was reviewed on an annual basis by a panel of senior doctors (*ibid.*). However, in 2010, the RCR launched a suite of workplace-based assessments which, for the first time in UK clinical radiology training<sup>1</sup>, offered a structured approach to the assessment of trainees' performance in the clinical setting. Trainees' performance in these assessments, along with their achievements in the three-stage FRCR exams, were to be evaluated yearly through a process known as the Annual Review of Competency Progression (ARCP) (RCR, 2010). The ARCP panel would consist of a number of senior doctors who would make a decision about whether or not trainees could move to the next year of specialty training. Given the link between these new assessments and high stakes decisions about progression through training, it might appear that WBA would have a mainly summative function. However, the RCR declared that these assessments should in fact have a primarily formative function. Specifically, the assessments were intended to provide feedback, which, according to the RCR, plays a key role in educational development:

The [workplace-based] assessment tools are designed to help doctors develop and improve their performance. Feedback is a key factor to enable this to happen. (RCR, 2010, p. 5)

Feedback is a key component of the interactions between supervisors and radiology trainees. Giving and receiving feedback...are...part of an effective professional learning environment. *Improvement in clinical radiological practice will only happen if regular review leads to constructive feedback...*It is essential that trainers provide, and trainees receive, structured feedback (RCR, 2010, p. 158, emphasis mine).

The basis on which the RCR make these claims is unclear, as no references are cited within the curriculum. However, the curriculum was 'designed in line with the GMC standards' for postgraduate curricula and workplace-based assessment (RCR, 2010, p.

3), and so an examination of the GMC's guidance is necessary to scrutinise the evidence upon which these claims were based.

## **1.2 The GMC and workplace-based assessment**

According to guidance from the General Medical Council (2010) on the principal function of workplace-based assessment:

The primary purpose of WBAs is to provide constructive feedback – assessment for learning for the trainee. (GMC, 2010, p. 5).

In issuing this guidance, the GMC have appropriated the term 'assessment for learning', which has been much researched and debated in the broader educational literature over the past 25 years, not least by the members of the Assessment Reform Group (see, for example, Gardner, 2012a). The GMC appear to use the term as a surrogate for 'assessment which provides constructive feedback,' which, as noted by McDowell *et al.* (2009), is common within research and other academic writing in education. However, researchers in school- and higher education-based settings often prefer to take a more holistic view of assessment for learning, encompassing pedagogical concerns such as the washback effects on teaching and learning (see Stobart, 2012, for example), or the impact of assessment on learning downstream from the assessment (see Messick 1996, and Gardner, 2012b, for example). Seen in this light, the GMC's statement might be said to offer a somewhat impoverished conception of assessment for learning. The RCR have largely echoed the GMC's concept in their current radiology curriculum:

WBAs are formative assessments – assessments for learning – principally intended to support learning by providing feedback to trainees and helping to identify strengths and areas for development. (RCR, 2014, p. 12)

Regardless of the precise concept of assessment for learning that is held by the GMC and the RCR, both bodies have asserted the particular importance of feedback within the WBA process. They have done so in the context of a reasonable measure of support within the medical education literature. For example, in Wilkinson *et al.*'s (2008) pilot of WBA with trainees in the medical (i.e. physician) specialties in the UK,

participants reported favourably on aspects of the educational benefit of the assessments. The assessments piloted were the mini-clinical evaluation exercise (mini-CEX), the direct observation of procedural skills (DOPS) and multi-source feedback (MSF). Between 60-80% of trainees stated that feedback from workplace-based assessments provided them with new information about their practice (63%, 75% or 79%, depending on the type of workplace-based assessment, n=128 for mini-CEX, n=59 for DOPS and n=230 for MSF, respectively). The majority of trainees (74%, 75% or 80%, for mini-CEX, DOPS and MSF respectively) also stated that the assessment process *overall* was helpful to their personal development, although it is not entirely clear what specific aspects were felt to be helpful. Another study by Johnston *et al.* (2008), involving core medical trainees in the UK, found that just over half of the participants (59/94) perceived WBA to be a valuable source of feedback. These broadly positive outcomes are consistent with what has been found internationally. In the US, Holmboe *et al.* (2004b) reported that assessments of internal medicine trainees' consultation and examination skills provided a useful opportunity for senior doctors to give developmental feedback to junior colleagues, whilst Weller *et al.* (2009), evaluating the same type of assessment in anaesthetics training in New Zealand, found that workplace-based assessments facilitated feedback that was perceived by trainees to be educationally beneficial.

However, despite the existence of this evidence, the GMC guidance seems to draw on only one of these studies – conducted by Wilkinson *et al.* (2008) – directly. Consequently, it is not clear whether the guidance reflects empirical evidence of the value of feedback in medical education, or the general consensus of opinion within the working group who constructed the document. In any case, as will be argued later, the education literature is not consonant on the question of how feedback supports learning. It may be the case, therefore, that the GMC's rhetoric around WBA and feedback is driven more by the acceptance of certain truisms – e.g. that feedback promotes learning; that WBA supports the provision of feedback – than by empirical evidence regarding written feedback in the context of WBA and less formal feedback in practice. This latter – feedback in practice – might be described as interactional feedback, and to varying degrees may be experienced by trainees through engagement with senior colleagues. Nonetheless, the fact remains that the medical regulator in the UK has declared an important formative role for WBA. Accordingly, the

RCR, like the other medical royal colleges in the UK, have been tasked with operationalising WBA in postgraduate specialty training.

### **1.3 Implementing WBA in clinical radiology**

The clinical radiology specialty training curriculum (RCR, 2010) contains guidance for radiology training programmes throughout the UK as to how WBA should be implemented. The guidance is not always clear, as the operational message is at times integrated with statements that are more conceptual than practical. For example, the clear procedural expectation that 'at least 50% of WBAs will be undertaken with consultants' is immediately followed by, 'Each WBA should also be considered developmental and an opportunity for learning and feedback' (RCR, 2010 p. 10) The full guidance on conducting assessments is in fact dispersed throughout the curriculum. The broad guidelines for the conduct of WBA, summarised on pages 11 & 12 of the RCR's 2010 curriculum, are that:

- participation in assessment is mandatory for trainees
- minimum numbers of certain assessments are required in order for a trainee to be allowed to progress to the next year of training
- trainees are generally expected to exceed these minimum numbers
- assessors can be drawn from a range of clinical backgrounds, as long as they are competent in the domain being assessed
- the 'pattern of evidence' (*ibid.*, p. 12) from WBAs will be used as evidence when ARCP panels meet to make decisions about progression to the next stage of training.

More specific guidance is then found at the end of the curriculum. For the radiology direct observation of procedural skills (Rad-DOPS) assessment – the only WBA that assesses the trainee radiologist's performance with a patient – the guidance is that:

- the minimum numbers should be six Rad-DOPS assessments in each year of training
- different assessors should be used for each assessment encounter where possible
- assessors must be trained in giving feedback (although no particular training is specified) and understand the purpose of the assessment
- the assessments should be used to 'sample' the curriculum content across radiological problems and procedures
- the assessment arrangements (timing, medical problem/procedure and assessor) should be agreed in advance with the trainee
- assessors may also carry out unplanned assessments.

In the case of the last two recommendations, the advice seems to be conflicting. Further complexity is introduced by the existence of separate, even more detailed guidance which is appended to the assessment forms themselves. In the case of the Rad-DOPS, the guidance (see Appendix 1) states that the assessment should be a direct observation of a trainee's performance in the authentic clinical environment, and should be an observation of a specific procedure, rather than an observation of the trainee's performance over an extended period of time. It also states that the rating scale on the form should be interpreted in the context of what the assessor would expect from a trainee of similar experience and at a similar stage of training.

These observation-based assessments are intended to culminate in the completion of the appropriate documentation. This includes the trainee being rated on a six-point scale, which extends from 'well below expectation for stage of training' to 'well above expectation for stage of training', on a range of different domains (see Appendix 2 for an example of the Rad-DOPS assessment form). There are also two mandatory free text fields. The first of these is for the tutor's written feedback, which should include 'specific written comments on areas of good practise [*sic*] and constructive feedback on areas for further development' (p. 1, Rad-DOPS Guidance for Assessors, Appendix 1). The second mandatory field is for the trainee's comments, which should capture their own reflections on the assessment event, and which may include their reflections on the assessor's feedback comments. The guidance from the GMC is that, 'All assessors should make written records of feedback given and actions taken,' (GMC, 2010, p. 7),



which implies that the written feedback record should reflect the substance of any verbal feedback. However, this expectation is not made explicit in the RCR's assessment guidance.

#### **1.4 Transforming assessment in clinical radiology**

What is clear from the clinical radiology curriculum, and the RCR's guidelines on specific assessments, is that there is an expectation that WBA should give rise to feedback. The rationale, stated repeatedly by the RCR, is that this feedback constitutes the means by which radiology trainees can become proficient practitioners; the RCR maintains that 'frequent and timely feedback on performance is essential for work-based experiential learning' (RCR, 2010, p. 156). The educational environment within which this feedback occurs is also held to be important, and WBA is seen as being an essential element constituting this formative setting. The RCR's assessment guidance alluded to above is therefore nested within an overall methodology that emphasises a 'continuous assessment' approach (*ibid.* p. 160). – frequent assessments, conducted by a range of different assessors, in a range of different, real world clinical settings, across different domains of radiological practice and spread throughout the duration of the clinical attachment. This approach is something of a departure for all medics, including radiologists, for whom the introduction of WBA has signalled major changes in educational process and culture. The extent of the transformation is explored in more depth later in the review of literature. Suffice it to say, assessment in the postgraduate stage of medical education was dominated until the late 1990s by high stakes professional (or 'college') examinations, conducted away from the clinical setting, with a large written component, and in relatively recent times a simulated clinical component. However, the the introduction of the new curriculum in 2010, as the RCR make clear, marked a significant change in not just the assessment process, but the underlying educational philosophy and culture:

The curriculum has undergone wholesale re-design since 2007. There are fundamental changes in terms of the underpinning educational ethos [and] the development of mapped assessments (RCR, 2015, p. 192).

In the same year that the re-designed radiology curriculum was introduced, the GMC's assessment guidance acknowledged that the success of WBA would require a significant culture change within postgraduate medical training:

Throughout medical training, particularly where there are large numbers of candidates for relatively small numbers of places in a particular training programme, a competitive culture exists. Competition can make people wary of assessment, and efforts to provide feedback on progress and attainment can unintentionally be seen as threatening. One aim of this guide is to emphasise that WBA requires a change in that culture (GMC, 2010, p. 1).

The GMC guidance goes on to highlight the risks of failing to achieve cultural transformation:

In order for WBAs to be valid and useful, trainees and assessors need to understand and value their role in the educational process. The assessment tools and findings from WBAs must be used formatively and constructively. Without this understanding, WBA tools will potentially become no more than a series of external requirements and hoops to be jumped through, and the educational validity of the process will be lost (GMC, 2010, p. 3).

Thus, in 2010, senior doctors and trainees in clinical radiology were expected to engage in an educational process that was clearly novel, and which would necessitate a transformation in their understanding of the educational ethos and practices of postgraduate medical training, as well as a transformation in their behaviour in fulfilling new, more formalised educational roles. This transformation would, according to the medical regulator (GMC, 2010) and the doctors' own medical royal college (RCR, 2010), ensure appropriate support for the professional learning of radiology trainees, without which, it was maintained by the RCR (2010), development would not occur. In an attempt to bring about this change, the RCR commissioned training from the Royal College of Physicians (RCP) in 2009, and undertook to train a minimum of 10% of the consultant radiologists in the UK in WBA and feedback. It was in my role as an educationalist at the RCP that I became involved in delivering this training for radiology assessors, and consequently became interested in the extent to which this new approach to assessment and feedback could be said to be 'working'.

## **1.5 My interest in workplace-based assessment and feedback – rationale for the study**

I have been employed as an educationalist at the Royal College of Physicians of London since June 2009, and in this role have worked extensively with doctors of all specialties to develop their practice as medical educators. An aspect of this work, from late 2009 to mid-2010, involved delivering a series of workshops on workplace-based assessment and feedback for consultant radiologists throughout the UK, on behalf of the Royal College of Radiologists. Most of this training was delivered prior to the launch of the workplace-based assessment programme in clinical radiology, and was intended to equip senior doctors to act as assessors for their trainees in the clinical environment. An emphasis throughout the training was on the formative dimension of workplace-based assessment, with a particular focus on the provision of feedback. Consequently, I was interested in the quality of the feedback provided to trainees and the extent to which the new workplace-based assessment system could be said to be supportive of trainees' learning. However, through my interactions with assessors in clinical radiology, as well as my work with several hundred assessors in other medical specialties, I have become increasingly concerned that the outcomes of these assessment interactions, and the environment within which they are conducted, are not as intended.

In addition to delivering this training for radiologists, over the last six years I have delivered in excess of 30 workshops on workplace-based assessment and feedback for doctors of all specialties throughout the UK. In the course of this work, doctors of all grades and clinical backgrounds have expressed a range of views on the effectiveness of workplace-based assessment in practice. These views have included positive reports on the utility of the assessment and feedback process, however my experience has been that a much greater proportion of the discourse has been negative. Consultants attending these workshops frequently describe: being asked by trainees for retrospective assessments, relating to clinical encounters that they can no longer properly recall; being asked for multiple assessments at the end of a training attachment, as trainees attempt to record a required number of each type of assessment; the assessments being little more than a tick-box exercise, which are more to do with satisfying a requirement for minimum numbers than gaining useful feedback; their belief that the assessments offer little of any real educational value, with the outcomes often not bearing much resemblance to the genuine capability of the

trainee concerned. These doctors also frequently identify themselves as being the educationally engaged members of their clinical teams, and refer to colleagues who will be expected to bear the same educational responsibility as being much less engaged, or even opposed to the changes being introduced. Trainees who have attended the workshops (albeit in smaller numbers) have also expressed discontent. They report: a lack of senior colleagues being available to complete their assessments; requests for assessments going unheeded even when the consultants do appear to be available; a lack of helpful feedback – verbal or written – when assessments are conducted. These trainees often also use the phrase 'tick-box exercise' to summarily dismiss the formative element of the workplace-based assessment process.

My concerns have been further fuelled by the findings from the national e-portfolio record for trainees in the specialties governed by the RCP. An initial (unpublished) analysis of assessment data recorded by these trainees, which I conducted as part of a preliminary investigation within the education department at the Royal College of Physicians, demonstrated that, of 30,969 mini-clinical evaluation exercise (mini-CEX) workplace-based assessments, undertaken by core medical trainees (CMT) between August 2009 and August 2010:

- 1926 (6%) contained no positive feedback
- 4958 (16.0%) contained no suggestions for improvement
- 6100 (20%) contained no action plan for further learning

On initial inspection, the figures appear encouraging – for example, 94% of assessments contained positive feedback; 84% contained suggestions for improvement. On closer inspection, the import of the findings is less compelling. For example, of the 80% of assessments that *did* contain an action plan, many of the action planning fields contained comments that were limited to phrases such as: 'Keep going,' 'See more patients,' 'Get more experience', and many associated variations.

The reasons for this paucity of feedback are likely to be complex, and will be explored further in the formal review of literature for this study. However, they appear to span, amongst other aspects, the environmental, relational, cultural, attitudinal, philosophical and intellectual domains of educational practice in postgraduate medical training. For example, anecdotal evidence from doctors with whom I have worked over the past six

years has suggested that reforms to the ways of working in most medical and surgical specialties (briefly, the expansion of shift working and the introduction of numerous different rota patterns – see Blundell *et al.* (2011) for further exploration of this) have fragmented the working relationships between trainee doctors and their senior colleagues. Accordingly, there is a lack of continuity in the educational relationship and, therefore, limited scope for follow-up on feedback. That said, when a degree of continuity *does* exist, as may be the case with the most senior trainees and their consultant colleagues, the relationship can often be one of interdependence: consultants often rely on the senior trainees to run many aspects of the day-to-day delivery of clinical care, and senior trainees rely on the consultants for educational support and guidance. The result is that, anecdotally, consultants report a fear of jeopardising a close working relationship through the provision of negative feedback, despite their desire to perform their educational role effectively. Even when a close working relationship does not exist, doctors are not keen to deliver formal criticism of their colleagues. As articulated by one medical assessor in Rees *et al.*'s (2009) exploration of doctors' reluctance to give negative assessment judgements, 'You don't want to sort of be the one who sticks the knife in them' (p. 5). Many consultant physicians have also reported being somewhat mystified by the new educational approach. In attending the workshops that I deliver, experienced consultants are often receiving assessment training for the first time, despite the WBA system having been in place for the physician specialties since 2007 (Wilkinson *et al.* 2008). Many appear to be unclear about the role of WBA, and their role as assessors, have spent the time between the introduction of WBA and attending training dealing with the new assessments as best they can.

These anecdotal accounts are supported by some of the findings reported in the medical education literature. Fernando *et al.* (2008), for example, found that, despite medical assessors in their study being provided with a structured approach to workplace-based assessment, feedback was often absent, or was of limited value when it was provided. Cohen *et al.*'s (2009) study with dermatology trainees in the UK found that feedback linked to a number of different types of workplace-based assessment - the mini-clinical evaluation exercise (mini-CEX), the direct observation of procedural skills (DOPS) and multi-source feedback (MSF) - was at times not useful. Holmboe *et al.* (2004b), whose study was mentioned previously, tempered their positive findings by reporting that feedback resulted in an action plan being formulated

in only 8% of the assessment encounters, despite 80% of the assessments containing at least one suggestion for improvement. Other criticisms in the literature include Archer *et al.*'s (2010) finding that MSF often fails to provide sufficient feedback for struggling trainees. Moreover, Bullock *et al.*'s (2009) finding that MSF feedback stringency tended to correlate positively with the seniority of the assessors raises questions about the accuracy and reliability of WBA – an assessment can hardly be said to be accurate or reliable if the seniority of the assessor has a demonstrable impact on the judgement recorded in the assessment. Researchers conducting an analysis of written feedback on MSF forms in the US found a failure to provide feedback of sufficient quality to support learning (Canavan *et al.*, 2010) with the authors judging quality against a theoretically-derived framework. A UK-based study on MSF feedback for so called 'staff grade' doctors (i.e. those who are neither on a recognised training programme, nor qualified as consultants), conducted by Vivekananda-Schmidt *et al.* (2013) found that assessors' comments 'rarely contain enough detail to illustrate the problem or to show where change to current practice is required and how it might be enacted' (*ibid.*, p. 1086).

It is yet more challenging to determine whether, even when perceived to be educationally beneficial, the feedback that is given to trainee doctors actually results in learning. Archer, McGraw & Davies (2010) failed to demonstrate a link between MSF feedback and any subsequent change in practice by the recipients. A comparative study conducted by Burford *et al.* (2010) which found a marked preference amongst trainees for written rather than numerical feedback nonetheless found that fewer than a third of participants intended to respond to the written feedback they had received. Sargeant *et al.* (2007) found that only half of hospital doctors participating in their study intended to change their behaviour in response to negative feedback on an MSF. The same authors had previously found a similar unwillingness to respond to negative feedback amongst family physicians (Sargeant *et al.*, 2005). In addition, I have been unable to identify any evidence in the literature that doctors who do intend to make a change do so in a way that is verifiable, with authors such as Saedon *et al.*, (2012) acknowledging the heavily confounded nature of any prospective studies that might be attempted.

Taken together, it seems that the weight of the anecdotal evidence, and a significant proportion of the medical educational literature, along with my own initial inspection of

empirical data drawn from the RCP e-portfolio, would support the conclusion that the workplace-based assessment process in clinical radiology is likely to be of little value. However, clinical radiology began the WBA process with two key advantages over other specialties, including the physician specialties. The first is the retention of much of the traditional 'firm' structure, in which trainees work closely with a well-defined team of consultants and fellow trainees. This, as reported by Blundell *et al.* (2011), is in contrast to the majority of medical and surgical specialties, in which the stable relationship between trainee and trainer has been fragmented, creating a 'lack of continuity in teaching' (p. 122). Consequently, it might be expected that the more stable, consistent training relationship in clinical radiology would provide the basis for a higher quality of assessment and feedback than has been reported in the physician specialities. The second advantage is that the RCR launched their curriculum having already trained in excess of 10% of the consultant radiologist cohort directly. They also provided a range of resources that could be used by workshop attendees to cascade the training to their colleagues. This was in contrast to the physicians who, given their much larger numbers and their early adoption of the new assessment system, had trained a much smaller proportion of their consultants by the time the system was launched, and then continued to deliver training through face-to-face workshops, with no formal support to allow delegates to 'cascade' the training at a local level.

In summary, therefore, my interest in the workplace-based assessment and feedback process was driven by four main considerations. The first was the anecdotal evidence from RCP and RCR workshop delegates which indicated that many doctors were struggling to get to grips with the new approach to formative assessment and feedback in postgraduate medical education. The second was my informal inspection of written feedback within the RCP's e-portfolio, which gave rise to considerable concern regarding the quality of feedback being provided to trainees in the real world setting of postgraduate physicians' training. The third was the dissonance within the medical education literature with regard to the quality and effectiveness of the formal feedback provided to trainees in the course of WBA. As the literature review chapters in this thesis will establish, a number of these articles were based on small scale studies, or on pilot projects which may have emphasised efficacy over effectiveness, thereby demonstrating a need for further empirical research. The fourth was the potential for the RCR, by virtue of their late adoption of WBA and their retention of much of the traditional apprenticeship structure within clinical training, to implement WBA in manner

that was more closely aligned with the original formative aims of the process. However, the discourse within the RCR's WBA training workshops was similar to that within the physicians' workshops, suggesting that these radiology doctors may also struggle to implement the process for reasons that were not dissimilar to those found within physicians' training. Consequently, I decided to research the matter using the approach outlined in Chapter 4.

## **1.6 Overview of the research**

This study centres on the relationship between workplace-based assessment and feedback and the development of competence in postgraduate radiology training. The picture painted by the literature in relation to workplace-based assessment is that it provides an opportunity for junior doctors to be observed by their senior colleagues and receive feedback on their practice. This formal approach to feedback clearly has the potential to be educationally beneficial, however the evidence to date is that this is often not the case. Clues that the WBA system may not be functioning as intended have arisen through my own work as an educationalist, both in my role as a course lecturer and facilitator as well as through initial exploratory work in WBA and feedback at the RCP. As will be demonstrated in more depth by the review of literature, the field currently suffers from a lack of empirical evidence, save for a small number of studies which have for the most part focused on data from pilot projects or small scale studies.

My research can therefore be summarised as seeking to answer the following central research question:

**Is the system of workplace-based assessment and written feedback in postgraduate clinical radiology training in the UK fit for purpose?**

### *1.6.1 Research objectives*

In considering how such a complex question may be addressed, it was clear that there should be several discrete axes of enquiry. Therefore, for the purposes of manageability and clarity, the overall question was broken down into the following



research objectives, each of which was aimed at supporting the consideration of an important aspect or aspects of the central question.

According to Stobart (2012), to ask the question as to whether or not an assessment is fit for purpose is to ask an important question about its validity. Consequently, the first objective of the research was:

- To identify the constituent components of validity in assessment, with a particular emphasis on formative assessment or 'assessment for learning', in order to construct a theoretical account of validity with which WBA in clinical radiology could be compared.

The review of literature that follows this chapter sets out a number of ideas regarding validity in assessment, and offers a backdrop for the analysis of the validity of WBA in radiology. As is argued in more depth in the same chapter, an intrinsic aspect of this validity question concerns the specific purpose for which the assessment has been devised, hence the second objective was:

- To analyse the claimed purposes of WBA, and the extent to which any multiple purposes exist and are in conflict.

This objective was addressed through the review of literature, and included the analysis of policy documents as well as research articles and medical education texts in order to examine the compatibility of various official, and other, assertions that have been made regarding the role of WBA. The GMC (2010) has acknowledged the likelihood for competing purposes to give rise to tensions that threaten the educational validity of the WBA process, and so any apparently divergent roles or purposes of WBA outcomes were analysed with respect to their potential to impact on the formative assessment and feedback process.

In seeking to establish the quality of the feedback that radiology assessors provided, an important element of the research was the analysis of empirical WBA feedback data. Given the lack of access to radiology trainees and assessors in order to objectively verify educational impact, it was necessary to construct a theory-driven framework which could be used to conduct content analysis of assessors' written feedback

comments and draw conclusions about their potential effectiveness. Therefore, a third composite objective was:

- To critically analyse the literature on formative assessment and feedback in order to identify the components of, or approaches to, feedback that have been found to be effective in supporting learning, and from this to synthesise and apply a coding framework for the content analysis of assessors' written feedback comments

Given the extent of the literature on learner engagement in the formative assessment process, through for example self-assessment, peer assessment, reflection and engagement in a dialogic feedback relationship, I was interested in whether and how trainees responded to the assessors' written comments. No existing analytical framework for coding trainee comments was found to exist in the review of literature, and so it was necessary to construct a coding framework inductively. The fourth objective was therefore:

- To analyse trainee comments accordingly to an inductively-generated coding framework, in order to draw conclusions about the extent to which the written feedback process might be said to evidence cardinal features of learner engagement in formative assessment, including reflection, self-assessment and dialogic interaction between the assessor and the learner.

Having completed content analysis of the feedback data, I was interested in elucidating what, if any, factors appeared to have an influence on the provision of high quality feedback to trainees. Accordingly, the fifth objective was:

- To establish what conditions, if any, appear to influence the provision of the highest quality feedback to trainees in clinical radiology.

A number of conditions were selected for statistical analysis, including whether the seniority of the trainee, their stage of training, or their performance in a given assessment were linked to the provision of high quality feedback.

### *1.6.2 Research questions*

The sub-questions of the main research question were therefore:

- What are the claimed purposes of workplace-based assessment in clinical radiology training, and to what extent do any multiple purposes appear to be in conflict?
- What are the documented features of the system of WBA and feedback in postgraduate clinical radiology training, and how do they compare with what is already known about effective formative assessment?
- What are the qualitative characteristics of the written feedback provided by assessors to clinical radiology trainees in workplace-based assessments, and how do these compare with the features of effective feedback found in the literature?
- What, if any, conditions appear to govern the provision of effective feedback in workplace-based assessments in clinical radiology?
- Can assessors in clinical radiology deliver feedback of sufficient quality to support the development of these trainee doctors?
- What is the nature of clinical radiology trainees' written comments, and is the written feedback process dialogical?
- Can workplace-based assessment and feedback in postgraduate clinical radiology training be said to support the learning of trainee doctors?

### *1.6.3 Design of the research*

In order to answer the research questions, the review of literature that underpins the study performs two main functions. Firstly, it fulfils the conventional role of establishing what is already known in the field of formative assessment (including workplace-based

assessment) and feedback, with a particular emphasis on the link between assessment, feedback and learning. Secondly, having established what is already known about the features of formative assessment and feedback that have been found to be effective in supporting learning, these features were synthesized in order to construct a theory-driven coding framework. The framework will subsequently be used to conduct content analysis of the assessors' formal feedback found in the national e-portfolio record for UK radiology trainees, and was modified inductively to ensure adequate representation of all relevant aspects of the written feedback comments contained in the e-portfolio.

Further analysis of the coded feedback data was undertaken in order to address the question of what, if any, conditions appear to determine the provision of the highest quality feedback, as adjudged against a theoretical construct of feedback quality. For example, are trainees who receive low assessment scores, or who are at earlier stage in their training, observed to receive high quality feedback more frequently than those trainees who are scored highly by the assessor, or who are close to the completion of training? Or is the provision of high quality feedback less systematic and predictable? The empirical dimension encompasses an analysis of trainees' comments, and also included descriptive statistical analysis of several aspects of the assessment system, such as the timing of assessments and the numbers conducted by trainees and assessors. Whether WBA and feedback can be said to support trainees' competence development was considered by comparing what was found in the empirical part of this study with what was revealed through the review of literature, in order to draw conclusions about the utility of WBA in clinical radiology. These conclusions then informed recommendations as to how the RCR might improve the effectiveness of WBA for radiologists.

## **1.7 Outline of the thesis**

The thesis consists of eight chapters. Chapter 1 provides the introduction and rationale for the study, as well as establishing the context of the work. Chapter 2 is the first of two literature review chapters, the focus of which is the pursuit of validity in assessment, including the assessment of doctors in training. Within the chapter I have drawn on the work of leading authorities in formative assessment, including Sadler (1989), Messick (1996), Newton (2007), Stobart (2012) and Black & William (2012) in order to critically analyse important principles of validity in assessment, including face validity, validity-

as-measurement-accuracy, washback and educational impact. This chapter therefore forms the 'validity backdrop' against which judgements are made, later in this work, regarding the fitness for purpose of formative assessment in clinical radiology.

In Chapter 3 I have expanded on one particular element of formative assessment – the provision of feedback – in order to establish what is currently known about the approaches to feedback that are effective in supporting learning. This chapter also includes a critical evaluation of a number of analytical frameworks found in the literature for appraising the quality of feedback provided by teachers or assessors in a range of educational settings, and underpins my synthesis of a deductive, theory-driven framework for use in the empirical aspect of my research. Chapter 4 contains a discussion of the factors that impacted on the choice of particular methods for conducting the research, and a rationale for the research design decisions that were made in this regard. In particular, there is consideration of an approach to qualitative data analysis first proposed by Ragin (1987, 2000, 2008) . Details are provided as to how this approach was modified and applied to the data in this study in order to reveal whether particular types of documented feedback could be said to be provided systematically.

Chapter 5 contains a presentation of three main categories of results: the descriptive analysis of aspects of the documented WBA process; content analysis of the written comments provided by assessors and trainees; and traditional statistical analysis, and modified Ragin analysis, of assessors' comments. Chapter 6 presents a discussion, which draws together the strands of the theoretical and empirical aspects of the work in order to answer the main research question. Chapter 7 considers the broader implications of the findings of my research, and presents my recommendations regarding the next steps for the RCR in developing their approach to formative assessment in clinical radiology training. The final chapter is a reflection on the limitations of the study and on my own development as a researcher over the course of this work.

## CHAPTER 2

### 2. Literature review Part I – Assessment in medical education

*Validity is central to any assessment. It is about the purpose of assessment, whether the form of the assessment is fit for purpose, and whether it achieves its purpose.* Stobart, 2012, p. 233.

#### 2.1 Introduction

The focus of this thesis, namely the fitness for purpose of workplace-based assessment (WBA) in clinical radiology training, was not derived from a speculative review of literature, yielding a 'gap' in the evidence to be duly explored. Rather, the starting point was my own awareness, as an educationalist delivering training on assessment and feedback for doctors, of the range of clinicians' views on the role and conduct of WBA and feedback in postgraduate medical education. Therefore, an important function of this literature review was to gather together the current evidence in relation to workplace-based assessment and feedback, in order to inform and shape the central research question. An important aspect of this exercise involved consulting official documents in order to identify the explicitly-stated purposes of WBA. However, it was important throughout the review process to remain alert to the existence of implicit or assumed purposes of WBA, as these may be as likely as any official pronouncement of purpose to influence how the assessments might be used in practice. Furthermore, a number of the stated or implied purposes may be misaligned, divergent or even in conflict, and therefore likely to affect the validity of the overall assessment process.

The concept of validity is key to establishing the fitness for purpose of the WBA system – authors such as Stobart (2012) argue that validity in assessment is actually *about* whether or not an assessment can be demonstrated to be fit for the purpose or purposes to which it is put. Therefore, a considerable portion of this chapter is given

over to exploring validity matters in assessment, and the challenges to WBA validity that exist in medical education as a particular type of professional education.

In building the validity argument, it was helpful to consider some often-used terms and ideas related to assessment, such as 'summative', 'formative', 'assessment of learning' and 'assessment for learning'. This was not done in order simply to go through the academic motions of addressing fundamental concepts in assessment. Rather, it has been included because the bodies that have driven forward changes in the assessment of doctors in the workplace have appropriated these terms and ideas, and have used them to explain their introduction of novel approaches to assessment into medical education. They have also adopted these terms in their guidance to medical educators – primarily doctors – as to the concepts behind certain approaches to assessment and how the assessments are therefore to be used. Consequently, I have drawn heavily on the literature that contains the most fully developed exposition and critique of these assessment-related concepts – the literature from school-based and higher education settings. This was deliberate, despite the nature of postgraduate medical education being professional rather than school- or classroom-based, since the GMC and RCR have adopted approaches to assessment that have arisen through research and application in the classroom-based context. In this chapter I will consider how the professional learning context of postgraduate radiology training squares with educational concepts adopted from school-based and higher education. This is another important element of the fitness for purpose or validity argument, and one without which something would be lacking from the analysis of WBA validity.

Having considered the validity issues in connection to the professional dimension of radiology training, another function of this chapter was to critically evaluate the *educational* context of WBA. Workplace-based assessment in postgraduate radiology education in the UK is situated within a curricular framework that is generally described as 'competency-based'. Therefore, in order to fully explore the notion of fitness for purpose, it was important to locate WBA within current concepts of, and approaches to, competency-based education in medicine. In doing so, it was necessary to critically analyse the academic discourse on competency-based education that exists within medical education, and explore the extent to which the approach aligns with contemporary thinking in practice-based or vocational education. Thus, I have aimed to shed light on the combination of theoretical and practical challenges that currently

beset workplace-based assessment, and draw conclusions about the extent to which WBA aligns with prevailing concepts of effective formative assessment.

The chapter that follows this one contains a more in-depth look at the evidence in relation to one specific facet of formative assessment practice – the provision of feedback – in order to synthesise a theoretical model of effective feedback. This theoretical framework was then used in the results and discussion chapters to provide a backdrop against which to judge the findings of the qualitative analysis of assessors' written feedback in WBA.

## **2.2 Assessment**

While there are many views in the literature regarding the role of assessment, the purposes to which assessment outcomes should be put, and how assessment should be done, there is generally good agreement about what assessment *is* in an educational context. Harlen (2012) sums it up thus:

It is generally agreed that assessment in the context of education involves deciding, collecting and making judgements about evidence relating to the goals of learning being assessed (Harlen, 2012, p. 87).

That is not to say that the assessment endeavour itself is necessarily easy or straightforward. There are likely to be different views as to what evidence should be collected, how it should be collected, and what judgements may or may not be supported by the evidence. The learning goals themselves may be unclear, either to the teachers, or to the learners, or to both. Furthermore, the measurement itself is likely to prove problematic. Gardner (2012, p. 115) quotes Dressel's (1983, p. 23) tongue-in-cheek characterisation of assessment outcomes, which Dressel describes as, 'an inadequate report of an inaccurate judgement by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite material'. Whilst this is likely to be something of an overstatement, it is sobering to consider the extent to which aspects of Dressel's portrayal could be said to be true.



Even taking assessment as 'measurement of learning', then, it is arguably the case that assessment presents significant challenges. However, a further challenge concerns how to make use of assessment judgments to inform teaching and learning. In other words, even if the primary role of an assessment is to 'summarize learning', a phrase used by Harlen (2012, p. 87), the assessment should be more than simply an adjunct to the educational effort - or as Gardner (2012) aphoristically terms it, 'the assessment tail' attached to 'the curriculum dog' (p. 104). On the contrary, as Gardner (2012) goes on to argue, the weight of evidence from assessment research over the last two decades is that assessment is a vital element of teaching and learning:

If we have learned anything from the last 20 years...[of assessment research]...it should be that assessment must be recognized as an integral part of the learning process (Gardner, 2012, p. 104).

However, embedding assessment within teaching and learning does not necessarily come naturally, even to dedicated classroom practitioners. As described by Pedder and James (2012), specific professional development is required in order for even classroom-based education practitioners to develop, and use successfully, approaches to assessment that support students' learning. It is likely that educators in less explicitly educational settings, such as hospitals, have an even greater need for professional development in this regard, given their more distal connection to leading edge educational research. For example, David Black, RCP Vice President for Education and Training (RCP, 2012) has expressed a view of assessment that is in stark contrast to that of Gardner (2012), and has proposed a separation, rather than integration, of assessment and learning:

An assessment should be summative, rather than developmental and formative. The GMC and colleges are now working to separate these activities in order to be clear about what is an assessment...and about what is likely to be called a supervised learning event (SLE), (RCP, 2012, p. 13).

Such is the remoteness of educators in other settings from the world of school- and higher education-based research, it is possible for them to hold views or concepts of assessment that appear significantly outdated, or at least restricted, to educationalists from more explicitly educational settings. However, as Boud (2000) neatly observes, 'as members of...a profession...we follow the norms of practice with which we are familiar' (p. 160). Medical doctors are typically familiar with assessment that is intended

to summarize learning at a point in time, normally to inform high stakes decisions about certification and progression. Accordingly, Black's (*ibid.*) statement most likely resonates with many members of his profession, including those who are currently in training, and provides a clue as to the cultural challenges and potential threats to systemic validity that might beset formative assessment in medical education settings.

None of this is to say that educators and educational researchers in traditional educational settings are in uniform agreement about what terms such as 'summative assessment' and 'formative assessment' should be taken to mean, or how either should be practised. Consequently, a consideration of terms such as 'summative assessment' and 'formative assessment' is useful prior to considering the issues in establishing fitness for purpose of any assessment – in this case WBA in medicine. This is particularly the case in this study as the WBA guidance from the GMC (GMC, 2010) and the RCR's clinical radiology curriculum (RCR, 2010) underscore the formative nature of WBAs, and yet insist equally emphatically that the amalgamation of WBA outcomes can be used to support annual summative decisions about progression through training.

### *2.2.1 Summative assessment*

To describe an assessment as being summative may be to attempt to convey a number of different ideas about the assessment. What the term summative is often taken to mean is that the assessment is intended to 'measure', or in some way gauge or capture, the performance or capability of the learner at a point in time. Harlen (2012) describes this as assessment which summarizes or reports learning. How the measurement is conducted may vary widely, depending on the nature of what is to be judged, but regardless of how the measurement is done, it tends to be made against some sort of standard or set of criteria. Harlen (2007) suggests that summative assessment tends to 'judge achievement against broader indicators, such as level descriptors or grade level criteria' (p.139). Of course, the standard may not always be as objective – Newton (2007) draws attention to norm-referenced summative assessment, which is concerned with comparing learners with their peers rather than a standard, and so returns a different type of measurement than a criterion-referenced summative assessment approach.

The term 'summative' may also be taken to denote something about the timing of the assessment. Bloom *et al.* (1971), often credited with introducing the term 'summative' into educational assessment, explain that,

We have chosen the term 'summative evaluation' to indicate the type of evaluation used at the end of a term, course, or program for purposes of grading, certification, evaluation of progress, or research on the effectiveness of a curriculum, course of study, or educational plan (p. 117).

The term summative may also imply that there is an element of finality about the assessment judgement. According to Sadler (1989):

Summative contrasts with formative assessment in that it is concerned with summing up or summarizing the achievement status of a student, and is geared towards reporting at the end of a course of study especially for purposes of certification. It is essentially passive and does not normally have immediate impact on learning, although it often influences decisions which may have profound educational and personal consequences for the student. The primary distinction between formative and summative assessment relates to purpose and effect, not to timing (p. 120).

What this tends to mean, in practice, is that summative assessment is often rendered a 'high-stakes' activity due to the limited opportunities to repeat the assessment and, crucially, the uses to which the assessment judgments are put e.g certification. Even if opportunities to repeat the assessment exist, learners may do so relatively blind to the aspects of their performance which were previously judged to have fallen short – these so-called 'passive' assessments (*ibid.*) are not usually geared towards providing information to candidates to support their future learning.

Sadler hints at an effect of high stakes assessment that may, in practice, be anything but passive – assessment decisions which have 'profound educational and personal consequences' for the student are unlikely to be perceived as passive, or necessarily benign. There might be negative effects on the learners, and the educational system. Consequently, summative assessment has at times been given a bad name, largely due to the effect that high-stakes testing can have on teaching and learning practices – so called 'negative washback' effects (see Messick, 1996, for example). Washback may be considered to affect the validity of an assessment. Similarly, impact – the effect that the assessment has on any aspects of the education system, or even the society within which it operates (Wall, 1997) – may also be considered a component of assessment validity, and both are considered in the context of WBA, below.

### 2.2.2 Formative assessment

The declarations of the GMC (2010) and the RCR (2010) that WBAs are primarily formative assessments means that the term merits some consideration. The GMC in particular uses the phrase *formative assessment* interchangeably with *assessment for learning* (AfL) (e.g. GMC, 2010, p. 8). As Gardner (2012a) observes, this is not uncommon, and he takes the pragmatic view that, 'in the final analysis there is little of substance to distinguish the two terms' (p. 3). However, the two terms have different origins and, at times, do refer to different types of assessment practice. Historically, *formative assessment*, the term preferred by the RCR (2010), pre-dates *assessment for learning* and as Gardner (2012a) explains,

[formative assessment] is sometimes used to describe a process in which frequent *ad hoc* assessments, in the classroom or in formal assessment contexts such as practical skills work, are carried out over time and collated specifically to provide a final (summative) assessment of learning (Gardner, 2012a, p. 2).

As it happens, this is a very apt description of WBA. One of the prescribed uses of WBAs is that they should be amalgamated in order to inform annual summative judgements on a trainee's readiness to progress to the following year of training: 'A series of WPBAs inform assessments of learning, which are essential waypoints for the judgement on progress throughout training' (GMC, 2010, p. 2). Yet this is not what the GMC or the RCR chiefly intend formative assessment to mean. In fact, their idea is more closely aligned with conventional notions of AfL. In their WBA guidance document, the GMC state that 'the primary purpose of WPBAs is to provide feedback – *assessment for learning for the trainee*' (GMC, 2010, p. 5, my italics). Although they limit their description of formative assessment practice here to the provision of feedback, the idea is in keeping with the commonly understood sense of assessment for learning. For example, the Assessment Reform Group (ARG, 2002) define assessment for learning as 'the process of seeking and interpreting evidence for use by learners and their teachers, to identify where the learners are in their learning, where they need to go and how best to get there'. Clearly there is more implied by the ARG's definition than simply the provision of feedback, but the GMC's concept of formative assessment, or assessment for learning, is arguably aligned with this emphasis on supporting learners' development.

### 2.3 Validity in assessment

In its simplest form, assessment validity may be taken to mean the extent to which an assessment measures what it sets out to measure, a concept often referred to by authors in the educational literature as alignment. As Case *et al.* (2004) put it: 'The alignment between an assessment and a set of content standards in a subject area has long been recognized as evidence of the assessment's validity'. Gardner (2012) provides some antithetical examples of alignment, such as science students being asked to correctly identify the features of a fair test in a multiple choice question when the intended learning was that the student be able *to conduct* a fair test. The principle of ensuring that an assessment is congruent with intended learning outcomes may appear to be a somewhat rudimentary one. Certainly, the later sections of this chapter will consider more complex aspects of assessment validity. However, the basic principle of alignment has not historically been well observed in the assessment of doctors in training. Similarly, reliability, a necessary (though insufficient) component of validity in any measurement system, has not been a particularly prominent feature of assessment in medical education until recent times. The examinations of the medical royal colleges are a case in point.

For the last 500 years, formal assessment of professional capability has been conducted via the knowledge- and practice-based examinations of the medical royal colleges. In the case of the Royal College of Physicians of London – the oldest of the medical royal colleges, instituted by Henry VIII in 1518 (RCP, 2015) – the examination took a format that was intended to confirm the successful candidates as capable clinicians, as well as 'men of wide learning' (RCP, 2015).

Conducted in Latin until the nineteenth century, the format of the exam in the earliest days is described thus:

Candidates would have three months to build their knowledge of the 17 volumes of recommended reading. The president, together with five Fellows, would pick out at random three questions from three different places in the books. The candidate would be allowed several hours to consider the questions, before returning to read out to the assembled college his identification of the passages and his answers to the questions (RCP, 2015).

Although greatly changed throughout the years between the sixteenth and the early twentieth centuries, the college examinations in the early 1900s suffered from what Van der Vleuten (1996) describes as 'subjectivity and poor measurement characteristics' (p. 42). This, along with ever-increasing numbers of students wishing to qualify as medical practitioners, meant that reform of the assessment process became necessary.

While the Royal College of Radiologists (RCR) is a much 'younger' medical royal college than the RCP – the RCR only received its royal charter in 1975 (RCR, 2016) – candidates have been sitting professional examinations in radiology since the 1930s, and the Fellowship of the Royal College of Radiologists (FRCR) examination since 1975. Changes in imaging technology in parallel with evolving medical knowledge have required an attendant evolution of the FRCR examination. As well as keeping pace with developments in science and medicine, college examinations have been modernised in an effort to match developments in educational practice. Consequently, modern day royal college examinations are generally held to be much more reliable than the college examinations of previous eras. For example, the first examination in the three-part sequence for Membership of the Royal College of Physicians (MRCP), known as MRCP Part I, was found to have been reliable over an extended period from 1984 to 2001 (McManus *et al.*, 2003). However, even at the turn of the 21<sup>st</sup> century, the overarching question of alignment remained. Salter and Smith's (1998) study of 59 trainees who had successfully passed the MRCP Part I exam revealed that 69% of the study participants believed clinical experience to be 'irrelevant to achieving success' (p. 34) in this portion of the exam. It intentionally and explicitly focuses on theoretical knowledge. The trainees' observations, however, are in stark contrast to the official rhetoric about the assessment:

Although the examination is a theoretical one this does not mean that clinical experience is not a very important contributor to the body of knowledge which will enable [the candidates] to achieve success...In general therefore, training for the examination demands wide reading *and extensive clinical experience* (RCPEGL, 1996, in Salter and Smith, 1998, emphasis mine).

Contrary to this assertion, all trainees who participated in the study identified that, rather than gaining extensive clinical experience, 'practice with former or simulated papers was...by far the most effective method of achieving success' (Salter and Smith,

1998, p. 34). In keeping with the theoretical focus of the exam, reading was also identified as being important by 63% of trainees. This was expressed with the caveat that it should be test-orientated reading (such as reading up on why practice questions had been answered correctly or incorrectly) as opposed to reading 'textbooks and leading journals with an emphasis on review articles and editorials' as recommended by the RCP (*ibid.*). The finding that the test failed to assess a prescribed body of theoretical knowledge would therefore give rise to questions about its alignment even as a knowledge-based examination, never mind as a test of clinical experience or capability.

### 2.3.1 Components of validity

It is worth taking a moment to consider the commonly-encountered aspects of validity in the assessment literature. Principally, four types of validity are frequently discussed, and are summarized helpfully by Newton and Shaw (2014). *Content validity* is typically thought of as the extent to which the content of an assessment samples the criterion. In education, this is usually taken to be the body of knowledge prescribed by a curriculum. An assessment's validity is therefore judged 'in terms of alignment between the content of the curriculum and content of the test' (Newton, 2012, p.265). *Predictive validity* is, as the name implies, the extent to which an assessment can forecast a future attribute. This is often challenging to establish, and at best the test is an indication or sign, rather than a sample, of its criterion. An example from school-based education might be the extent to which performance in assessments at age 14 predicts performance in public exams at age 16 or 18. *Concurrent validity* is said to exist when one assessment outcome corresponds to another measure to which it can reasonably be said to be conceptually related. Given that the two measures are unlikely to overlap completely (in that instance they would be for all intents and purposes the same test) the assessment is again taken as a sign, rather than a sample, of its criterion.

*Construct validity*, unlike the previous three aspects of validity, is the extent to which an assessment reflects a property of the learner which is a theoretical abstraction rather than a directly observable feature. This might be 'their understanding of a certain set of concepts or their attitude toward something' (Wilson, 2005, p. 6). Given the underlying, latent nature of the construct, this type of validity can only be established indirectly. As

described by Newton and Shaw (2014), both empirical and logical/theoretical demonstrations of its existence are required. This is not to say it is subordinate to the other types of validity that have been mentioned, with their more objectively verifiable aspects. Rather, as proposed by Loevinger (1957), construct validity is often taken to subsume the other types of validity, as it concerns 'constructs or explanations, rather than methodological factors' (Cohen, Mannion & Morrison, 2011, p. 213). In other words, construct validity is concerned with what is being tested, at a fundamental psychological level, rather than how it is being tested. That said, certain assessment methods may threaten construct validity, and it is these threats to validity that are of particular interest in my study. One threat comes from an approach to assessment that fails to represent important aspects of the construct – so-called construct under-representation (Cohen, Mannion & Morrison, 2007). Another threat to construct validity, known as construct-irrelevant variation (*ibid.*), typically occurs when elements are introduced into the assessment that are unrelated to the construct which is the intended focus of the test. In both cases, the inferences that might be made on the basis of the test results (termed consequential validity by Cohen, Mannion & Morrison, 2007) may be affected. However, it is the knock-on effect of these threats to construct validity on teaching and learning which are of primary interest in my study. That is to say, I am less concerned with the construct validity of WBA *per se*, and more interested in the extent to which the overall approach to WBA supports or undermines effective formative assessment.

My interest is therefore better described by the concept of *systemic validity*, first introduced by Frederiksen and Collins (1989). They describe systemically valid tests as 'ones that induce curricular and instructional changes in education systems (and learning strategy changes in students) that foster the development of the cognitive traits that the tests are designed to measure' (p. 27). In particular, I am interested in the extent to which the actual feedback data evidences assessors' and trainees' understanding of formative assessment, and their own role as teachers and learners in a formative setting.

### 2.3.2 *Validity as a function of assessment use*

Each of the approaches to validity described above tends to cast validity as an objective property of the test itself. However, as described by Stobart (2012), it is now



widely accepted that validity is not a property of tests, but of the uses to which the tests are put. Messick (1996) puts it thus:

Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment (*ibid.*, p. 6, original emphasis).

He goes on to clarify that:

Validity is not a property of the test or assessment as such, but rather the meaning of the test scores. Hence, what is to be validated is not the test or observation device *per se* but rather the inferences derived from test scores or other indicators (*ibid.*, p. 6).

Stobart (2008, p. 104) explains further that, ‘a well-constructed test can still be invalid if the results are misinterpreted or misused. If I use a test for an unintended purpose or I misunderstand the scores, then test validity is compromised’. Stobart’s statement suggests that it is the use of test outcomes for unplanned purposes that compromises validity. However, even *intentional* uses of assessment, particularly when those uses are multiple and at times conflicting, may compromise validity. This concern with whether the downstream uses of assessment outcomes can be said to be valid – so called consequential validity (Cohen *et al.*, 2007) – throws a spotlight on the issue of assessment purposes in general, and the uses<sup>2</sup> of WBA in clinical radiology in particular.

### 2.3.3 Uses and purposes of assessment

For some authors, such as Harlen (2012), the use or purpose of an assessment is sufficiently implied by the allocation of the term ‘summative’ or ‘formative’. Harlen categorizes formative uses as those which are intended ‘primarily to help students’

---

<sup>2</sup> At times in the literature the term ‘purpose’ has been used to describe the intended use of an assessment whereas ‘use’ has been taken to mean what has actually been done with these outcomes (e.g. by Harlen, 2007, and Mansell *et al.*, 2009). However, in talking about the uses, purposes or functions of assessment, I am following Newton’s (2007) example of using these terms interchangeably. The context should make clear whether I am talking about intended or actual use.

learning' (*ibid.*, p. 88) and summative uses as those that are intended 'primarily for reporting on students' achievement' (*ibid.*, p. 89). In doing so, she echoes Black and William's (2003) observation that 'from their earliest use it was clear that the terms 'formative' and 'summative' applied not to the assessments themselves, but to the functions they served' (p. 623).

Other authors, such as Newton (2007), have taken issue with the use of terms such as formative and summative as descriptions of assessment purposes. For Newton, the term summative refers to a particular category of assessment purpose – what he terms the first level, or 'judgement level' of assessment purpose (p. 150). Judgement level assessment discourse should be concerned, he argues, with the technical aim of the assessment e.g. to make a criterion-referenced judgement about learners. It says nothing, by itself, about how that judgement is then to be used. Newton's second level of purpose, is the 'decision level' (*ibid.*, p. 150). This is characterized by discourse about the 'decision, action or process' that may be supported by the assessment judgement e.g. the provision of feedback to learners, or selection decisions regarding entry to certain courses of study. For many authors, the former example (provision of feedback) would be regarded as a *formative purpose*, whereas the latter (selection) would be regarded as a *summative purpose*. For Newton, on the other hand, to talk about a 'summative purpose' is to make a category error – he argues that the term 'purpose' should be reserved for the second tier of his taxonomy. Therefore, according to Newton, all assessments have a judgement (i.e. summative) dimension, in that they all require some sort of judgement to be made about the learner's performance. Some assessments then have a formative purpose – the use of the assessment judgement to in some way inform next steps in learning.

Newton's point may appear a little academic, and indeed authors on assessment who are aware of his thesis (e.g. Harlen, 2012) nonetheless tend to revert to the language of 'formative purposes' and 'summative purposes,' believing the notions they capture to be generally understood. However, Newton makes the compelling point that these general understandings may not be shared. To wit, William (2004)

In the United States, the term 'formative assessment' is often used to describe assessments that are used to provide information on the likely performance of students on state-mandated tests – a usage that might better be described as 'early warning summative' (William, 2004, p. 4).

Similar approaches to so-called formative assessment have been encountered in medical education. In a recent pilot of proposed changes to the WBA system for trainee physicians, the three royal colleges of physicians in the UK recommended that trainees undertake several formative WBAs (known as supervised learning events or SLEs) before undertaking a single summative (i.e. pass/fail) WBA, known as an Assessment of Performance (AoP) (British Association of Dermatologists, 2015). Whilst the WBAs in radiology are not intended to head-off poor performance in pass/fail tests, the use of WBAs to detect underperformance might be said to be equivalent to the 'early warning' idea invoked by Wiliam (2004). If handled sensitively, and translated into an action plan, this could even epitomize the formative use of WBAs. However, a competitive culture and a desire to avoid being labelled as 'failing', both previously discussed, could limit this aspect of the WBA role.

More fundamentally, Newton (2007) questions the very notion of multiple assessment purposes. In particular, he questions 'the extent to which evidence elicited for any particular purpose can legitimately be used for any other' (p. 160). For example, WBA outcomes that are primarily supposed to inform learning, which are then used to inform a judgement about progression to the next stage of training. Newton argues that the multiple use of a single assessment outcome is particularly likely to occur when an assessment yields a summative output. For example, when assessments that are to be used to inform progression (a criterion-referenced process) are then used by selection panels in job interviews (normally a competitive, norm-referenced process). Yet, there is evidence that this kind of 'mission creep' occurs. Davies *et al.*'s (2009) evaluation of the UK foundation WBA programme arguably went beyond its original evaluative remit to recommend that 'collated [workplace-based] assessment data should form part of the evidence considered for *selection and career progression decisions*' (p. 74, my emphasis).

Newton recognises the use of formative assessments for summative 'purposes' as a sub-set of this role confusion. This is of particular interest to me, given that the use of 'formative' WBAs to inform 'summative' decisions about progression is not mission creep but is one of the originally intended purposes of WBA.

#### 2.3.4 *Formative assessment for summative purposes*

Harlen & James (1997) consider the dual use of assessments for what they call formative and summative purposes and conclude that, as long as caution is exercised, assessment intended for formative purposes *can* be used for summative purposes. This view would appear to be anathema to Newton (2007), who argues that even apparently similar uses of assessment, such as the short, medium and long term monitoring of an educational system, may not be supported by the same type of assessment. Harlen & James (1997) contend to the contrary that summative purposes can be served by evidence collected during teaching as long as the evidence is re-interpreted in light of the summative decision that is being made. In particular, they emphasise that a simple summation of judgements made in formative assessments is inappropriate, given the variable nature of learners' performance in the course of different teaching and learning tasks. Instead, they propose developing teachers' appreciation of the fact that learner performance does vary over time and in different contexts. A more sophisticated understanding of real world student performance, they argue, plus a greater appreciation of holistic assessment criteria rather than reductionist, itemised criteria would allow teachers to make more valid summative judgements about learners' progress.

Harlen and James' (1997) approach to collation, rather than arithmetic summation, of evidence aligns with the approach to decision-making recommended by the GMC (2010) in relation to WBA. Avoiding an algebraic approach, they refer more broadly to using assessment evidence collectively, to 'form an overall profile of an individual' (GMC, 2010, p. 1). They do not provide any detail as to how this is to be done, with this information instead being provided by the royal colleges. In the case of clinical radiology the RCR have provided a template, known as a decision aid, for the purpose (see Appendix 3). Whilst the decision aid appears to be quite detailed regarding the extent of expected curriculum coverage at the end of each stage of training, it stops short of specifying particular assessment outcomes, or requiring evidence of specific curriculum competencies having been achieved. The problem of overly simplistic use of formative assessments to inform summative decisions, against which Harlen and James (1997) have cautioned, might therefore have been said to have been avoided. However, given the consequences of making an incorrect decision about progression in a doctor's career, the question of dependability remains.

Harlen (2012) recognizes the potential problem with accuracy when using teacher assessment for high stakes decisions, referring to 'the known bias and errors that occur in teachers' judgements' (p. 100). This often translates into a concern about the reliability of these judgements. Whilst in the school context this might be addressed through approaches like standardized tasks and intra- and inter-school moderation, this level of calibration would be difficult to achieve in a professional clinical setting. It may not even be desirable, as the set tasks may no longer represent the reality of clinical practice, with direct consequences for the construct validity of the assessments and, importantly, their usefulness for formative purposes. This final point may be developed further, in a manner not addressed by Harlen & James (1997), to consider what the impact on *learners* might be upon finding out that their 'formative' assessments are also intended to contribute to a high-stakes summative judgement.

## **2.4 Washback and systemic validity**

*In all educational and psychological testing, what matters are not the processes that are operative in task performance, exemplary though they may be, but the processes captured in test scoring and interpretation. If it occurs, washback is likely to be oriented towards the achievement of high test scores as opposed to the attainment of facile domain skills (Messick, 1996, p. 5).*

For many leading authorities on the subject of assessment validity in education, their interest extends beyond a concern purely with the accuracy and reliability of the test. Their concern is also with the effect that any particular assessment, or programme of assessments, has on learners, teachers and the larger educational context within which they function. This broad influence on the education system has been labelled by some (e.g. Frederiksen and Collins, 1989) as *systemic validity*, which they describe as the validity that accrues when assessments introduce 'curricular and instructional changes that foster the development of cognitive skills that the test is designed to measure,' (*ibid.*, p. 27).

Washback (also termed 'backwash') is a more narrowly defined aspect of systemic validity. Alderson and Wall (1993), in a paper that is now considered a classic, define washback as simply 'the influence of tests on teaching' (p. 115). Messick (1996, p. 1)

extends this definition to encompass learners as well as teachers: 'washback...is the extent to which the test influences...teachers and learners to do things they would not otherwise necessarily do.' In linking the impact of an assessment on teaching and learning, Messick (1996, p. 2) underscores the exclusivity of the relationship: 'washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test, and not of other forces operative on the educational scene.' However, he goes on to acknowledge later that while these 'other forces' may not *create* washback, washback nonetheless 'appears to depend on a number of important factors in the educational system in addition to the validity of the tests' (*ibid.*, p. 6). Simply put, the 'washback hypothesis' (Alderson and Wall, 1993) is that there are features of the assessments themselves and the system within which they operate that can generate an effect on teaching and learning. The exact nature of this effect, including who or what is affected and the valency of the effect, is likely to be complex and variable.

#### *2.4.1 Washback valency - positive and negative washback*

According to Bailey (1996, p. 268), 'washback can be either positive or negative to the extent that it either promotes or impedes the accomplishment of educational goals held by learners and/or programme personnel.' Thus, positive washback is said to occur when an assessment encourages learners to engage authentically in the activities that comprise the construct at the heart of the educational endeavour. It should therefore also encourage teachers to direct their instruction towards this end, rather than towards artifices and tactics required purely to achieve success in the assessment. As Weigle and Jensen (1997, p. 205) put it, if a test has positive washback 'there is no difference between teaching to the curriculum and teaching to the test'. Negative washback will occur when assessments fail to represent the breadth of content described by the curriculum. It also occurs when, in some other way, success in the assessment is believed to require something more than a good command of the underlying construct that the assessment purports to test. These effects are likely to be particularly noticeable when teachers and learners perceive the judgement aspect of the assessment to matter. For learners, this may take the form of assessments for the purpose of progression or certification. For teachers, the use of assessments for accountability purposes may introduce a similar motivation.

It should be the case, therefore, that WBAs, as formative assessments conducted with the primary aim of providing feedback in the course of a trainee's authentic clinical work, are well suited to the creation of positive washback. However, as described earlier, alongside the intended formative function of WBA the assessments have been endowed with a number of other functions that may impact on the perceived importance of the assessment, potentially creating negative washback in the process.

#### *2.4.2 Washback as a feature of high stakes assessment*

Implicit in the description of washback effects given above is the notion that the effects are likely to be all the more acute when the assessment is perceived to be high stakes. According to Buck (1988, p. 17), 'There is a natural tendency for both teachers and students to tailor their [teaching and learning] activities to the demands of the test, especially when the test is very important to the future of the students, and pass rates are used as a measure of teacher success.' It is intuitively the case that the seriousness of the consequences of assessment failure is likely to correspond to the impact of an assessment on learning and teaching. Thus, it is worth considering whether radiology trainees are likely to view WBA as high stakes or low stakes assessments. The GMC state that it should be the latter, however they do so by conflating formative assessment with 'low stakes' assessment:

Assessment for learning is primarily aimed at aiding learning through constructive feedback that identifies areas for development. Alternative terms are Formative or Low-stakes assessment (GMC, 2010).

In practice, though, there are at least two elements of the GMC's and the RCR's published purposes of WBA that raise the stakes for trainees.

The first of these elements is the collation of WBA outcomes to inform judgements made by ARCP panels regarding progression to the next year of training. The guidance about this decision process, issued by the RCR within the curriculum, is in the form of a document known as an ARCP decision aid (RCR 2010, p. 164, see Appendix 3). The document defines the annual WBA-related requirements as:

- 6 mini-IPX assessments (2 per clinical attachment)
- 6 Rad-DOPS assessments (2 per clinical attachment)
- 1 multi-source feedback
- 1 Audit assessment
- 2 Teaching observations

It also states that 'WpBA should be undertaken in a timely and educationally appropriate manner throughout the training year' (*ibid.*, p. 164) and that progression will be 'predicated by [*sic*] satisfactory anchor statements' (*ibid.*, p. 164). No further guidance is given on what would constitute timeliness, educational appropriateness, or satisfactory overall outcomes ('anchor statements').

It seems that successful progression relies on trainees recording the correct number and distribution of each WBA, as well as achieving satisfactory overall outcomes in each one. There is further guidance that implies anyone achieving these targets might, even then, be considered to be merely 'getting by' in their training: 'A minimum number of WBA is specified in order to progress. It is expected that most trainees will undergo many more assessments demonstrating their engagement with reflective learning in practice' (RCR, 2010, p. 10). In other words, it is important for trainees to be seen to be addressing the requirements of the curriculum with respect to numbers and distribution of assessments, and ideally to record large numbers of assessments in order to demonstrate engagement with the process. The result of washback on learners could therefore be the tactical accumulation of assessments to meet these instrumental ends, realising the very 'target driven 'tick-boxing' approach' that the GMC (2010, p. 3) are keen to avoid. Numbers of assessments are only part of the formula, with 'satisfactory anchor statements' being another requirement for progression. Here, a tactical, or as Entwistle (1987) would describe it a strategic, approach to assessment may be encouraged. The trainee-led ethos of the WBA system would afford trainees the opportunity to delay assessment until they believed themselves to be competent, or to approach assessors whom they believed (or knew) would be less stringent in their judgements. Again, the washback effect here could encourage a certain degree of game-playing in order to achieve satisfactory outcomes. It is also possible that assessors and trainees alike concern themselves more with the judgement aspect of WBA, and less with the formative function of the assessments, given their collective familiarity with the gatekeeping role of assessment that has traditionally been so



evident in medical education. The result in this case would be little more than lip service being paid to feedback and reflection. The GMC recognise this risk:

Trainees are by their nature competitive. They want to achieve high scores and may therefore be very likely to choose to be assessed only towards the end of the programme (GMC, 2010, p. 3).

Whilst the GMC identify the motivation for delayed assessment as lying intrinsically with trainees, any such reaction to assessment is unlikely to be ameliorated by a system that emphasises satisfactory outcomes as one of the main aspects of WBA for informing progression judgements. Furthermore, delayed assessment may even have the effect of increasing the stakes. Trainees who request assessments late in their clinical attachment have a more restricted time within which to respond to any suggestions for improvement, or to request follow-up assessments in order to demonstrate improvement. Hence, assessors and learners alike may find themselves under pressure to record positive WBA judgements.

In support of the negative washback effects posited above, it appears that the WBA game is one that is worth playing. The RCR are explicit about the consequences of failing to meet the WBA requirements: trainees in this situation may be subject to a remedial action plan (RCR, 2010) or may even be asked to leave the training programme with no immediate possibility of return (GMC, 2015). These would appear to be compelling reasons for trainees to ensure that the WBA 'picture' painted by their e-portfolio is that of an engaged, high-performing learner.

The second element of the GMC's guidance that could function to raise the stakes is linked to another purpose of the assessments, which is to identify trainees who are falling short of the required standard. These doctors are often labelled as being 'trainees in difficulty' – a phrase that the GMC itself repeats in the course of its WBA guidance (GMC, 2010, p. 3). In other words, to receive a negative assessment judgement in a WBA is not just to receive an outcome that is personally disappointing, or which may present the challenge of generating a suitable number of counterbalancing positively-rated WBAs in order to ensure progression. It may also be to risk being labelled as a struggling trainee, with all of the attention that the label would most likely attract. This aspect of the GMC's guidance frames WBA as performing an important patient safety function, and so the emphasis here is on WBA as a clinical

governance tool, rather than an educational tool, with any particularly serious findings being potentially reportable to the GMC as the professional regulator. Cast in this light, it seems that the main objective of trainees is likely to be clearing WBA hurdles rather than gaining developmental input. In fact, developmental feedback may be genuinely unwelcome, given the potential for it to be viewed as flagging problems with performance, rather than informing next steps in learning and development.

A final consideration in examining how trainees perceive WBA is what they learn about the process through experience. For all that the above is true with regard to the potential and actual stakes of the assessments, the reality of the situation for most trainees is that their assessment judgements are likely to be substantially inflated. In a manner not dissimilar to what Stobart (2012) calls the Lake Wobegone effect, an early finding regarding WBA outcomes in UK foundation training was that the great majority of trainees were being rated as *above expectation for stage of training* (Davies *et al.*, 2009). Hinting at a similar finding amongst UK physician trainees, Crossley *et al.* (2011) describe the assessment data from over 4000 WBAs in their study as showing a 'skewed normal distribution' (p. 568) whilst simultaneously commenting on the reluctance of assessors to use the lower end of the scale. For all the potentially negative consequences of receiving low overall scores in WBA, if experience tells trainees that they are likely to be scored at the upper end of the scale a somewhat counterintuitive situation arises: an assessment system exists that is *potentially* high stakes while being relatively straightforward to 'pass'. Thus, WBA is rendered potentially valueless either as an assessment of learning, or as an assessment for learning.

What the washback effects may be, and how their existence may be verified, are challenging questions to answer. It is a challenge that has remained since Alderson and Wall (1993) first threw down the gauntlet to empiricists to demonstrate its existence. However, despite its empirical elusiveness, washback is a concept that can nonetheless prove useful in theorising the strengths and limitations of any particular approach to assessment.

### 2.4.3 Working for washback – the importance of authenticity and directness

For Messick (1996), the empirical study of washback is likely to be too heavily confounded to be productive in most educational settings. For example, while an evaluation of the change in learner outcomes after introducing a particular test *may* indicate an improvement in knowledge or skills, Messick points out that ‘a poor test may be associated with positive effects and a good test with negative effects, because of other things that are done or not done in the educational system,’ (*ibid.*, p. 2). Consequently, he recommends that researchers and educators would be better advised to attend to the conditions that, theoretically, should generate positive washback, in particular, *authenticity* and *directness* (*ibid.*, p. 2). Maximising these, he argues, will create the conditions for positive washback to occur.

#### *Authenticity in assessment*

According to Messick, (1996) authentic assessments are those which ‘pose engaging and worthy tasks (usually involving multiple processes) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world.’ In particular, he emphasises that the assessment of the focal construct should include everything that is relevant to the construct (Messick, 1994). In the language of construct validity, the threat to validity that Messick is keen to avoid is *construct underrepresentation*, where the narrowness of an assessment task fails to adequately capture all of the requisite aspects of proficient performance of the real world task. In this respect, WBA in clinical radiology has a lot going for it: the assessments are conducted in the real clinical environment, as trainee doctors go about their normal work with patients. As far as authenticity is concerned, therefore, it is difficult to conceive of a more authentic assessment task, and so it could be argued that the potential for positive washback is high. That is not to say that the assessment events are perfect representations of practice. The clinical setting presents a range of uncontrolled variables which may impact on the validity of any particular assessment in terms of the extent to which they allow the assessor to observe the genuine capability of the trainee. These may include individual factors such as assessment anxiety and other aspects of motivation. They may also include clinical or organisational factors such as: a long list of clinical cases that are routine or unchallenging, or (conversely) the unexpected presence of complex patients within the clinic list, the presence of

experienced or inexperienced colleagues (radiographers, nurses etc.). However, with the exception of assessment anxiety, it could be argued that the other elements *are* features of the authentic clinical setting, and so contribute to, rather than compromise authenticity.

A comparison with another popular approach to the assessment of doctors' clinical ability is useful here. The objective structured clinical examination (OSCE) was introduced into medical education by Harden and colleagues in the mid 1970s (Harden *et al.*, 1975) as an approach to assessing the clinical competence of medical students and trainee doctors that was reliable and offered good curriculum coverage (or content validity). The OSCE format originally proposed by Harden and colleagues has evolved over time, but the principles have remained consistent. Essentially, candidates rotate through a series of 'stations' (typically 5 to 12 stations in all, but 20 or more stations are not uncommon – see Brannick *et al.*, 2011) and are examined on some aspect of their clinical capability in each station. Stations often include simulated patients (actors), real patients, or body part simulators, and candidates are asked to demonstrate clinical skills such as taking a medical history, breaking bad news, taking blood samples and so on. The time spent at each station is short – 10 minutes is typical (see Hodges, 2003) – and candidates are observed and scored by examiners according to a pre-determined mark sheet. A composite of candidates' marks is used to determine whether or not they pass the examination.

In validity terms, the OSCE format has been presented as offering the potential for good content validity when compared with the traditional 'long case' (Harden *et al.*, 1975), due to the breadth of different clinical skills and medical conditions that can be represented within the one assessment. Along with this broader scope of assessment, the number of different assessor views on the trainees' capability (even on a short exam, pairs of examiners operating at five stations produces 10 views of the candidate's capability) makes it possible to generate highly reliable assessment scores. Indeed, Brannick *et al.*'s (2011) review of 39 OSCE-related studies revealed that reliability co-efficients of >0.8 were achievable. However, considered in terms of authenticity, a different picture of OSCE validity emerges. The very short, largely decontextualized encounters with patients, and demonstration of 'bits' of clinical capability, are not particularly reflective of the authentic clinical environment, or authentic clinical performance. As Teoh and Bowden (2008) put it, 'Could we conceive

of a professional music student who is told that her final acceptability as a musician will depend on a series of assessments of scales and short pieces but never on a recital of a complete piece of music?’

The brevity of each station is also not reflective of the time that doctors have with patients in reality. Despite the heavy clinical workload of doctors generally, hospital doctors often have half an hour for a first meeting with a patient in the outpatient setting. This is the clinical setting which is most naturally akin to the OSCE exam station setting, and so the time pressure of the OSCE is not particularly reflective of that reality. The range of settings that can feasibly be replicated in the OSCE is also limited – the ward-based or emergency department setting is not easily ‘staged’ within most medical schools or exam centres. Also, the practical or procedural skills that can be observed are limited to those which can be conducted within the 10-minute window of the OSCE station. Furthermore, other aspects of the genuine clinical environment are usually absent, such as the presence of medical and non-medical colleagues and the equipment and other artefacts present in the real world medical setting. Factoring in other cues linked to the assessment – it is common for bells or buzzers to ring when it is time for candidates to rotate to the next station, for example, and there are usually two examiners looking on while writing on mark sheets – the approximation to the genuine clinical environment begins to look somewhat superficial.

The influence of the OSCE assessment format on the learning of genuine clinical practice has been emphasised by Teoh and Bowden (2008), who report a change in medical students’ behaviour according to the format of the assessment. In one university, where the long case was dropped and replaced by a written examination, final year medical students ‘stopped seeing patients and spent most of their time studying for the written assessments’ (p. 336) – an example of negative washback, even if the authors do not identify it as such. Similarly, Gormley *et al.* (2011) report the impact at another university where the long case was replaced by the OSCE. Students in their study reported attempting to predict the types of patients who could be present in the OSCE, and admitted to concentrating on those types of patient during their clinical placements. Referring (albeit unwittingly) to positive washback, Teoh and Bowden (2008) concluded their piece by stating that ‘If we expect students to become doctors who take a “whole person” view of their patients, seeing them as more than the

sum of their diseased organ systems, then we must push them [via assessment] to learn medicine in an integrated manner' (*ibid.*, p. 336).

The potential influence of the OSCE is taken up in more general, theoretical terms by Hodges (2003), who highlights the 'transformative function of OSCEs' in redefining what medical competence *is* (p. 252). In other words, rather than *reflecting* the authentic clinical environment and testing extant authentic clinical practice, OSCEs *construct* a new (quasi-)clinical context in order to facilitate assessment, and require the performance of a version of clinical practice which is palpably different from the real world practice of clinical medicine in order for students or trainees to be deemed proficient (see Hodges 2003 for further illustration of this point). In fact, such is the power of these assessments, Hodges argues, they have been used to attempt to bring about systemic change in received notions of medical professionalism and clinical competence (*ibid.*). Examples that he cites include the efforts of the medical regulator in Canada to change clinical practice nationwide through an emphasis on aspects of clinical performance such as 'communication skills, inter-professional interaction, cross-cultural competence, patient-centred interviewing and sexual history taking, to name a few' (*ibid.*, p. 252) within certification and licensure OSCE exams.

Comparing WBA with OSCE assessment, then, WBAs would appear to offer a much more authentic assessment of clinical capability. The assessments are not based on constructed tasks, with decontextualized clinical skills demonstrated within specialized 'stations' – they are intended to be based on real time observations of trainees in the course of their normal clinical work with real patients who are receiving genuine clinical care (*cf.* the simulated patients of the OSCE assessment). In terms of time pressure in WBA, trainees are subject only to the time pressure present in the normal clinical context, and not the frequent, artificial deadlines required for the smooth running of a multi-candidate, multi-station practical exam. Thus, the setting for the assessment itself could scarcely be more authentic, and so it could be argued that the potential for positive washback is considerable. That is not to say that WBA are perfectly authentic. They are still assessments, and so assessment anxiety may exist, for example. WBA performance may therefore not mirror real world performance, the former being potentially either more or less proficient than the latter dependent on the individual's perception of the assessment, their response to stress, their familiarity with the assessor and so on.

In terms of construct underrepresentation, however, it seems that WBA generally manages to avoid many of the pitfalls of other approaches to the assessment of clinical capability. Where another challenge to positive washback, and hence systemic validity, may exist is in the directness of the assessments.

#### *Directness in assessment*

Direct assessments, according to Messick (1996), are those which allow participants to respond as freely as they would in the absence of a structured approach to assessment, unconstrained by response formats or limited choices of possible 'answers'. In validity terms, if *authenticity* is concerned with *construct underrepresentation*, then *directness* is concerned with *construct-irrelevant variation*. In other words, directness is compromised when an assessment is too broad, not in regard to its coverage of curriculum content, but in introducing unnecessary artefacts of the assessment process which could derail or in some way affect the candidate's ability to express what they actually know, or can do. Stobart (2008) cites the example of a mathematics exam in which the language of the questions is too difficult for the students to read. The consequence may be a low score in mathematics caused by a deficit in reading skills. More generally, the conventions of testing often introduce artefacts of the testing process itself that must be negotiated by candidates. For example, a learner may have a good grasp of the subject matter being tested, and yet be unfamiliar with the question rubric with the result that they are unsure of how to respond. The common instruction to 'illustrate your answer' may not indicate to a knowledgeable candidate what they need to do to satisfy the demands of the question. Accordingly, teachers can prepare students for tests, such that the validity threat posed by indirectness diminishes. However, this preparation can tip over into a different form of validity threat, in which learners are trained to respond in ways that do not necessarily relate to their understanding of underlying concepts. For example, Gordon and Reese (1997) found that learners can be taught to answer correctly questions that were intended to test application, analysis and synthesis abilities, without being able to apply, analyse or synthesise.

The caveat in all of this is that absolute directness is never achievable – Messick (1996), drawing on the work of Guilford (1936), points out that all assessment involves an element of indirectness, relying on processes such as judgement, comparison and inference. An example from the clinical setting is Holmboe *et al.*'s (2004a) finding that experienced medical practitioners often miss important aspects of performance when in an assessor/observer role. This was more than just a failure to observe important aspects of performance – a number of assessors had made the same observations of poor practice as their peers, but failed to classify them as serious or in need of correction. This is a good illustration of how, in all probability, an assessment judgement is at best an interpretation of an imperfectly observed performance. However, given that this is likely to be true of any assessment, the particular 'directness' features of WBA are worthy of consideration.

Again, the comparison of WBA with the OSCE approach is helpful in illustrating the strengths and limitations of the former. As previously mentioned, the OSCE format is one that has been contrived for the particular purpose of generating judgements of candidates' clinical skills that are as reliable as possible, and with the greatest degree of content validity (i.e. curriculum coverage) manageable. However, doing so in a format that is feasible for large cohorts of learners and which returns the type of reliability statistics that are desired has introduced a number of exam-specific characteristics that are not encountered in the real-world clinical domain, and which present particular exam-related challenges to the candidates. Some of these have already been highlighted, such as the artificial time pressure, the simulated nature of the patients and the obvious presence of examiners. There are additional features of OSCE assessments which lend themselves to the development of the 'testwiseness' that Messick argues is an indication of construct-irrelevant variance. For example, the format of the exam means that candidates can often rule out certain clinical scenarios or types of patient. Many clinical scenarios require the intervention of a team rather than an individual, for example, and most examination centres lack the resources to simulate these scenarios in an OSCE setting. If the candidates are aware that real patients, rather than simulated patients, are to be used, then it is highly unlikely that the patients will be acutely unwell, limiting the exam content to conditions that are chronic and stable. In addition, candidates may be able to make educated guesses about what domains are to be assessed in upcoming stations, despite Harden *et al.*'s (1975) original assertion that the format would actually rule out such a 'cueing effect' (p. 447).



This is even drawn upon by authors of textbooks aimed at helping students pass OSCE assessments:

As a rule of thumb, OSCE stations involving osteoarthritis are more likely to have an orthopaedic surgical focus, particularly [osteoarthritis] of the hip and knee, whilst inflammatory arthritides are more likely to have a rheumatological focus...A 'test wise' student who is taking a history like the one above will be prepared for a surgical rather than a medical discussion, either at the end of the station or at the next station...With this in mind, in the remaining time, the student should concentrate on [a list of surgical aspects of the patient's condition]. Byrne *et al.*, (2007, e-book).

In other words, well-prepared candidates will be able to narrow down potential diagnoses and predict likely questions based on features of the OSCE station. In so doing, they may outperform equally knowledgeable, clinically capable candidates who lack the same level of assessment-related tactical nous. This potential elevation of test scores for reasons that are not directly linked to the intended underlying construct may encourage negative washback, such that the attention of teachers and learners is diverted towards tips and tricks for succeeding in the assessment, rather than the genuine clinical skills of diagnosis and treatment. This is not to say that the clinical context of WBA lacks cues as to the nature of a patient's illness, or the best treatment options, however these cues are only the ones that would naturally arise from the authentic clinical environment, and are not features of a 'staged' assessment. Rather, WBA is conducted in the course of genuine clinical encounters which are much more akin than the OSCE to the unrestricted, open-ended assessment tasks recommended by Messick (1996). Of course, there are other aspects of the WBA process, not least the features of the educational and professional systems within which they operate, which may have an impact on assessment and learning in this environment. These systemic factors are considered next.

## **2.5 The influence of the system**

As Messick (1996) argues, the assessment itself is only one factor in determining washback. The other influences relate to 'the properties of the educational system, especially of the instructional and assessment setting,' (*ibid.*, p. 5). A key feature of this setting in radiology education is that it is a professional, medical setting. This contrasts

markedly with the school- or college-based classroom context within much research on the utility of formative assessment has been conducted. Another key feature of the educational system is that the syllabus component of the curriculum is comprised of numerous statements regarding the knowledge, skills and behaviours that have been deemed by the RCR to comprise professional capability as a clinical radiologist. These so-called competency-based curricula are not without their critics in medical education, and education more broadly, and so the implications for formative assessment are considered below.

### *2.5.1 Radiology training as professional learning*

A number of the principles of assessment and assessment validity discussed in this chapter (and of feedback, discussed in the next chapter) have been derived from the broad educational literature, which encompasses organisations or settings where education is the main focus of professional activity (schools, colleges etc.). I believe this to be legitimate, given that the GMC and medical royal colleges, including the RCR, appear to have adopted popular concepts from school-based education, such as formative assessment or assessment for learning, and introduced them into the postgraduate radiological setting. However, in doing so it must be acknowledged that there are important differences between overtly educational settings – where the primary function of the organization is to teach and support the learning of students – and settings where the primary function is something other than education (e.g. the treatment of patients). In the latter case, the approach to learning is more commonly conceived of as ‘professional learning’, rather than school- or classroom-based learning. Due account should therefore be taken of the particular contextual factors in professional learning generally, and clinical settings specifically, on which assessment practice may be contingent.

Yorke (2005) highlights some of the typical differences between practice settings and traditional educational settings: practice settings tend to lack well-rehearsed procedures for assessing learning; they often lack curricula or other standards against which learning can be assessed; and practice settings tend to be concerned with *performance*, rather than learning – failure to *perform* will often attract a great deal of organisational attention, whereas failure to *learn* (as long as performance is satisfactory) is likely to be viewed more leniently. In fact, according to Fenwick (2014)

learning may essentially be ignored if it is not necessary for, or clearly associated with, improved performance according to particular organisational measures such as productivity.

Yorke's (2005) description is a general one, and does not map perfectly onto the practice-based context of clinical education – for example, there *are* curricula in existence for every level of postgraduate medical training in the UK, the latter stages of which are separated into some 65 distinct specialties, one of which is clinical radiology. In addition, the curricula are broken down into numerous statements aiming to describe the knowledge, skills and behaviours required at each stage of training. Guidance in the form of the ARCP decision aid is also provided for educational supervisors as to the standard required at each stage. Furthermore, for trainee doctors, not to learn *is* not to perform. This is indicated by the GMC's requirement that WBA should be able to detect 'trainees who are struggling' (p. 2) and the RCR's decision aid which makes provision for additional training when trainees are deemed not to have achieved the required learning. However, the existence of 'reifications' such as curriculum documents, assessment tools and ARCP decision aids does not mean that the educational system that they are intended to support is functional, never mind 'robust'. This is particularly the case when it appears that an approach to assessment which was largely developed to support learning in school- or classroom-based contexts has been adopted by the professional regulator and imposed on a professional, clinical setting. Thus, one of the main aims of this study was to shed some empirical light on the extent to which the formative assessment system in postgraduate clinical radiology training in the UK appears to be functional, or fit for purpose.

### *2.5.2 Radiology training as competency-based education<sup>3</sup>*

Another feature of the UK postgraduate medical education system, and one which provides the educational backdrop for WBA, is the existence of curricula which are often described as being competency-based. What this typically means in practice is

---

<sup>3</sup> My interest in competency-based education is in the role of assessment, and particularly formative assessment, within these systems. A full consideration of the arguments about the nature of professional competence and whether it is genuinely reducible to lists of competences is beyond the scope of this thesis. See Lum, 2009, for a more comprehensive consideration of the issues at hand.

that the syllabus has been set out in detailed lists of statements, or ‘competences’<sup>4</sup>, which ostensibly describe aspects of observable performance. This is true of the clinical radiology syllabus (RCR, 2010), which extends to some 133 pages, with ‘competences’ being categorized as knowledge, skill or behaviour. These competences have also been grouped under two overarching headings: ‘generic competences’, which theoretically apply to doctors of all specialties, and ‘radiology-specific’ competences (RCR, 2010, p. 2).

The question of whether or not this statement-based approach can ever really capture ‘professional competence’ is one that has been tackled by a number of authors within and outwith medical education. For example, in developing his general critique of competency-based education and training (CBET), Lum (2009) turns initially to the discourse that developed around ‘the concept of education’ (p. 11) in the latter half of the twentieth century, and the references therein to the make-up of vocational training. Within this discourse, Lum perceives a pattern of thought that he refers to as an ‘orthodoxy’ (p. 12), in which allusions to vocational education are heavily laden with references to ‘skills’, often seeking to use the term to distinguish ‘training’ from ‘education’. The former is often cast as an activity that involves learners addressing a clearly specified programme of learning, in order to become proficient at clearly defined and rule-governed tasks. This is often then contrasted with ‘education’, which Lum (2009) argues is often framed as an activity that is distinct from, and superior to, training, precisely because of its complexity, lack of direct vocational utility, and its resistance to being narrowly specified. He characterises the vocational ‘orthodoxy’ thus:

Accounts of ‘training’ and ‘skill’...are characterized as follows: first, they are seen as related to specific or definite ends, and in this sense are characterized by a sort of confinement or narrowness of focus. Not only is it possible for these ends to be clearly specified, but so too can the skills required to achieve those ends, as can the processes of training necessary to impart those skills – skilled activity, in short, is something which can be tied down to clear-cut specifications and identifiable rules (Lum, 2009, p. 16).

---

<sup>4</sup> Medical education authors in the US tend to employ the term ‘competency’ to refer to overall capability and ‘competencies’ to refer to the individual behavioural statements that comprise certain curriculum documents. In this thesis I have adhered to the British convention of using ‘competence’ and ‘competences’, unless quoting from other authors. The exception to this is my use of the phrase ‘competency-based education’, as this phrase is used within the US and UK literature to refer to education that is described by lists of behavioural statements.

Whilst this account of the competency-based approach might be deemed unattractive, it could be argued that competency-based education offers a straightforward approach to capturing the elements of professional capability that comprise competence in medical practice. Authors such as Carraccio (2002) see this as a valuable alternative to the traditional approach to medical education, which relied more on the length of time served in training than the objective demonstration of capability in order for trainees to be deemed to have completed their formal professional education. However, not everyone is convinced that this is a suitable approach for medicine. Myerson (1998) puts it this way:

I would argue that we should avoid a system based on competencies. Such an approach trivialises professional behaviour with its complex reasoning and makes no attempt to assess these deeper aspects. Focusing on skills alone is easy. One can certainly be objective. A system of assessment based on skills is easier to organise, too. But the danger of focusing only on skills is that assessors will misvalue [*sic*] judgements and the complexities of how and when to use such skills, which are part and parcel of professionalism (Myerson, 1998, p. 1039).

Furthermore, some prominent voices within medicine perceive that the motivation for introducing a competency-based approach has more to do with accountability than pedagogy. For example, Grant (1999) observes that:

These new ideas very often are not about education at all. They are actually about external, managerial or political control (Grant, 1999, p. 272).

She may be at least partially justified in her view. lobst *et al.* (2010) describe the gatekeeping role of competency-based assessment in medicine, noting that 'regulatory organizations now require a demonstration of attainment of competency as part of their expectations; in some countries, this requirement now guides accreditation processes' (p. 651). It would be logical to conclude, therefore, that the role of assessment in such a system would be largely summative, aimed at confirming the achievement of competence. This would appear to be what Holmboe *et al.* (2010) mean when they state that, 'competency-based medical education (CBME), *by definition*, necessitates a robust...assessment system' (p. 676, my emphasis). They go on to say that, within a CBME context, 'effective assessment provides the information and judgement necessary to enable programme-level decisions about trainee advancement to be made reliably and fairly' (*ibid.*). However, this flies in the face of the GMC's and the

RCR's insistence that WBA is primarily formative. Indeed, Holmboe *et al.* (2010) go on to observe that, from the trainee perspective, 'CBME requires enhanced attention to formative assessment to ensure they receive frequent and high-quality feedback to guide their development and the acquisition of the necessary competencies' (p. 676). Therefore, it appears that to describe a system as being competency-based is not necessarily to say anything about the primary role of assessment within it, with the potential for strong arguments to be made for summative or formative approaches. However, in terms of how that assessment is conducted, it is possible to predict some of the potential impacts of a statement-based approach to education on any assessment, whether this is for summative or formative purposes.

### 2.5.3 Competency-based education and instrumentalist approaches to assessment

In terms of the role that the assessments play in practice, it is possible that the reduction of 'competence' to individual statements of competence, along with the requirement to have progression signed off annually, might encourage an instrumentalist approach to assessment. In other words, WBA might cease to function as assessments of learning or for learning and instead become assessments as *evidence of learning*, with trainees using them to demonstrate the achievement of sufficient competences in order to be deemed competent overall. Indeed, the GMC's guidance on WBA states that a trainee's portfolio 'should include comprehensive sampling across and within domains, using different [workplace-based] assessment methods and assessors, to build a clear picture of performance.' (GMC, 2010, p. 6). The consequence of this may well be a hoop-jumping or box-ticking culture, with the feedback aspect of the assessments receiving little attention. These fears have been articulated within medical education by the likes of Talbot (2004) and Leung (2002), and some evidence of a box-ticking culture has begun to emerge. Bindal *et al.* (2011) found that 9% of trainees in their study of WBA in paediatric training specifically referred to WBAs as a 'tick-box exercise' (p. 926). It is likely that the sentiment was shared by other participants, but was expressed more obliquely via comments such as, 'There is no interest in using the opportunity [of WBA] for education and feedback' or, 'Soul destroying, lots of very patronising paperwork. How can this raise the standard of medicine?' (*ibid.*, p. 924). In other words, WBA is seen by some trainees as little more than bureaucracy, required to demonstrate learning in a manner that is so prosaic as to generate little summative information or formative educational benefit.

A similar perception amongst assessors would serve to all but destroy WBA as a genuine assessment, whether summative or formative. However the format of the assessments may mean that even a more earnest engagement by assessors may stop short of being a genuinely formative process. Leading researchers in formative assessment, such as Black and Wiliam (2012), describe how the role of the teacher in formative assessment is to engage in a truly dialogical interaction with the learner. More is said about the nature of this process in Chapter 3, but it essentially involves the teacher taking an interest in the cognitive models, and metacognitive abilities, possessed by the learner. In practice, this often involves using skilful questioning which makes the learner aware of their own cognitive processes in order to improve their ability to self-assess in pursuit of their learning goals. Thus, genuine formative assessment requires assessors' engagements with learners to be 'constructed in the light of some insight into the mental life that lies behind the student's utterances' (Black and Wiliam, 2009, p. 13). In other words, formative assessment should involve assessors taking a cognitive constructivist view of learners and their learning. However, as James and Lewis (2012) point out, assessments that are aligned to behavioural competencies risk encouraging assessors to focus on the accurate observation of visible skills, and consequently take a much more superficial view of what constitutes learning. In support of this emphasis on the externally observable features of learning, the medical education assessment literature is replete with references to assessors needing training in order to improve their observational accuracy (e.g. see Herbers *et al.*, 1989; Noel *et al.*, 1992; Holmboe 2004; Margolis *et al.*, 2006; Cook *et al.*, 2009; Dath and lobst, 2010). Indeed, the training which I was involved in delivering for radiology assessors focused for the large part on assessors' attempts at scoring trainees' capability by watching video clips and using paper-based WBA forms to rate trainee performance. The remainder of the training addressed 'introductory' concepts in assessment (e.g. the difference between formative and summative assessment) and the use of a particular model for providing feedback. The emphasis in this latter task was often on how to go about communicating the assessor's judgement in a manner that might increase the acceptability of the message to the trainee. Little was said of the possibility of creating what Sadler (1998, p. 81) refers to as a 'non-convergent learning environment' for the exploration of trainees' underlying cognition or non-criterion-related capabilities.

This final thought, regarding the extent to which doctors have been trained to engage in a truly dialogic feedback process with trainees in the context of WBA, is one that will be taken up in the next chapter. For now, it is sufficient to say that there is scant reference in the medical education literature to this highly sophisticated approach to the provision of feedback which, as Perrenoud (1998) argues, is itself only one element of high quality formative assessment practice.

## **2.6 Summary**

The review of literature in this chapter has demonstrated that the concept of fitness for purpose is complex and multifaceted. It has been shown to be comprised of a number of components, and a consideration of whether or not the WBA system in clinical radiology is fit for purpose necessarily spans a number of domains. These include concepts of formative and summative assessment, the intended and unintended uses of assessments for different purposes, washback effects on teaching and learning, competency-based curricula and the professional setting within which WBA operates.

This complexity is not immediately apparent in the RCR's assertion that the assessments chosen for use in clinical radiology training are 'fit for purpose' (RCR, 2010, p. 161). In fact, given the lack of evidence presented by the RCR, this declaration appears to have been made prematurely. The review of assessment literature presented in this chapter strongly suggests that aspects of the WBA process are unlikely to be fit for purpose, not least because of the multiple and at times competing purposes that the assessments are intended to serve, as well as the complex environment into which they have been introduced. The review of literature in this chapter has therefore offered a partial answer to the main research question.

Perrenoud's (1998) observation that feedback is a necessary but insufficient component of formative assessment is noted. However, feedback remains an important element of formative assessment. Consequently, Chapter 3 contains a consideration of the literature on the role of feedback in formative assessment, which in turn underpinned the construction of a tentative framework against which the quality of assessors' written feedback in clinical radiology assessment could be compared.



## CHAPTER 3

### 3. Literature review II – Feedback in medical education

#### 3.1 Introduction

In asking the question about the fitness for purpose of the WBA system in clinical radiology, the previous chapter demonstrated: that any question of fitness for purpose in assessment is in essence a question of validity; that validity is a function not of assessments themselves but of the uses to which assessments are put; and that WBAs are intended to be primarily formative assessments, and thus should improve the process of teaching and learning. It could be argued, therefore, that any enquiry into the fitness for purpose of WBA should explore whether or not it can be demonstrated to improve learning. However, the task of demonstrating improved learning is highly complex, not to say potentially confounded by the number of factors that can impact on learning and the various meanings that might be ascribed to 'learning'. Thus, as I argue below, attempting to analyse learning as an end point may not be as logical as it may first appear when a suitably complex view of learning is adopted.

#### 3.2 Why analyse feedback in workplace-based assessment?

In his authoritative work on validity theory, Kane (2006) distinguishes between *validation* and *validity*. For Kane, *validation* is a largely theoretical exercise, in which hypothetical propositions about an educational intervention are analysed, or a theoretical case for the intervention is made. Taking formative assessment as one such example of an intervention, validation would involve examining the case for formative approaches, such as giving feedback, in supporting learning. Consequently, the review of literature in this chapter functions in part as a validation argument for feedback,

examining the theoretical case for feedback as a vehicle for delivering the improved learning that WBA is intended to facilitate. In doing so, the case is made for feedback being logically linked to improved learning, although not necessarily in a simplistic, predictable manner.

According to Kane (2006), a *validity* enquiry involves the examination of empirical evidence that the proposed intervention actually delivers on its theoretical promise. In the case of WBA in clinical radiology, this might imply that a validity enquiry would explore evidence that WBA results in improved learning amongst doctors. However, there are challenges to adopting this approach. One of the main challenges is the obvious difficulty in measuring improved learning when the measurement tool (i.e. WBA) is the educational intervention. Also, notwithstanding the earlier argument that positive washback should result in improved test performance, a number of other factors such as test-wiseness and teaching to the test could also improve performance in WBA independently of any contribution made by formative feedback. However, the definition of formative assessment proposed by Black and Wiliam (2009) renders this difficulty with verifying learning less important than it may first appear:

Practice...is formative to the extent that evidence about [learner] achievement is elicited, interpreted and used by teachers [and] learners...to *make decisions about the next steps in instruction* that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (p. 9, my emphasis).

As Black and Wiliam (*ibid.*) go on to make clear, their definition deliberately focuses on decisions about instruction, rather than the resulting outcome of instruction, in recognition of the unpredictability of learning. Linking a formative educational intervention to improved learning in a simplistic manner, they argue, fails to take account of the highly contingent nature of learning, as well as the agency of learners as participants or collaborators in their own learning. Thus, the lack of an objective measure of learning is not necessarily a weakness in an empirical validity study, as the intervention may not be linked to learning in a deterministic manner. This is somewhat contrary to Stobart's (2012) view that a validity enquiry in formative assessment *should* focus on learning (the outcome), yet Black and Wiliam (*ibid.*) make a compelling case for the complexity of learning rendering it less predictable and less amenable to objective measurement than some researchers may wish it to be. Consequently, blending Kane's (2006) view of a validity enquiry with Black and Wiliam's definition

(2009) of formative assessment, a validity study of a formative assessment intervention should involve an analysis of the instruction, not the learning outcomes linked to that instruction.

Applying Black and Wiliam's (2009) definition of formative educational practice to an assessment system has the effect of moving questions about the educational process to the foreground, and moving questions about learning outcomes to the background. In WBA, a key element of the process is the provision of feedback, and so the validity enquiry element of my research focusses on analysing the properties of the feedback provided to trainees as a process measure. Importantly, feedback is not being used here as a surrogate marker for improved learning at the level of individual trainees. Any finding of good quality feedback being provided cannot necessarily be taken to be a proxy for 'improved learning' – individual learners may fail to benefit from apparently excellent feedback for a number of reasons that are discussed later in this chapter. However, the validation argument aspect of this chapter makes the case for improved feedback being linked to improved learning *in general*, and so the provision of high quality feedback is taken to be important in its own right, whatever the specific downstream effects of the feedback may be for individual learners.

This approach to measuring elements of the educational process, rather than the outcomes of the process, is in keeping with approaches to measuring quality in healthcare for reasons not dissimilar to the arguments about the unpredictability of learning made by Black and Wiliam (2009). Consequently, a brief consideration of this approach to the measurement of quality in healthcare systems may help to illustrate the case for feedback as a legitimate gauge of the functioning of the WBA system.

### *3.2.1 Measuring quality in healthcare – a parallel case*

Measurement of how a clinical system is performing is an important aspect of healthcare quality improvement initiatives. However, it is often difficult to determine a suitable outcome measure due to the existence of a number of uncontrollable, unpredictable factors. These include, for example, the varying seriousness and complexity of patients' conditions and the various relative degrees of health to which they can be returned prior to discharge from hospital. It is common, therefore, for process measures to be used in place of outcome measures.

Mainz (2003), in his summary paper on defining clinical indicators in healthcare quality improvement, describes process as 'what is actually done in giving...care, i.e. the practitioner's activities in making a diagnosis, recommending or implementing treatment, or other interaction with the patient' (p. 525). Consequently, he argues, the measurement of these activities provides a legitimate indicator of the functioning of the healthcare system. There is a parallel here to analysing the performance of an educational system, in which practitioners carry out activities (such as giving feedback) which are aimed at bringing about an outcome (learning), and without which the outcome theoretically stands a lesser chance of being realised. This latter point is important. As Mainz (2003) points out, for a process measure to be useful, 'it must previously have been demonstrated to produce a better outcome' (*ibid.*, p. 525). For example, in the world of clinical medicine, there might be evidence from a randomised controlled trial that stroke patients experience better outcomes when they are admitted to a specialist stroke unit within four hours of arriving at the emergency department. If so, then recording whether or not stroke patients *are* admitted to specialist units within the four-hour time limit (a process measure) might be a more useful gauge of the state of a particular healthcare system than whether individual patients actually experience better outcomes several days or weeks down the line. This is because, due to a range of other factors (such as the patient's health prior to suffering a stroke, other medical complications that may arise during their stay and so on) patients may or may not achieve the outcomes that have been demonstrated to be linked to rapid stroke unit admission in more controlled circumstances.

A similar argument may be made for the measurement of the performance of an educational system or intervention – the apparently simple outcome (i.e. 'learning') may actually be too complex to provide a useful measure of the performance of the system. Consequently, it may be more suitable to substitute an evidentially-related process measure. Thus the validation argument element of this chapter, in which the case for feedback leading to improved learning is made, is important in establishing this evidence base and therefore underpins the element of empirical enquiry relating to validity in this work.

### 3.2.2 Feedback as a suitable process measure for WBA

In Chapter 2, WBA in clinical radiology was highlighted as being a primarily formative assessment, on the grounds that it is principally intended to support learning by the provision of feedback (GMC, 2010). It is important to note that feedback is not the only means by which formative assessment can be said to support learning. Mansell *et al.* (2009), for example, describe formative assessment much more holistically. They state that 'formative assessment is a central part of pedagogy' (p. 9), and cite examples of formative teaching and learning practices other than 'feedback', such as supporting students' metacognitive development and developing their self-assessment capabilities. Nevertheless, as acknowledged by Stobart (2008, p. 144), feedback 'is seen as the key to moving learning forward' in the context of formative assessment. Whether and how it does this in practice are matters of some considerable complexity, as is discussed later in this chapter. It is therefore challenging to dissect out the features of effective feedback in a general, decontextualized manner. However, as noted by Boud and Molloy (2013) in their review of feedback in higher and professional education, there are some principles of effective feedback that appear so recurrently in the educational literature that they would seem to offer a consensus view of effective feedback in education *in general*, if not a prescriptive formula for effective feedback in clinical radiology in particular.

In considering Mainz's (2003) condition that acceptable process measures should be evidentially linked to desired or intended outcomes, his concept of evidence is defined broadly:

linkages [between process and outcome] may be based on scientific literature; if little evidence exists, professional experience concerning these linkages may be distilled using consensus methods (p. 525).

Consequently, while my review of feedback literature focuses on empirical evidence from across a range of educational sectors and contexts, it at times includes reference to the expert opinion of leading authorities on the subject of formative assessment. Before embarking on a synthesis of the principles of effective feedback, however, it is useful to provide a backdrop to my study by establishing what is already known about the current state of feedback in medical education.

### 3.2.3 What is the current state of feedback in medical education?

In their now seminal paper on WBA, Norcini and Burch (2007), drawing on the work of Day *et al.* (1990), identified that, in the United States, 'the vast majority of first-year trainees in internal medicine were not observed more than once by a faculty member in a patient encounter where they were taking a history or doing a physical examination.' (Norcini and Burch, 2007, p. 855). Even with the introduction of WBA, this lack of observation has been identified by others in the field of medical education. Jackson and Wall (2010), in a study of 47 foundation trainee doctors in the UK, found that only 38% of the study population had been observed prior to their assessor completing their mini-CEX assessments. In research that demonstrates a lack of learner observation by senior colleagues, it seems intuitive to conclude that helpful feedback is unlikely to have been provided. Certainly, it is difficult to understand how assessors for the remaining 62% of Jackson and Wall's (2010) study population arrived at their decisions about assessment outcomes or the provision of feedback.

In reality, it seems that the answer to this question is that feedback is often not provided. The GMC's National Training Survey (GMC, 2013, p. 5) found that nearly a third of UK trainees (31.6%, n=52,484) had rarely or never received feedback from a senior colleague – a figure which was little changed from the previous year (32.7%). Other researchers in UK medical education have found corroborating evidence of a reported feedback deficit, but suggest that the shortfall may be linked to differences in perceptions of what constitutes feedback rather than a genuine lack of observation and feedback from teachers. For example, Sender-Liberman *et al.* (2005) found that only 17% of surgical trainees reported receiving helpful feedback, despite 90% of their senior colleagues reporting that they gave feedback which was, in their view, beneficial to the learners. These apparent differences in perception are not uncommon. Murdoch-Eaton & Sargeant (2012) found that while 96% of undergraduate students at a UK medical school (n=564) agreed that feedback on their work was important, only 58.8% reported receiving what they regarded as sufficient feedback in the course of their studies. In exploring the views of the participants, it became clear that some junior medical students undervalued or dismissed verbal comments as feedback, only identifying formal written feedback as being of any real worth.

It may be the case, therefore, that in surveys exploring the provision of feedback, respondents tacitly distinguish between verbal and written feedback. They may also distinguish between formal and informal feedback, with formal feedback more commonly being identified by learners as 'genuine' feedback, and informal comments on performance being comparatively undervalued and underreported. In the GMC survey mentioned above (GMC, 2013), there was an apparent attempt to differentiate between formal and informal feedback by enquiring about more objectively identifiable feedback vehicles, such as assessments and educational meetings, rather than asking about feedback itself. The result was that 27.2% of trainees (N=52,484) reported not having had a formal meeting with their educational supervisor, despite regular meetings of this kind being a mandatory requirement for doctors in recognised training posts. Furthermore, 30.1% of the same population reported not having had a WBA conducted during the training year, despite all UK postgraduate training curricula requiring several such assessments per year as a condition for progression to the next stage of training. It seems, therefore, that even when researchers enquire about mandatory feedback opportunities, learners report that these opportunities occur less frequently than required by the GMC and relevant medical royal colleges.

Difficulties with the provision of feedback in postgraduate education are not confined to medical training. As observed by Boud and Molloy (2013), student reports of insufficient feedback have persisted across the higher education and professional education sectors for some time. These authors are reluctant to embrace perceptual differences as a satisfactory explanation for inadequate feedback. In their view, this is tantamount to blaming the learners for failing to recognise feedback when it occurs. Furthermore, they are similarly scornful of attempts to address the problem by more effective signposting of feedback when it does occur. The premise here, they argue, is that there is nothing wrong with the feedback that is currently being provided, and so teachers do not take seriously any consideration of changes to their individual practice or to the system of feedback provision that is in place. Nonetheless, analysis of the feedback literature reveals that different concepts of feedback do exist, which may give rise to genuine perceptual differences as to whether or not feedback is happening. Thus, it is worth considering the feedback conceptions that are typically found in educational settings in order to explore more fully the nature and type of feedback that the WBA process in clinical radiology might be expected to deliver.

### 3.3 What is feedback?

Any exploration of the literature on feedback in education, medical or otherwise, quickly reveals that different concepts of feedback may be said to exist. For some authors, feedback is an entity, usually information about performance in a particular task, focussing either on the outcomes of the performance – which authors such as Kluger & DeNisi (1996) have called knowledge of results (KR) – or on some aspect of the execution of the performance (knowledge of performance, or KP), or both. Typically, the provision of this feedback information is characterised as an event, rather than a process, and the directionality tends to follow traditional hierarchical lines – teacher to pupil, lecturer to student, consultant to trainee. For other authors, feedback is better conceptualised as being a dialogic process - a conversation between teacher and learner, through which the learner comes to an understanding of their current level of performance and what is required in order to progress to the next stage. Other concepts of feedback also appear in the literature. For example, in their review of feedback concepts in clinical education, Van de Ridder *et al.* (2008) identified feedback concepts that they labelled ‘feedback as a reaction’ and ‘feedback as a cycle’ (p. 191). However, as these authors go on to state, ‘feedback as information is discrete, whereas both the reaction and cycle formulations are *processes*’ (p. 191, my emphasis). Thus, the feedback discourse in clinical education is dominated by two concepts – feedback as information and feedback as process, with the former being very much in the ascendancy.

For each of these two concepts of feedback, a range of aims or purposes has been invoked. Most commonly in medical education, improvement in professional knowledge or skills is the intended aim, but other stated purposes also exist. For example, Nicol (2013) adopts a metacognitive perspective, viewing the role of feedback as ‘progressively enabling students to better monitor, evaluate and regulate their own learning’ (p. 34). On the other hand, Webb *et al.* (2009) adopt a socio-cultural perspective, viewing the role of feedback in professional learning as being to support a process of becoming, by assisting learners in understanding the tacit rules, norms and value systems that constitute a particular professional community of practice. Analysis of the literature is therefore complex, not least because many authors in the field of medical education segue within one piece of work from one concept of feedback to



another, and from one purported aim or function of feedback to another, often without any reference to the potentially problematic nature of so doing. However, an exploration of the 'information' and 'process' concepts of feedback is useful in analysing the type of feedback that might be supported or promoted by the WBA system in clinical radiology.

### 3.3.1 Feedback as information

For Ende (1983), the author of what has been hailed by medical education researchers such as Bing-You and Trowbridge (2009) and Bernard *et al.* (2011) as a seminal paper on feedback in medical education, his primary concept of feedback seems to be that of feedback as an entity – specifically, feedback as information. Ende (1983) defines it thus:

In the setting of clinical medical education, feedback refers to information describing students' or house officers' performance in a given activity that is intended to guide their future performance in that same or in a related activity (*ibid.*, p. 777).

Ende is not alone in describing an information concept of feedback. In attempting to debunk the notion that feedback can be easily linked to educational impact – positive or negative – Latham and Locke (1991, p. 224) state that 'feedback is only information, that is, data, and as such has no necessary consequences at all'. A number of medical education researchers, such as Paul *et al.* (1998), Moorhead *et al.* (2004) and Rushton (2005) to name but a few, also describe feedback as being primarily the provision of information, and writers from the broader educational landscape are no less likely than medical educationalists to do so. In their review of feedback definitions in medical and non-medical education, Van de Ridder *et al.* (2008) cite definitions from general education handbooks that emphasise the 'feedback-as-information' concept. For example:

Feedback is information provided to the learner concerning correctness, appropriateness or accuracy. In short, feedback is information about a learner's performance (Meyer, 1995, in Van de Ridder *et al.*, 2008).

Bernard *et al.* (2011), in a review of feedback literature undertaken to make recommendations for training in emergency medicine, also conceptualise feedback as

information. This is not immediately apparent, as they appear to describe feedback as a process, however the process to which they are referring is actually that through which the educator comes to be in possession of information that they can then provide to the learner:

Feedback is the process by which the teacher observes a student performing an activity, analyses the performance, *and then provides information* back to the student that will enable the student to perform the same activity better in the future (p. 537, my emphasis).

For Bernard *et al.* (2011), and many authors, feedback is as straightforward as 'information provision'. This view does not resonate well with empirical evidence of researchers such as Rees *et al.* (2009), who highlight some of the challenges that can arise when giving feedback, especially when the message is (or is perceived to be) negative or unwelcome.

The difficulty with the 'feedback as information' concept is that when feedback is defined as narrowly as this, the link between information and learning is often assumed to be straightforward. Fullan (2001) acknowledges the difference between information and learning, and distinguishes between them by arguing that learning involves the social construction of knowledge in order *to make meaning from* data or information. Therefore, if feedback is information, a question for the educational researcher is whether and how that information becomes knowledge – in other words, how does feedback lead to learning?

### 3.3.2 *Feedback as process*

In considering how information becomes learning within educational settings, William and Thompson (2007), drawing on the earlier work of Ramaprasad (1983) and Sadler (1989) describe three processes that they view as being central to effective formative practice. These involve ascertaining:

- Where learners are in their learning
- Where they need to go
- What they need to do to get there.

Feedback that takes the form of a one-way transfer of information may have the appearance of addressing each of these processes. For example, in a Rad-DOPS assessment in clinical radiology, it could be argued that the assessor can observe where the learners are in their learning, as demonstrated by their technical competency in a given procedure. It might also be argued that the assessor, as an experienced practitioner, will know where the learner needs to go. Thus the role of feedback is to put the trainee in the picture according to each of these first two elements, and address the third by providing appropriate instruction. In fact, this example does not appear too far removed from Sadler's (1989) pronouncement that,

Formative assessment is concerned with how judgments about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning (Sadler 1989 p. 120).

However, this would be to ignore the active role of the learner in constructing their own learning. As Sadler (*ibid.*) goes on to argue,

The indispensable conditions for improvement are that the student comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced during the act of production itself, and has a repertoire of alternative moves or strategies from which to draw at any given point. In other words, students have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it (Sadler, 1989, p. 121).

This is fine as far as it goes, but Sadler's statement focusses on the learner's ability to self-assess or self-monitor, and does not address important issues of motivation and belief which, as is explored later in this chapter, can have powerful mediating effects on learning. His statement also appears to assume that the teacher holds the definitive concept of quality, and that it is for students to align themselves with this using a range of self-regulatory techniques. This contrasts to some extent with the view of Black and Wiliam (2012), who appear to acknowledge a more complex concept of the learner, and a more sophisticated, less predictable role for the teacher. Consequently, Black and Wiliam (*ibid.*) argue that a dialogical approach to feedback, as opposed to simple provision of information, is more likely to lead to effective teaching and learning. For Black and Wiliam (*ibid.*, p. 209) a 'dialogical interaction' is typified thus:

The teacher addresses a task to the learner, perhaps in the form of a question, the learner responds to this, and the teacher then composes a further intervention, in the light of that response (Black and Wiliam, 2012, p. 209).

Doing so, they argue, is the only effective way to take account of what Perrenoud (1998) describes as 'the cognitive and socio-affective mechanisms activated in the students' by the feedback (p. 85). However, Black and Wiliam (2012) also recognise that this dialogic approach requires a change in the traditional role of the teacher or assessor. Rather than acting purely as a subject expert, these authors describe the teacher as changing their usual teaching 'mode': focussing on what and how the learners will learn, rather than what and how they themselves will teach. This makes the teaching and learning process somewhat less predictable for the teacher:

In a formative mode, the teacher's initial prompt is designed to encourage more thought. The learners are more actively involved, but their responses are not predictable; thus formative interaction is a *contingent* activity. In such situations, the teacher's attention must be focused on what she or he can learn about the student's thinking from their response. However, what the learner actually hears and interprets is not necessarily what the teacher intended to convey, and what the teacher hears and interprets is not necessarily what the learner intended to convey. In a genuinely dialogic process, the teacher's own thinking may come to be modified through the exchange (Black and Wiliam, 2012, pp. 212-213, original emphasis).

This description of the role of the teacher in a formative assessment system represents a major departure from the teacher-centred approaches to feedback provision that are still so readily apparent in the medical education literature. That said, there are authors within the field of medical education who espouse a more dialogical approach, whether or not they label it as such. In their often-cited work on the teaching of medical consultation skills, Pendleton *et al.* (2000, p. 69) highlight the importance of gauging the intentions that lie behind the performance of the learner, rather than simply taking the performance at face value:

We can provide feedback about two distinct matters - the doctor's intentions and the attempts to bring about the intentions. We are required, therefore, to understand why the consultation was as it was (Pendleton *et al.*, 2000, p. 69).

They go on to give an example of how an assessor's judgement of a learner's performance may be wide of the mark if the learner's intentions are not taken into consideration:

Consider the example of a doctor who spends a considerable amount of time in a consultation exploring the possibility that a patient has a psychosexual problem underlying his physical complaint. If, at the end of the consultation, the patient has not revealed any such problem, it is possible that the problem does not exist. However skilful the doctor's attempts, the teacher may feel that they were inappropriate for that patient. If, on the other hand, the teacher feels that the patient did have a problem of this kind but that it was not discovered, he may decide that the doctor could have explored it more effectively (Pendleton *et al.*, 2000, p. 69).

In other words, if the assessor is aware of why the trainee is behaving as they are, they should be in a better position to make a judgement about the behaviour, and consequently better able to conduct a valid feedback discussion with the learner. In the example given above, the same performance could be judged as either *insufficient*, due to the failure to uncover an underlying psychological condition, or *inappropriate*, due to the likelihood that the patient's condition had a somatic, rather than psychosomatic, cause. Pendleton *et al.* (*ibid.*) maintain that discovering the intention behind the trainee's line of questioning will reveal to the teacher whether the feedback conversation should be focused on faulty clinical reasoning, which had led the trainee to develop an inaccurate working diagnosis, or on the consultation skills that had failed to reveal the underlying problem.

The message here is that observation alone is not sufficient to allow the teacher to provide appropriate feedback. More needs to be uncovered by the teacher in order to develop a better understanding of why the trainee chose to act as they did. Shortcomings in a learner's performance that are due to what Prins *et al.* (2006, p. 300) term 'availability deficiency' – the absence of a knowledge base or skill set upon which to draw – may be straightforwardly observed. However, in order to understand 'production deficiency' – the failure to deploy an aspect of an existing knowledge base or skill set – further exploration is needed. The risk of failing to explore the learner's intentions is that production deficiency is mistaken for availability deficiency, and the trainee is faced with remedial advice about improving their knowledge or skills when the source of their decisions to act as they have done may lie elsewhere. It is for this reason that Watling *et al.* (2012a, p. 602) advise:

Feedback [should] be treated as a conversation with the learner, in which the [teacher] seeks to understand not only the learner's perception of his or her own performance, but also the meaning of the task to the learner and the motivation with which he or she has approached it.

Clearly, according to this notion of feedback, a top-down, teacher-led approach is unlikely to yield the information that the authors believe to be important. Importantly, the Watling *et al.* (2012a) recommendation goes beyond gaining a simple understanding of the trainee's intentions, as proposed by Pendleton *et al.* (2000). They refer additionally to the importance of understanding the learner's motivation and the meaning they have attached to the activity around which the feedback discussion is based. As is discussed later in this chapter, motivation and meaning, as well as beliefs, are important factors in the learner's decision as to whether and how to respond to any feedback information given. Accordingly, an observation-based formative assessment system, such as the WBA system in clinical radiology, risks ignoring essential components of learning if it fails to take account of important aspects of the learner, such as intention, motivation, emotion and beliefs, which may not be readily assumed from observed behaviour.

In summary, it seems from the evidence presented thus far that feedback effectiveness is likely to be enhanced by the teacher and learner engaging with each other in a verbal, dialogic process. However, the WBA feedback system in clinical radiology training emphasises written feedback, and so in analysing the ability of the WBA system to deliver formative feedback it is important to consider the nature of the feedback that can be delivered in this manner, and the extent to which it is capable of being truly dialogical.

### **3.4 Written feedback in workplace-based assessment**

In considering the value of written feedback in general, a number of authors make important claims for its utility. Orsmond *et al.* (2005) and Carless (2006) have demonstrated that learners often review written feedback with the intention of making improvements to their work, and so the longevity of written feedback may thus support reflection, consolidation and repeated attempts to comprehend and apply the advice or

instruction of the tutor. Jolly and Boud (2013) highlight the potential for written feedback to be private, allowing learners to avoid the embarrassment of public criticism or even public praise: as Hattie and Timperly (2007) identify, even positive feedback can be perceived negatively if delivered in the presence of a social group whose collective values are not welcoming of individual praise. Whilst this is often the case in adolescent social groups, it may be also the case amongst groups of high performing professionals, such as groups of clinical radiology trainees.

Another advantage of written feedback over verbal feedback, as long as an immediate response is not required, is that it takes time to construct. This affords both teacher and learner with the opportunity to pause and reflect which, as Jolly and Boud (2013) observe, may be especially valuable if the learning or assessment episode has been intense or emotionally charged, as can often be the case in clinical settings. Furthermore, assessors can modify their original thoughts, rephrasing them if necessary to ensure that their comments are not as terse or critical as they might otherwise have been. Other linguistic amendments or revisions can also be made – Boud (1995) highlighted the paralysing nature of closed, judgemental statements which allow the learner no right of reply. Having time to redraft comments may afford assessors the opportunity to rephrase their feedback in order to ask open questions or make suggestions for further improvement instead.

The precise phrasing of assessors' comments may not be the only barrier to genuine dialogue in written feedback. As Crisp (2007) and Bloxham and Campbell (2010) point out, the written format itself makes it challenging to conduct a genuinely dialogic feedback conversation. This is especially true if it is delivered at the end of a term or clinical placement. Bloxham and Campbell (2010) also allude to the particular difficulty posed in professional settings, in which learning involves not just achieving mastery over a particular set of skills or body of knowledge, but participating ever more fully in complex communities of practice. Learning in this sense, they argue, is unlikely to be supported effectively by formal written feedback. Instead, professional learning occurs by immersion in the community itself, with extensive opportunities for 'observation, imitation, participation and dialogue' (Bloxham and Campbell, 2010, p. 292).

None of this is to say that verbal feedback, by comparison, is necessarily valuable or genuinely dialogic. As revealed by Murdoch-Eaton & Sargeant (2012), verbal feedback

may be dismissed by learners, who may think it of little value, or fail even to identify it as feedback. In particular, these researchers found that early stage medical students did not identify verbal comments as feedback, preferring written comments instead:

Very little feedback is given; most of it is oral and general (*ibid.*, p. 717).

Conversely, senior medical students in Murdoch-Eaton & Sargeant's (2012) study did appear to value verbal feedback, especially when learning in the clinical environment. The authors viewed this as a maturational difference, but there may be other drivers for the persistence of these perceptions. In particular, the early stages of medical school often emphasise 'traditional' academic performance, such as lecture-based teaching and assessment via essays and written examinations. These features of the curriculum may well generate a reliance on, or expectation of, detailed written feedback. By comparison, the latter years of medical school tend to emphasise clinical experience through extended clinical placements. Students in that phase of the curriculum are more likely to be aware of the difficulty of providing written feedback in busy clinical environments, and more appreciative of verbal comments on their clinical performance. This likelihood is supported by the remarks of one study participant, a senior medical student, who commented:

Feedback is more focused now, it's better i.e. towards clinical things and being a doctor rather than in previous years where it was more general and theoretical (*ibid.* p. 718).

In an echo of Murdoch-Eaton and Sargeant's (2012) study, Jolly and Boud (2013) observe that there are marked differences between feedback practices in higher education compared with professional education. They perceive that in the former, feedback is much more likely to be written than verbal, whereas professional environments, and particularly clinical settings, are far more likely to feature verbal feedback. Paul *et al.* (2013) observe that clinical medicine has a strong oral tradition of teaching and learning, and Jolly and Boud (2013) observe that verbal feedback tends to predominate even when these professional learning situations are 'staged' – such as simulation-based training – with the feedback model in these cases often involving small-group discussion between the teacher, learner, peers and even (at times) patients. Consequently, the arrival of an approach to workplace assessment in clinical radiology that emphasises written rather than verbal feedback may have created



something of a culture clash, in which a clerical model of feedback that has been previously applied in classroom-based learning contexts has been adopted and applied in a professional, clinical context. Furthermore, this may have been done with insufficient regard for the feasibility of providing effective written feedback in the busy clinical setting, along with insufficient conceptualisation of the type of learning that occurs in this context and the extent to which it might be effectively supported by episodic written feedback comments.

### **3.5 Why is feedback necessary?**

When introducing a radical new system of formative assessment and feedback into a professional setting, such as clinical radiology, with all of the additional demands it places on assessors and learners, it seems reasonable that a case should be made for its necessity. As Eva and Regehr (2008) observe, 'it is generally well accepted in health professions education that self-assessment is a key step in the continuing professional development cycle' (p. 14), and so it is useful to briefly examine the evidence as to whether or not self-assessment *can* be relied on to guide learning, or whether, in most cases, there is a need for feedback to support the learning process.

Whilst self-assessment is often held to be a key step in the process of professional learning, there is evidence that these assessments may not necessarily be accurate. For example, Kruger & Dunning (1999) have demonstrated how, in three apparently unrelated domains – humour, grammatical ability and logical reasoning – participants who were ranked in the bottom quartile in an objective test of their abilities in those domains systematically overestimated their own test performance, and overrated their performance with respect to their peers. Figure 3.1 shows this effect for participants (n= 84) who sat a short test of their grammatical ability. As can be seen, the students whose scores fell into the bottom quartile (n=17) overestimated their performance by nearly 50 percentage points (ie in the 61<sup>st</sup> percentile, compared with actual performance which fell in the 10<sup>th</sup> percentile) and rated themselves as being in the 67<sup>th</sup> percentile relative to their peers. It might be argued that these individuals, having no criteria or other objective frame of reference, might be forgiven their overestimates. However, it transpired that even when presented with the better quality work of their top-performing peers (i.e. those in the highest quartile) they failed to identify that the

work was of a much higher standard. Kruger & Dunning (1999) conclude, therefore, that ‘it takes one to know one’ (p. 1126) when asking students to identify the high standards of performance of their peers.

It might be argued that medicine, as a demanding profession with high academic entry requirements, might only be drawing from the upper percentiles of the population, thus the findings of Kruger and Dunning’s (*ibid.*) study may not apply. However, the population upon which these researchers drew was that of the undergraduate school of psychology at Cornell University in New York state, USA – an exclusive university with a reputation for high academic standards – and so the gap in academic capability that might normally be said to exist between doctors and the rest of the society from which they are drawn is unlikely to exist to the same extent here. Furthermore, a second finding in the Kruger and Dunning study was that students in the top quartile tended to significantly underestimate their performance with respect to an objective standard and the performance of their peers. Consequently, it seems that even high performing students may benefit from receiving external information about the quality of their performance.

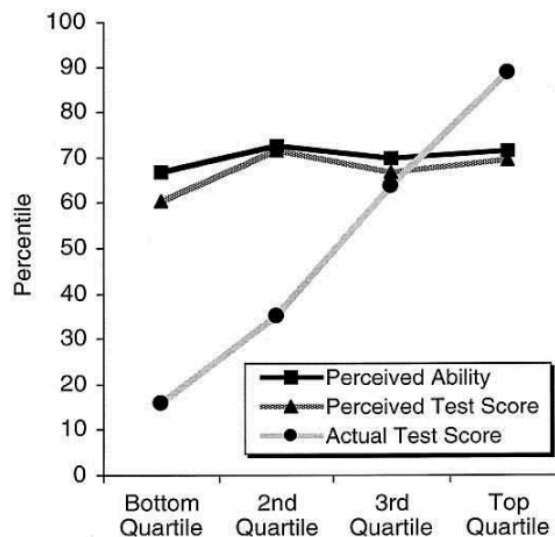


Figure 3.1 Perceived grammar ability and test performance as a function of actual test performance, Kruger & Dunning (1999, p. 1126) (Permission applied for).

Subsequently, Davis *et al.*'s (2006) systematic review of the literature on the accuracy of physician self-assessment found broad support for Kruger and Dunning's (*ibid.*) findings within medical education. Of the 17 studies that met the inclusion criteria, describing 20 comparisons between self and external assessments, 13 of the comparisons demonstrated either no correlation or an inverse relationship. In keeping with Kruger and Dunning's (*ibid.*) findings, physicians who performed least well in the objective assessment, and who were most confident of their performance, typically demonstrated the least accuracy in self-assessment. More recently, Sawdon and Finn (2014) were able to reproduce the 'Kruger and Dunning effect' amongst Year 1 and Year 2 medical students (n=74) in the UK who had recently completed a practical exam in human anatomy. Again, students in the lower two quartiles tended to overestimate their performance, while students in the uppermost quartile tended to underestimate their performance.

Eva and Regehr (2008), in their review of the phenomenon described by Kruger and Dunning (*ibid.*), attribute inaccurate self-assessment to a number of factors. These include social factors (individuals may, historically, have received overly generous assessments from others who have been keen to preserve relationships, leading to an inaccurate self-concept), cognitive biases (such as information neglect and imperfect recall of events) and socio-biological factors (such as the potential adaptive advantage of maintaining a positive self-concept). In fact, Eva and Regehr (2008) argue that in light of the multiple and diverse influences on the formation of self-concept, accurate self-assessment is not just difficult, it may be impossible.

It seems, therefore, that there are grounds to be wary of the sufficiency of self-assessment in professional learning. This is not to say that no-one can learn through a process that is driven by self-assessment rather than peer or teacher assessment. However, it seems that learners at all levels of capability may benefit from feedback from an external agent in order to, as Black and William (1998, p. 9) have phrased it 'make decisions about the next steps...that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence'. Thus, it might be said that a case for attempting to provide external feedback has been made. Whether or not external feedback is effective is another matter, and one that can be considered at least in general terms from a review of the evidence to date.

### 3.6 Is feedback effective?

In considering the question of feedback effectiveness, it is important to measure the impact or effectiveness of feedback on its own terms. That is to say, one needs to first ask the question, 'What job is feedback meant to do?' before attempting to answer the question of how well it does it. For example, feedback that is intended to develop the reflective capability of general practice trainees may have no bearing on the performance of those trainees in the Membership of the Royal College of General Practitioners (MRCGP) exams. In this example, a self-reported gauge of the value of feedback in supporting reflection or an objective assessment of some reflective writing may be more appropriate measures of feedback effectiveness than trainees' performance in the MRCGP exam. The focus of my study is the effectiveness of WBA as formative assessments. It is therefore useful to recall what formative assessment is intended to achieve before considering how and whether feedback may be said to contribute to this.

A useful definition of formative assessment, provided by the Assessment Reform Group (ARG, 2002), is that it is:

The process of seeking and interpreting evidence for use by learners and their teachers, to identify where learners are in their learning, where they need to go, and how best to get there (pp. 2-3).

In other words, formative assessment is to be used for the purpose of improving learning, and the feedback component which constitutes formative assessment should contribute to this goal. However, as alluded to earlier, evaluating whether or not formative assessment, including feedback, *does* improve learning is challenging, not least because of the the unpredictability of what is learned. As Black and Wiliam (2012) point out, a cognitive-constructivist view of learning implies that, 'because students are active in the construction of their own knowledge, what they construct may be very different from what the teacher intended' (p. 20). Socio-cultural perspectives such as Lave and Wenger's (1991) situated learning theory introduce further complexity into the concept of learning, by acknowledging the participatory nature of learning within communities of practice, as well as the distributed nature of learning and the idea that it is mediated by physical and cultural artefacts. Thus, a simplistic notion of feedback effectiveness must surely be rejected when considering the value of feedback in

supporting learning. Instead, when considering the published evidence for feedback effectiveness, it seems wiser to adopt a probabilistic stance. In so doing it is possible to consider the evidence of feedback effectiveness in one setting, or a number of settings, and suggest that on balance of probabilities it is likely (but not guaranteed) to have utility in a different setting. With this in mind, it is worth considering what has generally been found in the literature to comprise effective feedback, and what approaches to feedback have been found to be less effective or even detrimental to learning.

### *3.6.1 The importance of the feedback message*

Notwithstanding the important individual and socio-cultural factors that may impact on feedback effectiveness, a number of studies have demonstrated that certain features of the feedback message itself can either promote or inhibit learning.

#### *Feedback valency - positive feedback versus negative feedback*

As previously mentioned in section 3.4 in this chapter, feedback comments which are generally positive and which lack information about performance deficits or learning needs have been associated with a lack of learning effect in classroom assessment settings (see Smith and Gorard, *ibid.*, for example). A similar lack of learning effect in relation to positive comments has also been noted in medical education – Sargeant *et al.* (2007) demonstrated that doctors in their study tended to accept positive feedback at face value, with no reported changes in behaviour being made. That is not to say that positive comments *per se* are of little value in supporting learning. Hattie and Timperley (2007), for example, calculated that positive feedback taking the form of ‘reinforcement’ was responsible for an impact on learning that was equivalent to a mean effect size of 0.94 standard deviations (SD) (p. 84). For comparison, this is more than twice the effect size of educational interventions in general, which was estimated by Hattie (1999) to be around 0.4SD. Thus ‘reinforcement,’ as a type of positive feedback, appears to be capable of adding value to learning. On the other hand, Hattie and Timperley (2007) found that the mean effect size of a different type of positive feedback, ‘praise,’ was substantially lower than reinforcement, at only 0.14SD. It seems, therefore, that a degree of caution should be exercised when categorising feedback as ‘positive,’ as different types of positive feedback may be seen to have very different effects on learning.

The apparent impact of negative feedback is also interesting. According to Hattie (1999), one type of negative feedback, 'corrective feedback' (p. 9), was also associated with an effect size of 0.94SD, however this was revised downwards to 0.37SD in a subsequent paper (Hattie and Timperley, 2007). The effect sizes presented by these authors should perhaps be treated with a degree of circumspection – the dramatic downward revision of the corrective feedback effect was not addressed in their later paper and raises some doubt about the veracity of their other statistics. However, it is arguably useful to have some indication of the potential effectiveness of particular types of feedback intervention, as long as the figures themselves are treated with due caution. Hattie and Timperley (2007) themselves acknowledge that the true picture is substantially more complex than the impression conveyed by a single statistic, with additional factors such as specificity, timeliness, the complexity of the task and learner-centric factors all interacting in complex ways.

#### *Feedback specificity and connection to the assessment criteria*

The complexity involved in analysing feedback, referred to by Hattie and Timperley (2007), is important. For example, as previously illustrated, positive feedback can be resolved into components such as reinforcement or praise, each with its own potential impact on learning. Other important features may over-ride these components, and may act to modify their effectiveness for better or worse. One important feature of reinforcement versus praise, for example, is that the former is typically characterised by *specificity* – the learner in receipt of feedback that genuinely qualifies as reinforcement must be clear about the aspects of their performance that are being commented upon positively. This requirement for specificity has been noted by Van de Ridder *et al.* (2008) in their review of feedback literature in the field of clinical education. In fact, they regard unspecific comments as unworthy of the feedback label:

Feedback must contain a minimum amount of specification to serve its purpose. Utterances that cannot be understood by the feedback recipient in behavioural terms (i.e. in terms of what has been done well or what could be improved) should not be called feedback (p. 194).

Feedback that lacks specificity, argue Berglas and Jones (1978), can cause learners to become confused when attempting to make valid attributions about their success. As

Thompson and Richardson (2001) have demonstrated, confusion regarding appropriate feedback attribution can lead to self-handicapping techniques, such as learners reducing their effort in subsequent tasks, resulting in a deterioration in performance. Hattie and Timperley (2007) also connect unclear or unspecific feedback with confusion and uncertain attribution amongst learners:

Students' attributions about success or failure can often have more impact than the reality of that success or failure. There can be deleterious effects on feelings of self-efficacy and performance when students are unable to relate the feedback to the cause of their poor performance. Unclear evaluative feedback, which fails to clearly specify the grounds on which students have met with achievement success or otherwise, is likely to exacerbate negative outcomes, engender uncertain self-images, and lead to poor performance (p. 95).

Further evidence of uncertainty linked to unspecific or unclear feedback has been presented by Williams (2007), who described how learners on a university-based creative writing course found unspecific feedback unhelpful, particularly when laden with rhetoric or jargon. Fedor (1991) had previously described the link between unspecific feedback information and learners' lack of certainty in choosing how to respond to the feedback, and Sargeant *et al.* (2007) have demonstrated a similar lack of clarity amongst doctors who felt unable to respond to unspecific and unclear feedback in a clinical setting.

For learners, a clear link between feedback and the assessment criteria may help demystify the grounds upon which they have received feedback. This is not to say that to invoke assessment criteria is necessarily to provide specific feedback. Nicol (2010) classifies feedback that is clearly linked to assessment criteria or learning outcomes as 'contextualised', and discusses this separately from specificity, which he characterises as 'pointing to instances in the student's submission where the feedback applies' (p. 512-513). However, referring to the assessment criteria may at least have the effect of allowing learners to associate feedback comments with any mark or grade that has been awarded, such that they can safely predict that a change in the aspects of their performance that have been highlighted should result in improved performance in a re-run of the same assessment. The challenge posed by WBA in radiology, however, is that repeat or follow-up assessments are not necessarily conducted by the same assessor, due to the nature of the working environment with its reliance on shift

working and frequently-changing rotas. Consequently, new assessors who are unaware of previous feedback comments may inadvertently deliver a confusing or contradictory message. Of course, the risk is that sticking rigidly to assessment criteria cultivates what Torrance (2007) describes as a convergent approach to assessment in which learning that lies outwith the specifications of the curriculum is disregarded, thus limiting the assessment's formative potential. However, it does not follow that straying from the assessment criteria necessarily gives rise to divergent formative assessment, particularly in an environment where the teachers' primary expertise is typically clinical, rather than educational. Rather, departing from the assessment criteria may simply mean that the feedback is unhelpfully vague.

A synthesis of the literature on feedback specificity suggests that it is a good idea to be as specific as possible when giving feedback to learners, not necessarily to focus their attention narrowly upon criteria, but to avoid the risk of ambiguous or nebulous comments. Yet Kluger and DeNisi (1996) sound a note of caution, indicating that while they have found ample supporting evidence for the benefit of specific feedback information, they have also uncovered some evidence that it may be possible for feedback to be *too* specific. It is possible, they argue, that feedback about the intricate details of a task can lead learners to focus on minutiae in a manner which is actually detrimental to overall task performance. However, a potentially greater risk is the converse – that feedback is too general to be helpful. As Kluger and DeNisi (*ibid.*) point out, specific feedback may not guarantee learning, but 'non-specific [feedback] cannot accomplish it' (p. 268).

#### *Target of the feedback – person or performance?*

Kluger and DeNisi's (*ibid.*) finding that detailed feedback about task performance may be unhelpful is part of their broader analysis of published evidence on feedback interventions, in which their primary argument is that feedback may be classified not so much in regard to specificity but in terms of its intended or likely target. According to Kluger and DeNisi (*ibid.*) these categories are: *task level* feedback, which is primarily aimed at identifying whether the details of the task have been successfully accomplished; *process level* feedback, aimed at the underlying processes on which task success is based (such as error detection or cueing); *regulation level* feedback, aimed at aspects of self-regulation such as self-assessment, self-efficacy and



motivation; and *self level* feedback, aimed at important aspects of a person's self-concept. Of these, feedback at the self level, which draws learners' attention towards their self-concept, was found to be the least useful of all feedback. This is especially the case when the task involved is complex, as demonstrated by Baumeister *et al.*, (1990). Kluger and DeNisi (*ibid.*) hypothesise that this is due to affective changes that lead to the diversion of resources away from task learning processes. Based on their meta-analysis of feedback evidence, Hattie and Timperley (2007) concur:

...there is a distinction between feedback about the task (FT), about the processing of the task (FP), about self-regulation (FR), and about the self as a person (FS). We argue that FS is the least effective, FR and FP are powerful in terms of deep processing and mastery of tasks, and FT is powerful when the task information subsequently is useful for improving strategy processing or enhancing self-regulation (p. 90).

In fact, Hattie and Timperley (*ibid.*) go on to demonstrate that feedback about the self (FS) is so unhelpful that it seems to dilute the effect of other types of feedback, such as feedback about the task (FT), which would otherwise be a potentially powerful driver of subsequent learning. In doing so, they provide illustrations of feedback comments which they would classify as self-orientated and therefore unhelpful: 'You are a great student,' or 'That's an intelligent response, well done.' (Hattie and Timperley, 2007, p. 90). Stobart (2008) argues that this is important as it runs contrary to teachers' instincts to praise learners, whether in school-based or professional education. Investigating this aspect of feedback formed an important part of my research.

#### *Timing of the message – immediate versus delayed.*

In considering when the feedback message should be delivered, the literature paints a complex picture, and it seems there is no simple distinction to be made between the effectiveness of immediate versus delayed feedback. For example, the complexity of the task is an important modifying factor. According to Stobart (2008), immediate feedback seems to be effective in supporting the learning of new, complex tasks, as it acts to reduce frustration and ensure that the learner can make progress with the task in hand. Conversely, in simple tasks, early feedback interventions can cause 'feedback intrusion' (Stobart, 2008, p. 162) which can also frustrate learners and derail their learning efforts. The type of learning that is being promoted may be another important factor to consider. According to Shute (2008), immediate feedback is beneficial for

producing immediate learning gains. On the other hand, delayed feedback seems to be more effective in supporting transfer of learning to different tasks. Finally, the ability and stage of the learner may also be important factors. According to Stobart (2008), novice learners and 'low-achieving' learners (p. 162) are more likely to benefit from immediate feedback, an effect which Paas *et al.* (2004) explain by recourse to cognitive load theory. Novice learners who receive feedback early, they argue, are likely to experience a reduction in their cognitive load. According to Wulf and Shea (2002), this is necessary to avoid overload by reducing the cognitive load to levels that facilitate learning. Conversely, experienced or high-achieving learners may benefit more from delayed feedback, having been given the opportunity to exhaust all of their various problem-solving strategies before being helped to find the answer. These learners are less likely to be overwhelmed by what Paas *et al.* (2004) refer to as the 'intrinsic' cognitive demands of the task (p. 4) – i.e. dealing with the sum total of the interacting auditory/verbal and visio-spatial requirements inherent in completing the task. Consequently, their learning is more effective if it is allowed to proceed without interruption under the conditions naturally generated by engagement with the task itself.

Considering the timing of feedback within radiology WBA, the feedback process as currently constructed is such that feedback is given terminally, i.e. at the end of the activity that is being assessed. This has the effect of embedding a particular feedback approach within the WBA system, such that feedback is delayed regardless of the experience or ability of the learner or the complexity of the task. Thus there may be an impact on how useful this formative feedback may be, due not to the particular feedback practices of individual assessors but to the structure of the assessment system that has been implemented.

### *3.6.2 The importance of the feedback source*

#### *Credibility of the assessor*

Another important factor in determining how to respond to developmental feedback appears to be the judgement of the learners as to the importance of acting on the comments. Doctors who participated in Sargeant *et al.*'s (2007) study on responses to multi-source feedback found that the participants' first step when deciding how to respond to criticism was typically to analyse whether the critical comment had come

from a patient or medical colleague. For these doctors, it transpired that the patients' views were paramount. In other words, if the criticism came from a medical colleague and was not borne out by the patients' comments, then doctors reported being unlikely to change their practice. Conversely, if a patient gave, in the words of one doctor, 'a bad report' (p. 587), then doctors attended to this regardless of the favourable reports of their colleagues. The doctors in the Sargeant *et al.* (2007) study were fully qualified family doctors working in Canada, who were participating in a developmental multisource feedback exercise, and each of these factors may have had an influence on how they weighed up the degree to which the feedback mattered: fully qualified doctors may have less to lose than trainees, whose assessment outcomes inform judgements about their progression through training; successful family doctors have a particularly close and ongoing relationship with their patients, when compared to the often episodic, short term encounters that hospital-based doctors, such as clinical radiologists, have with their patients; and finally, the fact that the process was purely developmental meant that, in practice, all of the doctors would have been free to reject all of their feedback with no negative consequences.

Many learners, such as those enrolled on university courses or, in the case of my study, trainee doctors enrolled in a postgraduate training programme, are not as free as Sargeant *et al.*'s (2007) participants to decide on whether or not to respond to feedback. For example, Williams (1997) found that university students enrolled on an academic writing course were found to attend to formative feedback that came from their professors as a priority, rather than choosing to express themselves as they might otherwise have wished. In doing so, they appeared to accept, albeit temporarily, that their preferred way of expressing themselves was 'wrong', choosing instead to follow the recommendations of their assessors. Put another way, it seems that under the influence of high stakes assessment, learners do the things that they believe will allow them to be successful in the assessment. This is relevant to my study as trainee radiologists are in a particularly vulnerable position, in that they rely on positive WBA outcomes in order to progress satisfactorily through training. Thus, it might be the case that, regardless of the feedback accuracy, which was the focus of the last section, trainees may feel the need to respond to any feedback offered by senior colleagues in order to demonstrate engagement with the educational process.

Notwithstanding the tactical imperative of responding positively to feedback from particular sources, it seems that attributes of the assessor aside from their strategic or political importance are influential in learners' decisions as to how to act on feedback. In their interview-based study of the role of feedback in self-assessment, Sargeant *et al.* (2010) found that trainees valued feedback from 'trusted, credible supervisors' (p. 1218). The teacher's credibility can of course span several domains. Professional credibility might include the learners' perceptions of the teacher's clinical competence, or their professional or academic standing. Personal credibility might include whether the learner feels that the teacher is generally fair, holds them in high regard, and is interested in their development. For example, Watling *et al.* (2012a) found that learners needed to believe that their teacher was 'engaged in the creation and exchange of informed and accurate feedback' (p. 594) in order to accept the validity of the feedback they provided. Educational credibility might include whether the teacher is aware of relevant aspects of training (such as curriculum content and workplace based assessment requirements), or whether they have any educational training or qualifications. In support of the professional and educational aspects of credibility, Murdoch-Eaton & Sargeant (2012) described how early stage medical students greeted peer feedback with suspicion, believing it to lack validity and reliability. They expressed a preference for feedback from individuals that they believed to be credible, which in their context was restricted to academic staff from the medical school. Senior students, however, in conceptualising feedback as the opportunity for discussion and reflection, were more willing to value conversations with peers as legitimate feedback, emphasising personal credibility over professional standing or educational expertise. This willingness to embrace peer feedback echoes the earlier finding by Sargeant *et al.* (2010) that undergraduate and postgraduate trainees must often rely on peer feedback in the clinical setting, due to the absence of formal feedback from supervisors.

Interestingly, the failure to include peer feedback in the WBA process is a notable point of departure from the classroom assessment/AfL origins of the formative assessment ideas and language that have been adopted by the GMC and the RCR. According to RCR guidance,

Most raters/assessors should be supervising consultants, doctors in training more senior than the trainee under assessment and experienced radiographic, nursing or allied health professional colleagues (RCR 2010, p. 161).

The requirement for trainee assessors to be more senior than the trainee being assessed provides a clue as to where the summative/formative balance may lie in these assessments, as may the use of the phrase 'the trainee *under* assessment' (*ibid.*, my emphasis). It seems that, despite the reality of peer-peer or near-peer learning in everyday clinical practice, this avenue for the provision of feedback is not one that is to be facilitated by the formal WBA process.

### *Credibility of the evidence*

If the WBA process as constructed by the RCR is therefore actually more about the verification of ongoing competence development than being the main vehicle by which trainees receive formative feedback on their day-to-day learning, the trainee doctors being assessed will want to be sure that the feedback they receive is based on accurate observations. However, there is evidence that trainees do not believe this to be the case. For example, Bindal *et al.* (2011) found that paediatric trainees in the West Midlands Deanery were unconvinced about the relationship between their workplace-based assessment outcomes and their capability as doctors. In fact, trainees may be right to question the accuracy of the assessments on which their feedback is based. For example, Herbers *et al.* (1989) conducted a study in which 32 medical education faculty members were asked to assess the performance of a trainee who had agreed to be videoed performing a simulated clinical encounter for the purposes of the research. There was wide variation in overall ratings of the trainee's competence, which could not be explained by disagreements amongst faculty as to the standard of clinical practice required: 50% of the assessors rated the trainee's performance as marginal, 25% failed him, and 25% viewed his performance as satisfactory. In a similar study conducted by the same group of researchers, Noel *et al.* (1992) found that the assessors only identified around 30% of the standardised trainees' strengths and weaknesses using open-ended assessment forms. This level of accuracy improved to 60% or better when structured observation forms were introduced, which appears to be an argument in favour of WBA checklists such as those featured in the Rad-DOPS forms used by the RCR. However, there were still wide discrepancies in the assessors' judgements of the trainees' overall clinical competence. For example, one of the trainees was judged to be barely or insufficiently clinically competent by 31% of assessors, with the other 69% of assessors rating the

same individual's skills as being satisfactory or superior. A second standardised trainee was judged to be either minimally competent or incompetent by 48% of assessors, with the other 52% viewing the same trainee's skills as satisfactory or superior. The assessors in this US study were some 209 senior doctors with dedicated educational roles, and so experience did not seem to relate well to accuracy. Of more concern was the researchers' finding that the accuracy of the assessors' scoring failed to improve after training in the use of the assessment tools.

The finding that training has been shown to produce little improvement in the reliability or accuracy of assessors' scoring has been replicated in several studies. For example, Holmboe *et al.* (2004a) found that training produced a significant increase in assessors' stringency when assessing aspects of trainees' clinical skills, compared with their counterparts in an untrained control group. However, there was no indication that these more stringent assessments were more accurate or reliable than the others: the intervention group reported being significantly more comfortable with the assessment tool than the control group, and so the more stringent scores may simply reflect a greater confidence in administering harsh judgements, rather than necessarily reflecting more accurately the performance of the trainees. There is a tendency amongst some medical educators to regard more stringent assessments as being more likely to be accurate. However as noted by Lösel & Schmucker (2003), stringency, and its opposite, leniency, are potentially both types of assessor bias, each of which can lead an assessor to score a learner less accurately than they should.

It would appear, then, that assessment accuracy in WBA is difficult to achieve, and has remained resistant to training effects. This may provide trainees who wish to dismiss negative feedback with ample reason to do so, regardless of its accuracy in their case. This particular example of learners demonstrating agency in choosing and constructing their own learning is a reminder of this essential, ultimately decisive, factor in the potential effectiveness of feedback. Yet the active role of the learner is often ignored in the literature about feedback, which tends to focus instead on the feedback message or the feedback-giving behaviours of the assessors. Thus it is worth considering the aspects of learners that are particularly pertinent in relation to the potential effectiveness of feedback.

### 3.6.3 What are the important characteristics of learners in responding to feedback?

In presenting a model of how feedback might be said to function in the field of human learning, Weiner (1967) draws on his engineering background to offer the analogy of system control:

If...the information which proceeds backwards from the performance is able to change the general method and pattern of the performance, we have a process which may very well be called learning (p. 84).

His efforts to extend this cruise control concept to the teaching and learning process illustrate a common assumption made by many educators and researchers when considering the role of feedback in promoting learning, and the analogy is one to which medical educationalists (e.g. Ende, 1983) and non-medical educationalists (e.g. Harlen, 2012) alike have instinctively been drawn. However, Weiner's notion shines a spotlight on the ability of *the information* to change the 'pattern of performance' (*ibid.*) without referring to the entity that is producing the performance in the first instance – the learner. The well-recognised agency of learners in constructing their own knowledge has already been referred to in this chapter. Thus, it is appropriate to consider some of the personal characteristics of learners that may have a bearing on how they choose to respond to feedback.

#### *The role of learner beliefs and motivation*

In weighing up whether and how to respond to developmental feedback, it is arguably the case that learners do so based at least in part on a self-evaluation of the sufficiency of their cognitive or intellectual capacity. However, there is an emerging body of evidence which suggests that individuals' *beliefs* about the demands that are placed on their cognitive capabilities are more important than whether they actually have the capacity to carry out sustained cognitive effort in pursuit of their improvement goals. These beliefs are not typically articulated by the individuals concerned, hence they are referred to by researchers such as Miller *et al.* (2012, p. 1) as 'tacit theories'. Evidence of the influence of these beliefs has been established experimentally within the field of psychology. For example, Job *et al.* (2010) explored the extent to which participants who were engaged in a task requiring sustained levels of self-control (an experimental gauge of participants' cognitive resources) were able to maintain high levels of self-

control in a subsequent activity. These researchers found that participants who believed that they could only exercise self-control for a finite length of time demonstrated lower levels of self-control in the second activity than those who believed that the capacity for self-control was limitless.

In later work, set in a more explicitly educational context, Miller *et al.*, (2012) explored the impact of these 'tacit theories about the nature of intelligence' (p. 1) on the ability of college students to engage in working memory-intensive learning tasks. They found that participants who believed that willpower is a 'limited resource' (*ibid.*, p. 1) – i.e. that it is capable of being depleted – behaved in controlled conditions as though this were the case. These participants were found to be unable to sustain their capacity to learn beyond the first half of a standardised test. On the other hand, study participants who believed that willpower was unlimited continued to demonstrate a capacity to perform intensive cognitive activity for an extended period of time, persisting with their efforts to the end of the test.

A number of authors – most notably Dweck (2000) – describe these learner-held beliefs about cognitive capacity as being either entity theories or incremental theories. According to Dweck (2000, p. 2), learners who hold an 'entity theory' of intelligence believe that intelligence is a fixed property of individuals, and that success in learning can be attributed to having either sufficient or insufficient quantities of intelligence. Thus they are likely to believe that some things cannot be learned by them (or anyone else lacking the requisite measure of intelligence) and so do not persist with their attempts at learning. Learners who hold a more malleable concept of intelligence – so called 'incremental theorists' (Dweck, 2000, p. 21) – are characterised by their belief that intelligence can be increased with effort, and so are more likely to display mastery-oriented learning characteristics: enjoying learning, actively seeking challenge and persisting in the face of difficulty.

These self-theories are not the only important psychological aspect of learners when determining how to respond to feedback, as there is evidence that other aspects of motivation, such as the meaning that learners attach to a task, can also impact powerfully on learner responses. In exploring the importance of the meaning of different learning tasks to learners, Watling *et al.* (2012a) researched the extent to which regulatory focus theory, first proposed by Higgins (1997), might provide a



framework for understanding learners' perceptions of and responses to feedback in medical education settings. Regulatory focus theory posits that learners operate out of one of two motivational states, which are characterised as promotion focus and prevention focus. Learners who are engaged in an activity that is concerned with aspiration or accomplishment are likely to exhibit a promotion focus. In other words, they seek or value confirmatory (or positive) feedback, and may employ strategies to avoid, minimise or dismiss critical or negative feedback. On the other hand, individuals who are engaged in routine or obligatory tasks are more likely to exhibit prevention focus, in which case they are likely to value more critical comments on their performance and dismiss or in some other way devalue positive remarks on their accomplishments. Watling *et al.* (2012a) found that when tasks were clearly identifiable as either aspirational or, conversely, routine, regulatory focus theory offered a useful construct for exploring learners' responses to feedback (p. 593). However, they also identified that the nature of professional learning in the clinical context was so multifaceted and multi-layered that it was difficult to identify the majority of activities as being clearly either aspirational or routine. Hence the predictive power of the regulatory focus concept was limited in practice.

To the extent that these studies demonstrate a general point about the power of learners' privately held beliefs about themselves, about intelligence, and about the meaning of particular learning tasks, it would seem to be challenging to predict how any learner in particular may choose to respond to a given feedback episode. It is for this reason (although it was not necessarily articulated as extensively as has been done here) that Black and Wiliam (2009) were loath to specify improved learning in their definition of formative assessment. Yet for all of the complex, contingent factors that impact on the effectiveness of feedback in the case of individual learners, this chapter has demonstrated that it is possible to draw some generalised conclusions about the likely educational impact of certain types of feedback. These have included the value of positive and negative comments, as long as they are specific, clearly linked to transparent assessment criteria, and task-focused rather than person-focused, as well as the value of feedback which is timely enough to be acted upon by the learner and not delivered at the end of a course or placement.

The next step in my research, having elucidated a number of general properties of high quality feedback, was to construct an initial theory-driven framework for the coding of

assessors' written feedback comments in clinical radiology WBAs. To this end, the literature on written feedback in WBA was reviewed with a view to identifying previously-validated frameworks that might be appropriately adapted and used in my study. The broader educational literature was also explored for the same purpose.

### **3.7 Judging the quality of feedback**

In attempting to make a judgement about the quality of formative feedback, researchers in a range of educational contexts and sectors have taken various approaches to the analysis of the feedback message itself. Some of these approaches have involved the a priori construction of deliberately weighted or judgmental frameworks, informed by literature reviews or the consensus views of experts. At other times, feedback data have been approached with an initial framework that was more neutral in its tone, with judgements being applied to the findings *a posteriori*. This final section of Chapter 3 analyses examples of each of the above, and identifies a framework which provided a starting point for the analysis of the empirical feedback data in my research.

#### *Judgemental analytical frameworks*

In keeping with the former approach – the a priori construction of a judgemental framework for analysing feedback – Van de Ridder *et al.* (2008) conducted a systematic review of feedback concepts in literature drawn from the social sciences, medical education, and what they term the 'general literature' (p. 190): dictionaries, encyclopaedia and other general reference texts. In doing so, they synthesized their findings in order to propose two broad categories, 'weak feedback' and 'strong feedback' (*ibid.*, p. 195), with the intention that they be used to support research into the quality of feedback in clinical education (see Table 3.1, below).

Table 3.1 Characteristics of weak and strong feedback, after Van de Ridder *et al.* (2008, p. 195).

<b>Weak feedback</b>	<b>Strong feedback</b>
Competencies that are not observable	Well observable tasks and competencies
Uninformed or non-expert observer	Expert observer and feedback provider
Global information	Highly specific information
Implicit standard	Explicit standard
Second hand information	Personal observation
No aim of performance improvement	Explicit aim of performance improvement
No intention to re-observe	Plan to re-observe

In doing so, these authors categorised as ‘strong’ a number of features of feedback that concur with the findings in the literature reviewed within this chapter: ‘strong’ feedback, they argued, should be based on observed performance, should be specific, should allow comparison with an explicit standard, should be aimed at improving performance, and should be part of an ongoing educational process. It is worth noting that Van der Ridder *et al.* (*ibid.*), like many authors in this field, gravitated towards the concept of feedback as information, with the feedback process being conceptualised as the provision of this information. To wit, their definition of feedback:

[Feedback is] specific information about the comparison between a trainee’s observed performance and a standard, given with the intent to improve the trainee’s performance (Van de Ridder *et al.*, 2008, p. 189).

This definition reflects the dominance of the ‘feedback as information’ concept revealed by their review, and as with Black and Wiliam’s (2009) definition stops short of including improved learning as a component of the definition. Accordingly, Van de Ridder *et al.*’s (*ibid.*) framework provided a potentially useful approach to the analysis of data in my study, given that my research was focused on the analysis of written feedback comments in radiology WBAs.

However, even within the concept of 'feedback as information', the judgemental element of the framework ('weak' versus 'strong' feedback) seemed problematic, not least as it necessarily consigns some important learner capabilities to the realm of weak feedback. These capabilities include the cognitive and affective competences that experienced medical educators, such as Ende (1983), argue are essential elements of proficient clinical practice. In addition, the judgement of feedback as either 'strong' or 'weak' suggests that it can be simplistically assigned to a binary category, despite the authors' own illustration of the composite nature of feedback within their framework (see Table 3.1). The authors also provide no guidance as to how feedback that satisfies only some of the criteria in one category, or some of the criteria in both categories, should be classified. Nesbitt *et al.* (2014) appear to have encountered this particular limitation in their use of Van de Ridder *et al.*'s (*ibid.*) framework to classify the written feedback provided to UK medical students within a particular WBA known as a supervised learning event (SLE). In their research, Nesbitt *et al.* (*ibid.*) created a third category, which they labelled 'neither strong nor weak' (p. 281), although again it is not clear how the thresholds for distinguishing between 'strong', 'neither strong nor weak' and 'weak' feedback were determined.

A more sophisticated approach to analysing feedback in medical education was taken by Prins *et al.* (2006) in their exploration of the written feedback provided to trainee general practitioners in the Netherlands (see Table 3.2).

The criteria that comprised the framework were developed prior to coding by a group of four medical practitioners, and were refined according to a Delphi process. As can be seen, their framework was intended not only to guide the coding and categorising of feedback, but also to support the researchers in coming to a judgement about *how well* each assessor had addressed each of the researchers' feedback quality criteria by assigning a score to the feedback comments. The coding scheme and scoring system, in combination, illustrated the expectations of the researchers with respect to the quality of written feedback provided by assessors. These expectations appear to have been that the feedback should include 'substantial doctor-patient communication-related remarks', and the 'description of behaviour and explanation of remarks throughout the report' (p. 295). In addition, for feedback to attract the highest scores it had to contain a balance of positive and negative remarks, questions that promote reflection throughout the report, examples given from the practice of others (including

the assessor's own practice) and constructive advice for improvement. Furthermore, it was expected that the first person should be used throughout the feedback report, that the report should be clearly structured, and that the report should be pieced together from short descriptions of what the assessor has observed.

Table 3.2 Coding framework used by Prins *et al.* (2006, p. 295) in their analysis of written feedback provided to general practice trainees.

Main category	Sub category	Good achievement		Average achievement		Minimal achievement	
1. Criteria	Content	Substantial medical and doctor–patient communication related remarks	30	Some medical and some doctor–patient communication related remarks	15	No or hardly any medical and doctor–patient communication related remarks	0
	Explanations	Description of behaviour and explanation of remarks throughout the report	20	Some descriptions of behaviour and some explanation of remarks	10	No description of behaviour and no explanation of remarks	0
2. Nature	Remarks	Balanced number of positive and negative remarks	10	Positive remarks dominate	5	Negative marks dominate	0
	Posed questions	Questions fostering reflection throughout the report	10	Some questions that stimulate reflection	5	No questions in the report	0
	Repertoire	Good external examples (e.g. own experiences)	5	Unclear examples	2	No examples	0
	Advice	Good and clear suggestions for improvement; constructive advice	10	Some suggestions for improvement	5	No suggestions for improvement; no constructive advice	0
3. Writing style	Structure	Clear structure e.g. chronology	5	Unclear structure	2	No structure	0
	Formulation	Short descriptions	5	Key words dominate	2	Only key words	0
	Style	First person throughout the report	5	Sometimes first person	2	No first person, judging	0

In considering this framework as the basis of an initial coding framework for my study, it was clear that despite having been arrived at by expert opinion rather than a review of evidence, it did contain some features that were supported by the feedback literature. These included: the provision of positive and negative feedback comments; making reference to specific, observed, behavioural aspects of performance; and the importance of being as clear as possible in providing feedback, such that the assessor and trainee have a shared understanding of what is being discussed. However, upon further consideration, the framework was not a particularly good fit for my research on practical or theoretical grounds. Firstly, Prins *et al.*'s (*ibid.*) framework was clearly intended to be applied to a fairly lengthy, summary feedback report, whereas the written comments on clinical radiology WBAs are intended to be more focused and, in general, not particularly lengthy. The scoring system also appeared problematic, as the scores were not justified by recourse to any theory – for example, clarity of structure was afforded a maximum of 5 marks, whereas reference to doctor-patient communication was afforded a maximum of 30 marks. Arguably, a poorly structured report that makes extensive reference to doctor-patient communication could be of lesser educational value than a well-structured report that only refers to a few instances of doctor-patient communication. There was also no justification given for some of the gradation in certain scores. For example, while a 'balanced' report (equal numbers of positive and negative remarks) was awarded 10 marks, a report that was more positive than negative was awarded 5 marks, whereas a report that was more negative than positive was awarded 0 marks. This appears ideologically rather than theoretically driven as there is no clear evidence to support the dominance of positive comments in feedback versus the dominance of negative comments. In fact, as Kluger and DeNisi (1996) have demonstrated, there is often no easy link between feedback valency and impact on learners.

The aspects of the Prins *et al.* (2006) framework that were aligned to feedback theory, and which could reasonably be expected to be present in radiology WBA feedback, were considered for inclusion in an initial coding framework (without awarding any scores to comments). These aspects were: the presence of positive and negative comments; comments that were based on observed behaviours; and comments that were developmental in nature. The concept of assessors providing feedback that was specifically intended to stimulate reflection idea was interesting, however the format of the Rad-DOPS form suggested it was unlikely that assessors would have posed open-

ended reflective questions in their written feedback. However, the possibility was not ruled out, and trainees' comments in particular were analysed for evidence of reflection.

The literature search revealed one further approach to the coding of written feedback comments in WBA, utilised by Canavan *et al.* (2010) (see Figure 3.2). As with the Prins *et al.* (2006) framework, these researchers set out to make judgements about what they termed the 'quality' of written feedback (p. S106). However, unlike Prins *et al.* (*ibid.*), Canavan *et al.* (*ibid.*) made this judgement retrospectively and qualitatively rather than prospectively and quantitatively. In other words, whilst the framework was constructed in advance of the coding process, it was not aligned to any scoring system, and the researchers made their judgements about feedback quality *a posteriori*, based on the overall patterns of feedback that emerged through their study. In addition, the wording of the coding framework was non-judgemental in its description of potential feedback characteristics. This degree of objectivity made it an appealing option as a starting point for my study. In addition, the Canavan *et al.* (2010) framework was theory-driven in the first instance, and therefore supported the authors' attempts to compare the features of written feedback that they found in their research with what had been found in the education research more globally. This was well aligned to my own interest in judging the quality of written feedback in Rad-DOPS assessments against a standard derived from educational theory, and so again this framework offered a potentially useful basis for the initial coding of the data. Consequently, this framework was adopted as a starting point for my analysis. Even so, some differences between the Rad-DOPS assessment and the particular type of WBA at the centre of Canavan *et al.*'s (*ibid.*) study meant that modifications to the framework were required even before initial coding. The framework was also modified inductively throughout the coding process. Details of these modifications are provided in Chapter 4.

### **3.8 Conclusion**

This chapter has set out the theoretical case for feedback being linked to improved learning. In doing so, an argument has been made that feedback that meets certain criteria is likely to be linked to improved learning, although the relationship is probabilistic rather than deterministic. This theoretical argument for the value of feedback in supporting learning aligns with Kane's (2006) notion of a validation

argument, and so comprises one element of my analysis of the fitness for purpose of the WBA system in clinical radiology.

To support the empirical element of my research, the published evidence on effective feedback was synthesised in order to construct an initial coding framework for the analysis of written feedback comments in clinical radiology WBA. The decision to analyse feedback as a measure of the fitness for purpose of the WBA system was based on the unpredictability and lack of verifiability of any learning outcomes that may have been linked to individual WBAs. This approach was justified using a parallel case argument, in which the practice of measuring elements of process is accepted as valid in healthcare quality improvement circles due to the difficulty of measuring an objective outcome of the system. My approach to analysis was therefore to use a theoretically-derived (and then inductively modified) framework to code large samples of written feedback in WBAs from the first three years of the newly-introduced WBA system in clinical radiology. In addition, other key metrics such as the numbers of assessments undertaken by the learners and the timing of these assessments were calculated and analysed. The chapter that follows sets out the methodological considerations and decisions that shaped the empirical aspect of my research.



## CHAPTER 4

### 4. Methodology

#### 4.1 Introduction

This chapter sets out the details of the study design which was aimed at answering the main question at the centre of this research. In doing so, key considerations in the selection of the overall approach to the research and in the choice of particular methods are discussed. The strengths and limitations of the chosen methods and the implications for the study are also examined. The resulting research process is described in detail, with consideration being given to measures taken to enhance the validity, reliability and generalisability of the findings.

#### 4.2 Summary of the research questions

The aim of this study was to generate empirical evidence regarding the validity of the system of workplace-based assessment and feedback that has been implemented in postgraduate training in clinical radiology in the UK. In particular, I was interested in exploring the patterns of use exhibited in the national e-portfolio record, as well as analysing the quality of the written feedback provided to trainees and the nature of trainees' engagement with assessors' written comments.

Therefore, the main research question was:

**Is the system of workplace-based assessment and written feedback in postgraduate clinical radiology training in the UK fit for purpose?**

The sub-questions that supported the main research question were:

- What are the claimed purposes of workplace-based assessment in clinical radiology training, and to what extent do any multiple purposes appear to be in conflict?
- What are the documented features of the system of WBA and feedback in postgraduate clinical radiology training, and how do they compare with what is already known about effective formative assessment?
- What are the qualitative characteristics of the written feedback provided by assessors to clinical radiology trainees in workplace-based assessments, and how do these compare with the features of effective feedback found in the literature?
- What, if any, conditions appear to govern the provision of effective feedback in workplace-based assessments in clinical radiology?
- Can assessors in clinical radiology deliver feedback of sufficient quality to support the development of these trainee doctors?
- What is the nature of clinical radiology trainees' written comments, and is the written feedback process dialogical?
- Can workplace-based assessment and feedback in postgraduate clinical radiology training be said to support the learning of trainee doctors?

It was clear at the outset of the process that the complexity of the main research question and related sub-questions would require a number of different approaches to the research, resulting in a multi-methods study design. In addition, a number of aspects of the questions could be perceived to be problematic in terms of the meaning attached to particular terms or phrases, such as references to 'fitness for purpose' or 'quality'. Further complications arose from the nature of the data set, as described below. These and other challenges, and the steps taken to address them, influenced the study design and are discussed below.

### 4.3 Access to data

Having determined to undertake an empirical study, the principal challenge in the first instance was gaining access to relevant data. Having approached the RCR at the outset of the research, the only data to which they were willing to provide access was that held within the national clinical radiology e-portfolio. This national e-portfolio is the combined record of all radiology trainees' individual e-portfolio records, and therefore contains all of the assessment data for trainees within the specialty in each training year. Accordingly, once anonymised by the RCR it proved to be a rich source of the formally recorded outputs of the WBA process for radiology trainees throughout the UK. However, there are potential limitations associated with a purely documentary source of data, and so other approaches to data collection were considered.

One of the key considerations was whether it would be possible to evaluate the assessment encounter directly. Certainly, there is often a compelling degree of face validity associated with naturalistic observational studies, with authors such as Adler and Adler (1994) labelling observation 'the most powerful source of validation' within the methodological spectrum (p. 389). However, serious threats to validity exist. As observed by Denzin and Lincoln (2000), 'most social scientists have long recognised the possibility of the observer affecting what he or she observes' (p. 674). Thus, observational research in the context of WBA may well have impacted on how the assessment and feedback process was conducted, especially if the researcher is associated with the participants' medical royal college. Furthermore, observer bias can serve to limit the validity of the data in ways that may be anticipated or unanticipated by the researcher. According to Robson (2011), a prime example of anticipatable bias is selective encoding, in which the observer's expectations of a situation colour their perception of what is happening. This may be countered to some extent by the researcher being aware of this threat to validity, but it is difficult to offset this bias effectively as it is often unconscious. Less anticipatable (and arguably less perceptible) threats also exist. As Denzin and Lincoln (2000) point out:

...the plain fact is that each person who conducts observational research brings his or her distinctive talents and limitations to the exercise; therefore the quality of what is *recorded* becomes the measure of usable observational

data...rather than the quality of the observation itself' (p. 676, original emphasis).

Consequently, there are methodological limitations that may compromise the veracity of observational data, however persuasive the data might appear to be.

Another barrier to conducting an observational study was presented by the nature of the clinical environment. This setting typically includes patients, who may have complex or serious medical conditions, and also includes other health professionals whose role was not under investigation in this study. Thus, the observation of real-time assessment events was ruled out - the consent of assessors, trainees, other health professionals and patients would have been required, and this proved to be a significant barrier to this approach being practicable.

Another approach, which would have removed some of the ethical and practical barriers present in an observational study, would have been to employ self-reported measures of assessment effectiveness. Self-reported measures typically include methods such as questionnaire surveys, interview studies, and more longitudinal approaches such as asking teachers to maintain descriptive logs or reflective journals of their assessment activities (see, for example, the work of Stiggins and Conklin, 1992). Of these methods, the one that was primarily considered for inclusion in this study was the use of interviews to explore the perceptions of radiology assessors and trainees with regard to the workplace-based assessment process. This would have offered the potential for in-depth analysis of the rich qualitative data typically yielded by the interview approach. However, given the anonymised nature of the e-portfolio data set, it would not have been possible to relate the interview responses of these participants to any of the specific examples of assessment and feedback found in the data set. Neither would it have been possible to draw generalisable conclusions from what would have been a relatively small number of assessors in order to explain the patterns of assessment practice and feedback apparent from the analysis of the whole data set.

In any case, my consideration of the strengths and limitations of these approaches proved to be largely academic. The RCR did not grant access to radiology assessors or trainees and so none of these approaches was viable. Rather, the e-portfolio record of WBA outputs comprised the only data that the RCR were able to provide, and

consisted of the anonymised assessment scores and written comments of the assessors and trainees throughout the UK for each year since WBA had been launched. My approach was therefore to examine empirically, and on a large scale, the formally recorded outputs from the WBA process to determine whether they have the potential to be an effective means of supporting the development of competence in trainee radiologists.

#### **4.4 Which workplace-based assessment?**

Trainees in clinical radiology undertake a number of different types of workplace-based assessment throughout the training year. The clinical radiology curriculum (RCR, 2010) makes it clear that of these assessments, two are of particular importance and comprise the majority of formal workplace-based assessment events within any one year of training. These assessments are known as the mini image interpretation exercise (mini-IPX) and the radiology direct observation of procedural skills (Rad-DOPS). Therefore, in asking the question about written feedback in the context of workplace-based assessment, it was clear that the bulk of the assessment data in clinical radiology training was associated with these two assessments. However, it became apparent at an early point in the research process that, given the scale of the data held within the e-portfolio and the time required to code the qualitative feedback data, the analysis of two or three years' worth of data for both of these assessments would not be feasible. Consideration was given to analysing the data associated with both assessments within a single training year, but it was decided instead that exploring one of the two assessments over three years would offer more insight into whether and how the WPBA system had changed over the initial stages of its introduction. Thus the question that arose was which of the two assessments to use for my study?

My initial intention had been to use the mini-IPX, due to its apparent similarity with the mini clinical examination exercise (mini-CEX), an assessment that is commonly used in the medical specialties both in the UK and the USA, and about which there is an appreciable amount of published research evidence. However, despite the superficial similarity of the assessment names, the two are in fact 'false friends,' as the mini-CEX is a real-time assessment based around a patient consultation, whereas the mini-IPX is

an assessment of a radiology doctor's ability to analyse medical images (such as plain film x-ray images), come to sound diagnostic conclusions, and write a clear report on the findings. Crucially, none of this involves any interaction with a patient. The Rad-DOPS, on the other hand, is a real time assessment of a doctor's performance in conducting radiological procedures, usually on fully conscious patients with whom the doctor must communicate throughout the process. I therefore decided to use the Rad-DOPS assessment for my research, as the dimension of the doctor-patient interaction that it contains seemed more intuitively comparable with the clinical work that most doctors do. This also offered the potential of comparing my findings with what had previously been found in research on other workplace-based assessments that involved doctors interacting with patients, such as the mini-CEX assessment and the DOPS assessments used in other specialties.

## **4.5 Study design**

### *4.5.1 What type of study?*

In setting out to answer the research question at the centre of this study, it became clear that no single method was likely to be sufficient. Consequently a mixed methods approach was adopted. However, rather than choosing mixed methods at the outset, my approach aligned with what Creswell and Plano Clark (2011) refer to as an 'emergent' study design (p. 54), arising out of a dynamic approach to the research rather than an approach that was pre-determined and 'typology-based' (*ibid.*, p. 55). My original intention had been to focus on the content of assessors' written feedback statements, analysing these qualitatively in order to draw conclusions about the potential educational value of the feedback provided. However the review of literature prompted me to consider other aspects of the WBA that could usefully be analysed in order to develop a more complete picture of how this formative assessment system was functioning. Features such as the timing and frequency of the assessments could add additional context and were therefore included, resulting in a more multi-faceted approach to the research. Crucially, the analysis of pairs of assessor and trainee comments was necessary in order to establish whether written feedback exchanges could be said to be dialogical.

It is important to consider the appropriateness of adopting a mixed methods approach, as it is apparent from the research methods literature that there is a range of views on whether and how methods should be mixed. On one hand, some authors take a pragmatic view of mixed methods approaches, accepting as a rationale that different methods have relative strengths and weaknesses, which may compensate for each other when used in combination. For example, Greene *et al.* (1989) are content to define mixed methods as any research design which uses 'at least one quantitative method...and one qualitative method' (p. 256). Creswell and Plano Clark (2007) expand on this definition, stating that:

As a method, it focuses on collecting, analysing and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches, in combination, provides a better understanding of research problems than either approach alone (p. 5).

Other authors are keen to emphasise a more holistic notion of 'mixing', thus establishing the mixed methods approach as methodology rather than method. Tashakkori and Teddlie (2003), for example, suggest that mixed methods research should be typified by mixing throughout the research process, from conception to inference, and Johnson *et al.* (2007) have produced an integrative definition of mixed methods research that similarly emphasises mixing at all points of the research process. This notion has not been universally welcomed, with some authors seeing the mixing of worldviews, philosophies and paradigms as being problematic. Creswell and Plano Clark (2007), for example, point out that 'different paradigms give rise to contradictory ideas and contested arguments – features of research that are honoured but cannot be reconciled' (p. 27). However, the mixed methods researcher, they argue, need not be derailed by these contradictions, but rather acknowledge and embrace them as 'different ways of knowing about and valuing the social world' (*ibid.*, p. 27). In taking this view, which they describe as pragmatic, Creswell and Plano Clark (*ibid.*) contend that the research *question* is therefore foregrounded, with research *methods* and their underpinning philosophies and worldviews being duly subordinated. This was the view that I took when conducting my study, in that rather than emphasising a particular world view or paradigm and conducting my research from that perspective, I was interested in forming a view of the fitness for purpose of the formative assessment system in clinical radiology that was informed by the best use of the data at hand.

None of this is to say that a pragmatic approach permits the indiscriminate assembly of methods to qualify as mixed methods research. Rather, as Symonds and Gorard (2008) emphasise, the selection and blending of methods should be purposeful. A helpful description of five primary purposes has been provided by Greene *et al.* (1989), of which one – complementarity – was of key importance in my research. According to these authors, complementarity is concerned with seeking the ‘elaboration, enhancement, illustration [and] clarification of the results from one method with the results from another method’ (p. 259). This, they highlight, is different from ‘development’ (p. 259), which is a more linear use of the results from one method in order to then inform the development of another method. Rather, complementarity may take a more sophisticated form, employing some methods in parallel as well as in sequence, in a manner which is at times not easily distinguishable from the pursuit of the goal of ‘expansion’ (p. 259) – a deliberate attempt to expand the scope of the research via the introduction of additional methods. The defining feature of complementarity versus expansion or development is the rationale, which is that the range of methods used is employed in order to ‘increase the interpretability, meaningfulness and validity’ of results by drawing on the differential affordances of the different methods used. In my study, the methods utilised were selected in order to explore a number of aspects of the WBA system that could each be said to be linked to the overarching concept of ‘fitness for purpose’, thus creating a more complete, and therefore a more valid, picture of the functioning of the assessment system. The reasons for the decisions that were made about the use of each particular method, which thus gave rise to the overall research design, are discussed below.

#### *4.5.2 Planning the research*

In designing the study, I considered that I needed to establish a reference point against which to compare the assessment and feedback data that were found in the e-portfolio record in order to allow conclusions to be drawn regarding the WBA system’s fitness for purpose. Consequently, the first phase of the study involved conducting a narrative review of literature in order to establish the stated purpose (or purposes) of workplace-based assessment according to literature that included official documents and relevant academic publications. A second important function of this review was to establish what had previously been found to be effective in the implementation of formative



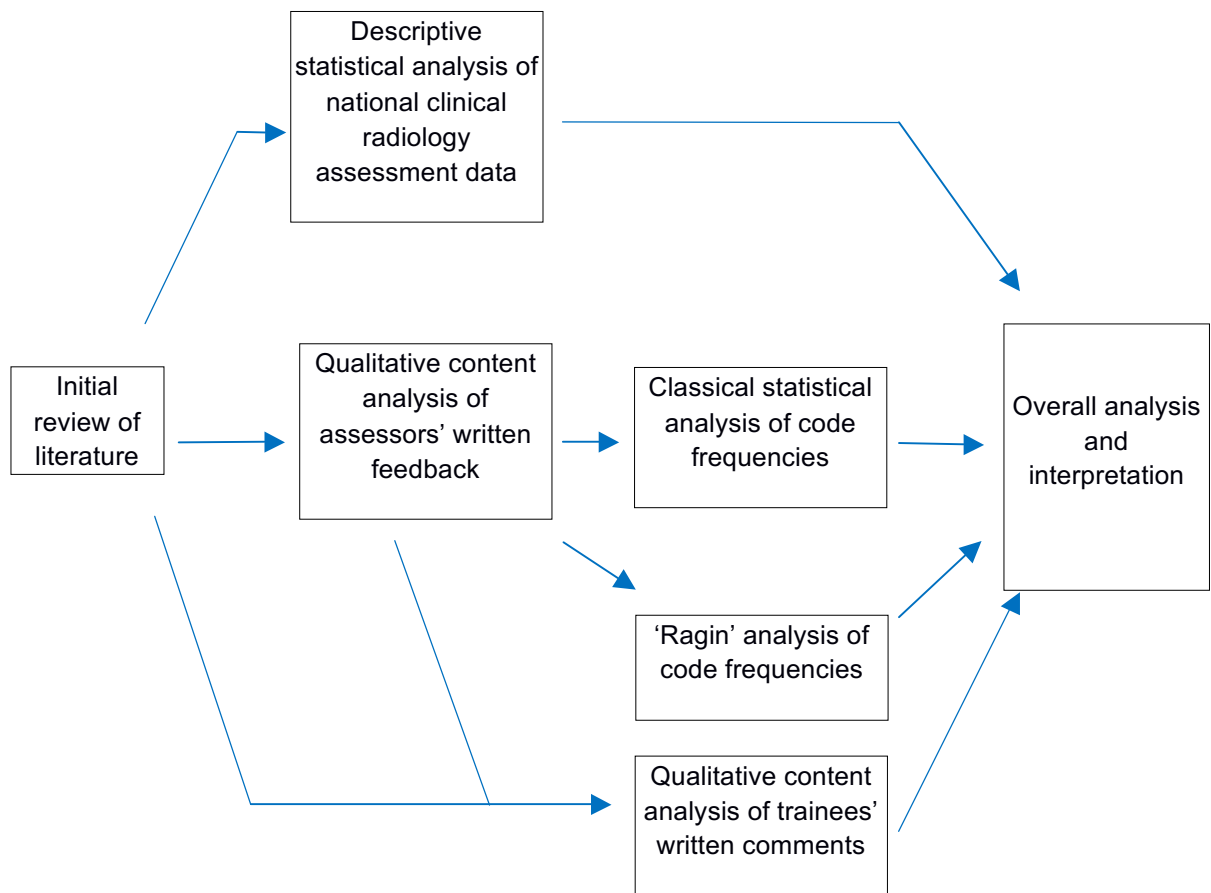
assessment and feedback interventions, in order to provide an objective set of criteria against which Rad-DOPS assessment and feedback data could be judged.

According to Baumeister and Leary, (1997) narrative literature reviews are useful 'when one is attempting to link together many studies on different topics, either for purposes of reinterpretation or interconnection. As such, narrative literature reviewing is a valuable theory-building technique' (p. 312). They contrast this with meta-analysis which, in their view, is aimed at supporting or refuting a clearly stated hypothesis, and depends on the studies which constitute it being focused on the same (or a very similar) hypothesis and exhibiting a large degree of methodological consistency.

Whilst the purpose of the literature review in my research stopped short of generating new theory, it was intended to perform the integrative function of drawing together research from different educational contexts and perspectives in order to create a coherent theoretical picture of effective formative assessment and feedback. To this end, the medical education literature was searched in order to build a picture of the current state of WBA assessment discourse within medical education. This was augmented by the inclusion of more purposively-sampled publications by notable authors in the field of formative assessment who have conducted their research primarily, although not exclusively, within the domains of school- and university-based education. This was done in order to place the concept of formative assessment in medicine within the broader formative assessment context.

The empirical aspect of the study involved a two-stranded approach: descriptive statistical analysis of a number of facets of the Rad-DOPS assessment data recorded in the e-portfolio; and qualitative content analysis of the written feedback comments of assessors and trainees. In addition to conventional statistical analysis of the coded assessor feedback statements, I chose an analytical approach described by Ragin (1987, 2000, 2008) to determine the necessary and sufficient conditions for specific types of feedback to be provided. The final stage of the study involved integrating the findings from each of the research components in order to draw conclusions and address the main research question. The activities involved in the study are displayed in Figure 4.1.

Figure 4.1 The methodological elements of the study design.



#### 4.5.3 Which paradigm?

In order to answer the research questions, the overall approach included both qualitative analysis of features of the assessors' written feedback statements, and quantitative analysis of the resulting coding frequencies and other aspects of the assessment and feedback process. Comparisons were then made with the stated purposes of workplace-based assessment and with the characteristics of effective assessment and feedback that were yielded by the review of literature. Thus the work straddled two paradigms: the interpretive paradigm, in seeking to understand and code the written assessor feedback statements and trainee responses that comprised the raw data, and draw conclusions about the nature of the written feedback; and the

positivistic paradigm, in using descriptive and inferential statistics (such as  $\chi^2$ ) to explore relationships between variables, and in using a theory-driven framework, largely constructed *a priori*, against which to compare the findings and to draw conclusions as to the likely effectiveness of the feedback process in Rad-DOPS assessment. These two paradigms were at times blended during the exploration and analysis processes, and in particular when using Ragin's (1987, 2000, 2008) approach to qualitative comparative analysis (QCA). Purists may balk at the blending of research paradigms, however some authors, such as Cohen *et al.* (2000), recommend a degree of paradigmatic flexibility, and caution against becoming 'paradigm-bound', in order to avoid 'stagnation and conservatism' (*ibid.*, p. 106).

## **4.6 Data collection**

### *4.6.1 Narrative review of literature*

As stated in Chapter 2, the review of literature that informed my research was not undertaken in order to identify a gap in current literature which could duly be explored. Neither was it conducted solely to answer a previously-determined question or hypothesis. Rather, the review was undertaken in keeping with the traditions of a narrative review, instead of the systematic review methodology so often preferred in medical and scientific circles.

This is not to say that the review was unstructured – the approach to seeking out relevant medical education literature in particular followed a clearly defined process in order to ensure that there were no important omissions. Relevant databases – Pubmed, Ovid Medline and Embase – were searched using the following keywords: feedback, assessment, workplace-based assessment, WBA, WPBA, evaluation, formative, medical education, clinical education, radiology. The search was initially limited to the years 2000 to 2012 – this represented the period from ten years prior to the launch of WBA in clinical radiology (which occurred in 2010) to the time of the literature review being undertaken. This window was chosen as authors such as Augustine *et al.* (2010) have described the ten-year period prior to 2010 as encompassing the transition towards outcome-based education in medicine, and the accompanying introduction of WBA across the medical specialties in the UK. Articles published prior to 2000 were

subsequently included if they had been cited in the reference lists of those that were returned in the search and found to be particularly relevant. These included 'seminal' articles such as a frequently-cited paper by Ende (1983), amongst others.

As a rule, articles were excluded if they did not focus on feedback or workplace-based assessment in the context of postgraduate medical education. However, due to the recent introduction of WBA into undergraduate clinical placements, some articles from undergraduate medical education were deemed to be relevant to the discussion of the postgraduate context and were duly included. Articles from the US medical education literature were included despite the differences between the postgraduate training process in the US and the UK. This was because WBA in its current form is widely recognised as having originated in the US. However, I emphasised the UK-based literature within my review in recognition of the focus of my study, which was concerned with clinical radiology training in the UK. Thus, I drew on the US-based literature only when it seemed particularly relevant, or seemed to offer an alternative point of view, or where there was a dearth of UK based literature on some aspect of WBA and feedback. The reference lists of all the selected publications were also hand-searched for relevant articles, and monthly updates provided by the library of the Royal College of Physicians were used to update the literature that had been retrieved initially. Given the role of official bodies in the introduction of WBA to clinical radiology, the publications issued by these organisations in relation to WBA were purposively sought out.

The review of the broader educational literature was less formally structured, and was guided by recommendations made by my supervisor as well as my own awareness, developed over 20 years working in education, of the leading authors on the subject of formative assessment and feedback.

In summary, the literature that informed the review consisted of empirical research articles, opinion pieces, and official documents relating to assessment and feedback in the postgraduate medical and radiological education context, as well as literature that encapsulated the expert opinion and empirical research findings of leading authors from the world of classroom-based education in the school and university sectors.

#### *4.6.2 The nature of the Rad-DOPS workplace-based assessment and feedback data*

The empirical data at the centre of this research were drawn from the national e-portfolio record for UK clinical radiology trainees. While it was not possible to ascertain whether every clinical radiology trainee was using the e-portfolio at the time of conducting the research, it was the case that its use had been mandated by the Royal College of Radiologists since 2010, and so it is likely that the majority of trainees in clinical radiology training in the UK – and particularly those in the early stages of training – were using the e-portfolio to record their workplace-based assessments. Certainly, the number of trainees who had logged assessments in the e-portfolio was large: 595 trainees recorded one or more Rad-DOPS assessments in the first year of the programme, with even greater numbers recording assessments in the second and third years (N=1028 and N=974 respectively). Thus, the data set afforded the opportunity of yielding robust, generalisable, empirical evidence about the workplace-based assessment and feedback system in clinical radiology nationally.

The complete data set comprised information on a number of aspects of the Rad-DOPS assessment process. These aspects included: the number of assessments recorded by individual trainees; the time within a trainee's attachment that the assessment was recorded; the number of assessments undertaken by individual assessors; the feedback field in which assessors enter their comments on the observed performance of trainees. It was also possible to see the scores awarded to trainees against the individual assessment criteria within each Rad-DOPS assessment – these scores were allocated on a scale of 1-6, with 1 equating to a judgement of 'well below expectations for stage of training' for a particular assessment criterion, and 6 equating to 'well above expectation for stage of training'. Finally, the training grade of individual trainees was also apparent from the data.

To comply with ethical requirements, the data were anonymised with respect to the trainees and the assessors prior to being passed to me. This was appropriate from an ethical perspective, but anonymisation introduced a degree of limitation into the study which impacted on the selection of research methods.

#### *4.6.3 Preparing for analysis of assessors' written feedback*

The main data at the heart of the study were the written feedback statements provided to clinical radiology trainees in the course of their 'radiology direct observation of procedural skills' (Rad-DOPS) assessments. In seeking to draw conclusions about the potential usefulness of assessors' written feedback, I considered that a judgement about its quality would have to be made. The question of how to make such a judgement was a complex one, and several approaches were considered.

Inductive thematic analysis of the assessors' written comments was considered on the grounds that it would offer an approach to analysis that could be applied to relatively large volumes of data, and remain flexible enough to construct a coding framework that reflected the content of assessors' feedback comments (see, for example, Braun and Clarke, (2006) on the affordances and limitations of a flexible approach to thematic analysis). This approach was used to inspect the data initially, and a number of initial themes were identified (see Appendix 4 for examples of the initial codes that were constructed). However, in order to be able to make judgements about feedback quality, it was necessary to take an approach to analysis that was also theory-driven in the first instance, rather than being purely data-driven. This was important because in seeking to come to a judgement about the quality of written feedback, a comparison with some sort of standard would be necessary. A theory-driven approach to analysis would therefore allow a framework to be derived against which the extant Rad-DOPS feedback could be compared. Moreover, in the interests of increasing transparency and limiting bias, taking a theory-driven approach would help me to recognise any latent prejudices and preconceptions about what the characteristics of the feedback should or might be. I could then counter the tendency of these notions to exert an undeclared and hidden influence on an inductive, data-driven coding process. Thus the theory summarised in the review of literature could be reflected in the coding framework that was constructed. Accordingly, the coding framework was initially theory-driven, but then modified inductively through immersion in, and familiarity with, the data. The chosen approach was therefore content analysis of assessors' written feedback statements. The particular content analysis approach employed in this study is discussed in section 4.7.

#### 4.6.4 Sampling

The Rad-DOPS assessment data from 2010-11 contained 4798 assessments. In seeking to reduce the data while retaining the integrity of complete assessments, I chose a randomised sampling approach. In the first instance I selected 500 assessments, which represented an approximate 10% sample of the population. I then chose a second sample of 500 assessments and compared my analysis of the two statistically to determine the representativeness of the original sample.

Having found good agreement between the analysis of these samples, I subsequently selected 500 assessments from 2011-12 and 2012-13 for analysis. It was interesting to note that in Vivekananda-Schmidt *et al.*'s (2013) study of written multisource feedback comments provided to non-trainee doctors, the researchers also chose a sample size of 500 assessments. In their case, however, the sample represented less than 5% of the total population of 11,483 assessment forms. In addition, there was no attempt in their study to demonstrate the representativeness of the sample.

### 4.7 Data analysis

#### 4.7.1 Descriptive statistical analysis

Descriptive statistical analysis was employed in order to generate a number of metrics that were likely to be indicative of the functioning of the WPBA system. These metrics, along with the rationale for calculating them, are set out below.

##### *Assessment scores*

In Rad-DOPS assessments, trainees are awarded a score from 1-6 for each criterion addressed within the assessment. Each of the scores is qualified according to the assessor's 'expectation for stage of training'. For example, for the criterion named 'Technical ability', a trainee may be rated anywhere on a scale which ranges from 'well below expectation for stage of training', which is equivalent to 1/6, to 'well above expectation for stage of training', which is equivalent to 6/6 (see figure 4.2). No overall score is awarded. Instead, assessors are asked to provide an overall qualitative

judgement about the trainee. This overall competence rating is phrased in terms of the trainee’s readiness for independent practice<sup>5</sup> (see figure 4.3). As can be seen in table 4.1, there was very little consistency between the numerical scores awarded to trainees who had the same overall rating. Therefore, to provide a summary of the trainees’ scores in their individual assessments, I calculated an average (modal) score for each assessment for each trainee.

Figure 4.2 The rating scale and the first six criteria on the Rad-DOPS assessment form.

	<i>1. Well below expectation for stage of training</i>	<i>2. Below expectation for stage of training</i>	<i>3. Borderline for stage of training</i>	<i>4. Meets expectation for stage of training</i>	<i>5. Above expectation for stage of training</i>	<i>6. Well above expectation for stage of training</i>	<i>Unable to comment*</i>
<b>1. Demonstrates understanding of indications, relevant anatomy and technique</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>2. Explains procedure/risks to patient, obtains/confirms informed consent where appropriate</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>3. Uses appropriate analgesia or safe sedation/drugs</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>4. Usage of equipment</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>5. Infection prevention and control</b>	<input type="checkbox"/> Unsatisfactory		<input type="checkbox"/> Satisfactory		<input type="checkbox"/> Not applicable		
<b>6. Technical ability</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 4.3 Overall competence ratings on the Rad-DOPS assessment form.

	<b>Rating</b>
<input type="checkbox"/>	Trainee requires additional support and supervision
<input type="checkbox"/>	Trainee requires direct supervision (performed at level expected during Core training)
<input type="checkbox"/>	Trainee requires minimal/indirect supervision (performed at the level expected on completion of Core Training)
<input type="checkbox"/>	Trainee requires very little/no senior input and able to practise independently (performed at level expected during Higher Training)

<sup>5</sup> This was a deliberate decision taken by the RCR in the light of work done by Crossley *et al.* (2011). This work demonstrated that qualitative overall competence ratings, phrased in terms of so called ‘anchor’ statements – phrases that referenced ‘increasing independence’ or other constructs aligned to overall professional development as a doctor – were more reliable than previously-used summary scores.



The median and mode are typically preferred to the mean when calculating the central tendency of ordinal data, such as the 1-6 scoring system employed in these assessments. Of these options, the mode was felt to provide the best indication of the weight of the assessors' scoring: it was unlikely to be the case that assessors' scores on individual assessments were normally distributed and, as Manikandan (2011) observes, the mode is typically used to indicate the most frequently occurring value(s) in skewed or non-normally distributed data. Furthermore, as Manikandan (*ibid.*) also notes, the mode is the only measure of central tendency that can be used with nominal data. The ordinal scores of 1-6 actually relate to a nominal scale (see figure 4.2) in which each value corresponds to a qualitative statement about how the trainee has performed against each assessment criterion (e.g. 1 = 'well below expectation for stage of training'; 2 = 'below expectation for stage of training'; 3 = 'borderline for stage of training'...). Thus I felt that the mode was the most valid indicator of an assessor's overall pattern of scoring for a particular assessment.

#### *Frequencies of assessments recorded by trainees*

In order to provide an indication of the degree to which trainees were following curriculum guidance as to the number of Rad-DOPS assessments that should be undertaken, the frequencies of Rad-DOPS assessments recorded by all trainees in 2010-11, 2011-12 and 2012-13 were calculated in Excel. From these numbers, the mean, median and modal numbers of assessments recorded by trainees in each training grade (ST1, ST2, ST3...ST6) were calculated<sup>6</sup>. It was apparent that some trainees were recording substantially lower, and substantially higher, numbers of assessments than recommended, and so the range and standard deviation was also calculated in order to examine the extremes of assessment activity.

---

<sup>6</sup> These calculations were done in order to determine the average number of assessments recorded by trainees in different grades – this should not be confused with the previously-mentioned modal calculations, which were conducted at the level of individual assessments in order to give an average score for each individual assessment.

Trainee id	Training grade	1. Demonstrates understanding of indications, relevant anatomy and technique:	2. Explains procedure/risks to patient, obtains/confirms informed consent where appropriate:	3. Uses appropriate analgesia or safe sedation/drugs:	4. Usage of equipment:	5. Infection prevention and control:	6. Technical ability:	7. Seeks help if appropriate:	8. Minimises use of ionising radiation:	9. Communication with patients/staff:	10. Quality of diagnostic images:	11. Judgement/Insight:	12. Quality of report of procedure:	13. Overall Competence – please select one of the following options which best describes the trainee:	Modal score
mi653oy	ST3	5	5	5	5	Satisfactory	5	5	U/C	6	U/C	6	6	Trainee requires very little/no senior input and able to practise independently	5
un571i	ST1	4	4	U/C	5	Satisfactory	5	5	U/C	5	5	5	5	Trainee requires direct supervision	5
ma418 l	ST5	5	4	4	4	Satisfactory	4	4	4	5	5	4	5	Trainee requires very little/no senior input and able to practise independently	4
ar939Pe	ST2	5	5	5	5	Satisfactory	5	5	5	5	5	5	5	Trainee requires minimal/indirect supervision	5
mi662ar	ST1	4	4	U/C	4	Satisfactory	4	4	4	5	4	4	4	Trainee requires direct supervision	4
ei805oh	ST1	5	4	4	5	Satisfactory	4	4	4	4	5	5	5	Trainee requires minimal/indirect supervision	4
ar087 B	ST5	4	4	4	4	Satisfactory	4	4	4	4	4	4	4	Trainee requires very little/no senior input and able to practise independently	4
an593ng	ST2	4	4	U/C	4	Satisfactory	4	5	U/C	5	U/C	5	U/C	Trainee requires direct supervision	4
ho333 S	ST1	6	5	U/C	5	Satisfactory	6	5	4	4	5	5	6	Trainee requires minimal/indirect supervision	5

Table 4.1 Summary of assessment data for a random sample of trainees, including scores for individual assessment criteria, the overall competence rating and the modal score that was calculated for each assessment. Shaded rows in particular highlight how trainees with the same modal score (4) may have very different overall competence ratings, which may not be explained by differences in training grade.

### *Frequencies of assessments conducted by assessors*

Senior doctors who attended the WBA assessor training workshops that I formerly delivered for the RCR commonly complained about the lack of time available for them to undertake assessments with trainees. As discussed in Chapter 2, even professional classroom teachers need significant staff development input and the opportunity for repetition and reflection in order to appreciate and fully embed formative assessment in their daily classroom practice. I postulated that medical professionals may lack these opportunities to engage in frequent formative assessment, which could have implications for their skill development, or skill retention, in conducting this type of assessment. Consequently, I was interested in the numbers of assessments conducted by individual assessors. Pivot tables were used in Excel to extract and aggregate this data from within the large data sets provided by the RCR.

### *Timing of the assessments*

As previously established in the review of literature, formative assessment should come at a point in time when the feedback is potentially of use to the learner. Thus, an assessment is not simply formative because the assessor providing feedback intends it to be. For example, in Sinclair & Cleland's (2007) research into medical student engagement with formative feedback, they found that less than half of the student cohort collected the written feedback provided by faculty. However, whilst the feedback was described as being formative, it was given at the end of the course of study, immediately prior to the holiday period, and was linked to an assignment that bore little resemblance to the students' other coursework assignments. It is understandable, then, that students may not have perceived the value in the written comments given at the end of the module when there is little or no scope for improvement. Similarly, feedback that is given to radiology trainees at, or close to the end of their clinical placement is unlikely to lead to useful learning, regardless of its quality, due to the limited opportunities for trainees to respond. It is important to acknowledge that developmental feedback provided at an earlier point in a learning process may still not be formative, as it may not be of sufficient quality to support development, or may be rejected by the learner for a number of different reasons. However, feedback that comes too late in the learning process cannot achieve its intended aim, regardless of its quality. Therefore, the timing of the assessments recorded by trainees in clinical radiology across the UK

was of interest, and was considered to be an important metric for the functioning of the formative WBA system.

The raw data extracted from the e-portfolio by the Royal College of Radiologists was used to calculate when the Rad-DOPS assessments were recorded within each training attachment. This was done in Microsoft Excel, and patterns of assessments which were recorded at different times in trainees' attachments were displayed graphically.

#### *4.7.2 Content analysis of assessors' written feedback*

Content analysis is a term that has been used to describe a number of different approaches to data analysis. According to authors such as Graneheim and Lundman (2004) it is an approach that was originally devised in order to conduct analysis of the lexical features of mass media and public speeches. At its most straightforward, it consists of counting the frequencies of key words and phrases, and reporting these descriptively. This approach to content analysis tends to deal with what has been termed by authors such as Downe-Wamboldt (1992) and Kondracki *et al.* (2002) the *manifest* content of the text. In other words, the text is addressed at the semantic level, and deals with what Graneheim and Lundman (2004) term the 'visible, obvious components' (p. 106). Accordingly, the analysis of the text tends to focus on the presence or absence of particular words and phrases, and these can either be coded by researchers with the assistance of a coding scheme or guide, or, given the purely lexical nature of the exercise, the task can be undertaken by purpose-built analytical software. Presentation of results is often done through simple frequency tables, however advances in technology have allowed researchers to take more innovative approaches to their presentation and analysis. For example, Gill and Griffin (2010) conducted a comparative content analysis of historic and present-day General Medical Council (GMC) documents using tag clouds (also known as word clouds) and used these as a way of exploring how the concept of medical professionalism has evolved over time. More conventionally, once the framework has been applied and coding frequencies generated, these frequencies may be subjected to further analysis, for example by looking for statistical relationships between codes.

However innovative the style of reporting results, this word-counting approach to content analysis can appear to be somewhat quantitative in nature, and at times this is reflected in the research methods literature. A number of authors on content analysis identify it as a quantitative method, and tend to be concerned with aspects of the research process that are more frequently encountered in positivistic approaches to research. For example, Krippendorff (2004) emphasises replicability, which is a measure of rigour more likely to be encountered in the natural sciences than the social sciences. Cohen *et al.* (2007) similarly highlight the potential for verification through reanalysis, and draw attention to the systematic, explicit and transparent rules for analysis as a strength of the approach. However, replicability becomes more challenging when one takes the view that content analysis can be applied not just to manifest content, but also what is often described (e.g. by Graneheim and Lundman, 2004) as *latent* content. Latency refers to the idea that words and phrases do not necessarily encode information in obvious and straightforward ways. Therefore, an approach to content analysis that recognises the subjective and contingent nature of language, and employs a more interpretivistic approach to coding the data, is at heart a *qualitative* research method, even if the codes, once assigned, are analysed through statistical or other quantitative means.

The content analysis aspect of the study, therefore, was essentially qualitative in nature. However, in exploring and analysing the data, the research embraced aspects of both the qualitative and quantitative research traditions. This is in keeping with Weber's (1990) assertion that the inclusion of both qualitative and quantitative analyses is a feature of the highest quality content analysis. Furthermore, the quantitative exploration of the coded data, rather than making the research 'more quantitative', may actually support a more qualitative approach to analysis by allowing the researcher to construct relationships between aspects of the data that are not apparent from simple frequency counts or statistical tests of significance. One such approach is that proposed by Ragin (1987, 2000, 2008), which is described in more detail below.

In their review of content analysis concepts and procedures, Graneheim and Lundman (2004) point out that, in fact, both manifest and latent content require interpretation, even if the interpretations 'vary in depth and level of abstraction' (p. 106), and so content analysis is in essence an interpretive exercise, however straightforward an approach the researcher appears to take. These same authors indicate that this

position – that all linguistic analysis is to some degree interpretative – has implications for issues of trustworthiness in data analysis, and these issues are explored below. It also has implications for particular approaches to analysis, such as computerised data analysis, in that content which to the casual observer seems easily recognisable and easily coded is almost impermeable to data analysis software. For example, computer aided qualitative data analysis software (CAQDAS) can identify the occurrence of the word 'competence' within written feedback statements, but cannot infer anything about the sense in which it is being used. Sentences such as 'has achieved competence in this procedure', 'struggles to demonstrate competence' or 'is progressing towards competence' are undecipherable and indistinguishable using automated, word-recognition approaches, emphasising the essential role of the researcher in interpreting even these apparently straightforward uses of language. That is not to say that the meaning of some of these phrases is therefore highly complex or obscure – most readers would agree that the phrase 'has achieved competence in this procedure' carries an unambiguous, positive meaning, which is that the required proficiency in the procedure has been attained. The meaning of other phrases is more nuanced. For example, should the phrase 'is progressing towards competence' be regarded as positive, because the trainee concerned is making progress? Or should it perhaps be regarded as negative, because they have clearly not yet achieved competence? The meaning is likely to depend on, amongst other aspects, the assessor's view of competence, the assessor's view of the stage at which trainees of similar experience and typical ability would be, and any objective landmarks or benchmarks indicated in the curriculum. The true meaning may or may not be revealed by qualifying statements in the surrounding text, and the role of the researcher in all of this is to study the entire feedback statement and arrive at a view of what is actually being communicated by the assessor. It is also the role of the researcher to share their own decision-making process in regard to these more problematic judgements with the reader, and so in section 4.8.2 I have provided examples of 'grey cases' in which the application of codes was particularly challenging.

#### *4.7.3 The content analysis research process*

According to Cohen *et al.* (2007, p. 476), the content analysis process 'takes texts and analyses, reduces and interrogates them into summary form through the use of both

pre-existing categories and emergent themes in order to generate or test a theory.' As such, it was suitable for this study, given the largely theory driven nature of the coding framework (used to generate pre-existing categories), and the potential to modify the framework in response to the researcher's interpretation of the data (emergent themes). However, the three-step approach articulated by Cohen *et al.* (*ibid.*) does not quite do justice to the research process, which is better described by the work of Graneheim and Lundman (2004). Their work does not identify a prescriptive list of steps, but instead identifies important concepts and decisions that should feature in the outworking of a robust content analysis approach. Their ideas informed the decisions made at a number of points in this research, and these are described and discussed below.

#### *4.7.4 Constructing an initial framework for coding assessor comments*

The first step was familiarisation with the data, which was done by reading and re-reading a random sample of 500 of the 4798 feedback statements written by Rad-DOPS assessors in the training year 2010-2011. Initial codes were developed in order to identify broad conceptual categories, which reflected both content-related themes (what the assessors said) and process-related themes (how they said it). This was done both deductively, using the important features of feedback as derived from the literature, and inductively, using the observations and 'working codes' from the data itself. The framework was also revised during the coding process as new aspects of the assessors' feedback came to light (see Appendix 4 for an illustration of how the coding framework evolved over time).

It was clear from this initial phase that the codes would be likely to be overlapping. For example, comments on the observed performance in an individual procedure could be either positive or negative and, similarly, comments on overall progress within the training attachment could also be positive or negative. Developmental comments, by their very nature, tended to imply deficiencies in the observed performance, or at the very least a need for further development, and so these would generally also be coded as negative. However, there were some developmental suggestions which the assessor took some trouble to indicate were purely comments about what the next stage of normal development would tend to involve, and so these (although few in number) were not coded as being negative. The use of overlapping codes contravenes

the assertion by Miles and Huberman (1994) that codes should be distinct and mutually exclusive. However, the nature of the feedback was such that discrete codes did not always provide for unique coding of any one comment, and this complexity was addressed by introducing an element of combination. For example, a feedback comment could be coded as being 'positive; linked to assessment criteria; general behavioural,' or 'negative; general behavioural; unspecific recommendation.'

#### 4.7.5 Analysing published frameworks

The review of literature had revealed a number of validated coding frameworks that were potentially useful in approaching the data in this study. These were analysed in order to confirm whether or not they reflected the aspects of feedback that were identified in the literature as being educationally valuable, as well as whether they were flexible enough to be adapted in order to adequately describe the data in my research (see Chapter 3 for discussion of these frameworks). The approach that was best aligned to my own research, and which offered a robust starting point for the development of my own framework, was that described by Canavan *et al.* (2010). Modifications were required to make the framework more intuitive: whilst the researchers in Canavan *et al.*'s study (*ibid.*) undertook several rounds of review to develop a shared understanding of the codes being applied, some of the terminology they used is still not immediately clear to others. For example, the meaning of the 'non-behavioural/global assessment' code was not immediately apparent. In practice, it was used to mean a reference to the personality or other personal attributes of the trainee that were not linked to observed behaviours. Examples cited by Canavan *et al.* (2010) included "She is competent, she is caring, and she is compassionate" (*ibid.*, p. S107). Firstly, these comments most likely *do* relate to one or more of the domains on the multi-source feedback (MSF) assessment form. This cannot be ascertained directly, as their article does not include an example of the MSF assessment tool, but a 'consideration for patients' domain (or similar variant) is often featured in MSF exercises in healthcare. Therefore, even if the authors considered these comments not to be behavioural, there is a good chance that they did align to one or more of the domains being assessed in the MSF. In addition, their example includes a remark made about competence, which albeit general, would seem to be a comment on observed behaviour, rather than an aspect of the trainee's character or personality. In



fact, when the authors refer to global comments in their results and discussion, they tend to identify these types of comment as 'self-directed'. This resonates with Kluger and DeNisi's (1996) consideration of feedback which causes recipients to direct their attention towards the 'ultimate goals of the self' (p. 262) as opposed to lower-level (yet important) processes such as *task motivation* or the detailed focus on specific components of the task – so-called *task learning* processes. In fact, Kluger and DeNisi (1996) refer to feedback that relates to the person rather than the task as 'personal feedback' (p. 255), and this is the term that was adopted in my research.

On the other hand, the term 'global feedback' did feature in my coding framework. My initial interpretation of Canavan *et al.*'s (2010) 'global' code was that it referred to progress over a period of time, as opposed to a point in time. However, progress over time was more or less implied in Canavan *et al.*'s (2010) research, as the focus of their study was a multi-source feedback exercise. Such exercises are normally only conducted once a trainee has been in post for a period of time, and the expectation therefore is that raters' comments refer to the trainee's performance over the time they have been on placement. Hence, they did not use the code in this way. In the context of my research, the Rad-DOPS is intended to be a point-in-time assessment of a directly observed procedure, but initial inspection of the data showed that some assessors had made comments that were linked to the trainee's progress over the whole course of their placement. The 'global' code was therefore used in my study to refer to this type of comment on a trainee's overall progress.

The creation of the 'global' code was required to ensure that my coding framework was complete and was not systematically excluding certain types of comments from the analysis process. An additional pre-ordinate code – 'linked to assessment criteria' – was created for my study. This was done in order to establish the extent to which assessors invoked particular assessment criteria in their feedback, and was driven by the observation of Nicol and Macfarlane-Dick (2006), amongst others, that feedback providers and feedback recipients should have a shared understanding of what the feedback is about. Reference to particular assessment criteria may provide a basis for shared understanding, and this code was created accordingly. Initial inspection of the data had also suggested that making reference to the published assessment criteria was a feature of the feedback provided by a number of assessors, and so the code was included in order to capture this aspect of the feedback statements.

There were codes in Canavan *et al.*'s (2010) framework that I deemed to be redundant for my purposes. For example, in my research there was no helpful distinction to be made between 'specific behaviour' and 'specific instance of behaviour'. There was also no requirement for the 'remark regarding lack of exposure' code – feedback in Rad-DOPS arises from an assessment of an observed clinical encounter, and so none of the assessors involved provided written feedback to the effect that they had not witnessed the trainee's performance. The 'hearsay' code also proved to be redundant, as assessors restricted their comments to those aspects of trainee performance that they had personally observed, and did not comment on second-hand information. Canavan *et al.*'s (2010) original coding framework is reproduced in figure 4.3.

Figure 4.3. Summary of the codes used by Canavan *et al.* (2010, p. S108) in their content analysis of written feedback provided to trainees in the context of a formative multi-source feedback (MSF) exercise.

Non-behavioural/global assessment
References to observee's behaviour
- General behaviour
- Specific behaviour
- Specific instance of behaviour
Statements indicating valence of feedback
- Positive
- Negative
Comments offering strategy for improvement
- General strategy for improvement
- Specific behavioural strategy for improvement
Remarks on ability to rate feedback recipient
- Remark regarding limited/lack of exposure
- Hearsay

The coding framework developed by me for use in my research, along with the rules for applying each code and examples of feedback comments to which the codes were applied, are shown in figure 4.4.

Figure 4.4 Coding framework applied to assessors' feedback comments, with a description of the criteria for assigning each code to a particular comment.

<b>Code</b>	<b>Criteria for applying code to assessors' comments</b>
Valency	
Positive	The comment was clearly intended to be positive in nature
Negative	The comment was negative in nature. This included any suggestion that improvement would be necessary, however constructively expressed.
Performance	
General comment on observed performance	The assessor commented on an aspect of the trainee's performance in a manner that may have required further explanation
Specific comment on observed performance	The assessor made a comment that was sufficiently clear as to make it unlikely that the trainee would have needed further explanation
Linked to assessment criteria	The comment clearly invoked one or more of the assessment criteria on the Rad-DOPS form
Descriptive	The comment is limited to a description of the procedure undertaken by the trainee, and lacks any judgement of their performance or suggestions for further development
Developmental	
Specific recommendation	The assessor made a suggestion for improvement that is unlikely to need further clarification
Unspecific recommendation	The assessor made a suggestion for improvement that was unclear or ambiguous
Personal	The comment referred to some aspect of the trainee's personality or personal qualities
Global assessment	The comment referred to the trainee's overall progress within the training post
Assumed improvement	The assessor made a comment to the effect that time, or experience, or continued practice would necessarily bring about improvement
Absent	The assessor failed to provide a comment

#### *4.7.6 Applying the coding framework*

The initial coding framework was applied to a random sample of 500 Rad-DOPS assessments from the training year 2010-11, which were extracted, prepared and exported to MAXQDA for analysis. In applying the initial framework, assessors' written feedback was analysed for both manifest and latent content, with adaptations being made to the pre-ordinate coding framework as necessary. The initial coding process involved iterative coding 'sweeps' through the same set of 500 feedback statements, to ensure that the coding framework covered all of the relevant aspects of the assessors' comments, and to develop clarity about the application of the codes. Rules were developed about the scope of each code, and these are described in Figure 4.4. Reflecting the fact that both manifest and latent content require interpretation, there were 'grey' cases in which the application of codes was not straightforward, and examples of these are described later in this chapter in the interest of transparency and, as I argue below, the demonstration of validity.

#### *4.7.7 Defining the unit of analysis*

There is a great deal of variation in the content analysis literature as to what researchers consider to be their unit of analysis. Some units of analysis are taken to be objects present in the field – for example, Mertens (1998) identified individuals, programmes, organisations or clinics as potential units of analysis. According to Graneheim and Lundman (2004), other researchers have defined their units of analysis as:

- particular artefacts that are constructed through the research process itself, for example, entire interviews and diaries;
- abstracted aspects of these artefacts, such as phrases that have been coded in a particular way, or elements of the data that have somehow been selected for particular consideration;
- or all of the individual words and phrases that comprise the research material.

Graneheim and Lundman (*ibid.*) take the view that entire interviews or observational protocols should comprise the unit of analysis, as long as they are 'large enough to be considered a whole, and small enough to be possible to keep in mind as a context for the meaning unit' (p. 106). For my study, the unit of analysis was taken to be a written feedback statement, and a feedback statement was taken to be the entire block of written feedback provided by a single assessor on an individual Rad-DOPS assessment form. In the interests of consistency and clarity, from this point onwards the term 'feedback statement' will be taken to indicate this unit of analysis.

#### 4.7.8 Defining the meaning units

In exploring methods for breaking down the units of analysis further, I encountered the concept of 'meaning units' – subsets of units of analysis that are more manageable than the overall unit of analysis, but which contain enough meaning to be usefully analysed in their own right. For some authors, such as Elo *et al.* (2014), the notion of 'unit of analysis' appears to overlap completely with the concept of meaning units. Other authors distinguish between these two ideas with the latter, meaning units, forming composite parts of larger units of analysis. For example Brann and Mattson (2004), used the synonymous term 'coding unit' and defined this as 'a basic unit of text that consisted of a complete idea' (p. 156). Consequently, coding units, or meaning units, in their study could consist of single words, phrases or sentences. Graneheim and Lundman (2004) define meaning units as 'words, sentences and paragraphs containing aspects related to each other through their content and context'. These latter two definitions captured the meaning units that were being studied in this research, and had much in common with the approach taken by Canavan *et al.* (2010). In defining their meaning units though, Canavan *et al.* (*ibid.*) chose to do so linguistically – they described their approach as parsing complex feedback statements into phrases. This use of the specialist linguistic term 'parsing' is potentially misleading, as the research team did not obey the conventions of linguistic analysis when doing so – rather, their examples made it clear that what they were actually doing was separating off phrases, or often one or more sentences, in which assessors were commenting on a particular aspect of a trainee's performance. Regardless of the linguistic exactitude of the use of 'parsing', their approach was similar to mine, in that the meaning units in my research were taken to be individual feedback comments,

which were in turn defined as any phrase or sentence that contained feedback focused around a particular aspect of the trainee's personality or performance. Therefore, in the context of my research, a feedback statement (the unit of analysis) could be comprised of one or more feedback comments (the meaning units) (see figure 4.5).

Figure 4.5. Illustration of how codes were applied to feedback comments (the meaning units) within feedback statements (the units of analysis).

**The following feedback statement (unit of analysis)...**

'A very competent examination of the abdomen as well as the soft tissues and muscle. Good use of machine settings, using several transducers to obtain maximum information in order to answer the clinical questions asked by the referring clinician. More scanning of patients with complicated clinical pictures would help to adapt scanning technique'.

**...was divided into the following feedback comments (meaning units), and was coded as shown in italics:**

'A very competent examination of the abdomen as well as the soft tissues and muscle.' – *Positive valency; Linked to assessment criteria; General behavioural comment.*

'Good use of machine settings, using several transducers to obtain maximum information in order to answer the clinical questions asked by the referring clinician.' – *Positive valency; Linked to assessment criteria; Specific behavioural comment.*

'More scanning of patients with complicated clinical pictures would help to adapt scanning technique' – *Negative valency; Unspecific recommendation; Assumed improvement.*

Thus, data were analysed at the level of these meaning units, or 'feedback comments', but given my interest in the quality of the assessors' feedback provision within each assessment, coding frequencies were reported at the level of the units of analysis i.e. the feedback statement (see Chapter 5, tables 5.3 and 5.4).

#### 4.7.9 Reducing the data

Data reduction is often necessary in order to make large volumes of raw data manageable, as well as facilitating exploration of the data on a 'higher' logical or analytical level. Graneheim and Lundman (2004) prefer the term 'condensation' (p.

106), as they argue that the term 'reduction' says little of the quality of what remains after the reduction process, whereas condensation, they feel, conveys the sense that the 'core' of the data (p. 106) has been preserved. They give the example of the following meaning unit, 'There is a curious feeling in the head in some way, empty in some way,' being condensed to 'curious feeling of emptiness in the head'. This condensed meaning unit was then coded as 'emptiness in the head'. This condensation approach to data reduction was not particularly feasible in my study, due to the nature of the data being analysed; they were not comprised of discursive or narrative accounts of assessors' or trainees' experiences. Rather, the data mainly comprised relatively short statements that were written, rather than spoken, and so the meaning of individual units was often clear without the need to reduce, condense or in some way edit out the artefacts of normal speech patterns. However, given that the data pool consisted of some 4978 to 8013 feedback statements, depending on the year of the programme being analysed, it was clear that some form of data reduction would be required in order to make the research manageable.

In seeking to preserve the entirety of the data set from each of the three training years for analysis, an approach to content analysis of large data sets was sought. At first, a technological approach to data reduction and analysis was employed. The search function within the MAXQDA™ analysis software was used to return feedback that contained words or phrases that might indicate the presence of particular types of feedback. The approach was piloted with terms such as 'competence', 'progress' and 'confidence' (and associated variations), however it was found that there were no search terms that would allow even apparently simple feedback phrases to be identified reliably. For example, the search term 'competence' (and its variations) returned feedback comments of divergent, or even opposing meanings. Phrases that were returned for three different trainees were as follows, '[Name] shows a competency which is beyond her level of training', 'Increase practice to consolidate competencies', and 'should further read up about the procedure to gain competence'. This approach to data reduction was therefore rejected in favour of a sampling approach to the data.

In employing a sampling approach to data reduction, I was conscious of the fact that I was adopting a method that is more commonly encountered in quantitative studies, in which the considerations typically include a concern for representativeness and generalisability. Whilst representativeness and generalisability were not preoccupations

in the design of this study, I was nonetheless aware that having access to a complete, national-scale data set created the possibility of making generalisable observations about the quality of feedback being provided nationally within one particular medical specialty. It seemed that a probabilistic approach to sampling would afford the opportunity to reduce the data in a way that preserved entire assessor comments while retaining the potential to make generalisable statements about the feedback at the end of the research.

#### *4.7.10 Problems with content analysis*

One of the features often cited as a strength of content analysis – its adherence to a transparent and consistent coding framework – may at times present significant limitations as a result of inflexibility. Cohen *et al.* (2007) recognise this risk, and point out that while the initial coding framework is 'usually derived from theoretical constructs or areas of interest devised in advance of the analysis' (p. 475) – an approach that they term 'pre-ordinate categorization' – such frameworks can be modified throughout the research process in accordance with the features of the empirical data. This was the approach taken in my study and accordingly, as previously mentioned, the framework was revised on a number of occasions rather than being followed rigidly from the outset.

Another risk, linked in part to the development of pre-ordinate categories but which could exist in a study that only used an inductively generated coding framework, is that the coding framework fails to take account of all relevant features of the data, resulting in systematic exclusion of relevant aspects of the data from the study. An awareness of this possibility allowed me to check and re-check during the iterative coding process that there were no comments that were simply being systematically overlooked or ignored as a result of not fitting into one or more of the categories present in the framework. Where new features of the data did emerge during the coding process, new codes were created accordingly.

Thirdly, in attempting to reduce and encode data, the researcher can inadvertently 'murder to dissect', in that they fragment the original whole to such an extent that the context and meaning are lost in the process. In my research, the nature of a lot of the feedback data was that assessors' overall comments were relatively brief compared



with, say, an in-depth interview transcript, and so fragmentation was less of a concern in my study. The risk in my study was that reporting the prevalence of certain types of *comment* throughout the sample would perhaps prevent the reader from developing a feeling for the holistic feedback *statements* provided by assessors. In order to address this, it was necessary to present a number of whole feedback statements in the results chapter to allow the reader to observe individual feedback comments in context and appreciate the nature of the raw empirical data. Similarly, it was necessary to present pairs of assessor comments and trainee responses to allow the reader to view the feedback exchanges which occurred, and weigh for themselves their dialogical nature.

## **4.8 Trustworthiness of the research**

### *4.8.1 Establishing rigour in qualitative research*

Despite the essentially interpretivistic nature of qualitative content analysis, qualitative researchers may often resort to concepts and approaches more normally found in positivistic, quantitative research in order to establish the trustworthiness of their findings. For example, the work of Prins *et al.* (2006), who looked at written feedback in general practitioner assessments, contains prominent references to reliability and seeks to establish this property quantitatively. Conversely, some researchers deliberately set aside these quantitative approaches and constructs, and instead seek to establish parameters for robust research that are different both categorically and practically. For example, Frambach *et al.* (2013) have suggested alternative concepts, which they term 'quality criteria' (p. 552), for addressing issues of rigour in qualitative, as opposed to quantitative, research. The latter, they argue, is concerned with *internal validity* (through power calculations, estimates of effect size, standardised treatments, controlled study designs etc.), *external validity* (through large sample size, representativeness of the sample, generalisability and predictive validity), *reliability* (through calculation of consistency coefficients such as kappa scores and Cronbach's alpha) and *objectivity* (including the removal of personal bias through blinded study designs, the quantification of results, and the production of value-free information). Whilst some of these ideas are compatible, or at least not entirely incompatible, with the philosophy and values of qualitative research, others are in direct opposition to the aspirations and tenets of the field. For example, the idea that personal bias *can* be

removed from the analysis of results is not only at odds with the underlying philosophy of qualitative research (which is interpretivistic in nature and therefore alert to the subjectivity inherent in the research process) but is also at odds with the *ambition* of a lot of qualitative research which, believing reality to be constructed and therefore inherently subjective and open to multiple interpretations, positively *embraces* subjectivity and seeks to understand the perspective of different individuals and groups, not least the researcher's own perspective, as part of the research process. As Cohen *et al.* (2013) observe:

Qualitative enquiry is not a neutral activity, and researchers are not neutral; they have their own values, biases and worldviews, and these are lenses through which they look at and interpret the...world. (p. 225).

This is in contrast to the positivist position, which often seeks to demonstrate that no such subjectivity exists.

The discourse in the research methods literature around rigour in qualitative research often seeks to establish parallel criteria to those previously described for quantitative research. For example, Frambach *et al.* (*ibid.*) offer *credibility* (or believability) as an alternative to internal validity, and state that it is established through approaches such as triangulation (of data sources, methods, researchers and theories), prolonged engagement (through longitudinal studies) and participant checking of data and interpretations. However, almost all of these suggestions are potentially problematic in themselves. For example, triangulation of methods is likely to result in an increase in complexity of the data rather than some sort of zeroing in on 'the truth' – the methods employed in qualitative research tend to be expansive rather than reductionist, and there is an appreciation in qualitative research that data are constructed rather than discovered or seen to spontaneously emerge.

The involvement of co-researchers risks introducing complexity for similar reasons, and Sandelowski (1993) argues strongly that it is an approach that fails to take account of the nature of qualitative research:

One of the most important threats to the...construct validity of qualitative projects is the assumption that validity rests on reliability. Investigators often claim that their findings are valid when, for example, they can show that...a panel of experts or persons other than the investigator coded information the

same way. What is embedded in these examples is the notion of reality as external, consensual, corroboratory and repeatable. What is being sought in these examples are coefficients of agreement or consensus on the nature of that reality. What is forgotten is that in the naturalistic/interpretive paradigm, reality is assumed to be multiple and constructed rather than singular and tangible (p. 2-3).

Consequently, any decision to involve additional investigators to demonstrate internal reliability of the analysis and interpretation of data must be considered carefully, and argued on a footing that is consistent with the qualitative paradigm. My own decisions regarding whether and how to involve a colleague are discussed in the following section (4.8.2).

Regarding external validity, Frambach *et al.* (2013) offer *transferability* as an alternative to traditional notions of generalisability. In doing so, they recommend 'thick description' (*ibid.*, p. 552) of the context, which allows others to judge if the context is applicable to them, and a clear explanation of the sampling strategy – 'typical case sampling or maximum-variation sampling' (*ibid.*, p. 552). Despite the change in terminology, the authors still emphasise the idea of being able to apply research findings in one context to other populations that were not the focus of the original research, as their recommendations are essentially aimed at allowing readers of the research to come to a judgement as to whether the findings of the research are applicable to the reader's particular setting. As with quantitative approaches to demonstrating reliability, transferability is an ambition that is not necessarily in keeping with the fundamental principles of qualitative research, and so it is not clear that Frambach *et al.*'s (*ibid.*) suggestions are particularly helpful here.

In considering the extent to which transferability was a useful concept for my research, my aim was primarily to establish the fitness for purpose of the WBA system within clinical radiology in the UK, and so my principle concern was to be able to argue that the findings were generalisable to UK clinical radiology trainee population. As indicated in the literature review, I was also interested in the degree to which my findings aligned with the findings of any similar WBA research, and as I have argued in chapters 1 and 2, the system in operation in clinical radiology shares many of the features of the WBA system that has been introduced across the medical specialties in the UK. Consequently, I expect that a number of the findings of my research would generalise to the broader UK postgraduate training context. However, I have not set out to

demonstrate this in any verifiable way, and so, in keeping with Frambach *et al.*'s (*ibid.*) suggestion, it is largely up to the reader to appraise the findings for themselves and evaluate whether, based on my description of the WBA system in the clinical radiology context, the results may have anything to say about their own area of interest.

#### 4.8.2 Establishing validity in this research

When considering how validity and reliability should be addressed in my research, it was clear that the blending of methods meant that my approach could not be driven by adhering to the validation principles and practices of a single paradigm. In any case, as I have already demonstrated, there is ongoing and at times vigorous debate *within* individual paradigms as to how validity should be established. In practice, researchers often opt for their own particular approximation to one or more recognised approaches to demonstrating rigour. For example, Prins *et al.* (2006) undertook a reliability exercise as part of their efforts to allocate numerical scores to the quality of written feedback comments on assessments used in the course of general practitioner training. Two researchers scored a third of the written feedback statements and their scores out of 100 for each statement were compared using Cronbach's alpha ( $\alpha$ ) as a measure of inter-observer consistency. Having established what they termed an 'acceptable' level of consistency (an average of 0.84 across all variables, with a minimum consistency of 0.74) they then disregarded the scores of the second observer and based their subsequent analysis on the scores allocated by the first observer. By contrast, they did not undertake a similar exercise for another portion of their study, in which they report that they categorised (but did not score) the 'style' of written feedback according to particular criteria (p. 294). They did not provide a rationale for the difference in approach.

In their analysis of written comments in workplace-based assessments, Canavan *et al.* (2010) undertook a process of establishing, refining and applying a coding framework which involved the input of four researchers in the design of the framework, and two researchers coding their entire sample of feedback 'phrases' (p. S107). However, they made no attempt to make statistical comparisons between the coding undertaken by the two researchers. Instead, they appear to have undertaken the exercise in order to ensure consistency of approach *throughout* the coding exercise (ie to counter the

potential influence of individual coder 'drift' when applying the framework) and in particular to gain agreement when making key decisions about 'borderline' cases:

two of us...independently coded the entire set of phrases and identified and resolved discrepancies; only minimal modifications to clarify differences between coding categories were made at that stage (e.g., *refining the parameters of what was considered a global assessment versus a general behaviour*) (Canavan *et al.* 2010, p. S107, emphasis mine).

This aligns with the view of Graneheim and Lundman (2004) that, while they do not believe it necessary to establish statistical reliability between the coding efforts of multiple researchers, it is nonetheless a good idea to conduct a 'dialogue among co-researchers' (p. 110), not solely for the purposes of verifying that the data have been labelled consistently, but also to gain agreement with regard to the way in which the data were coded and sorted. Even when the goal *is* verification of a consistent approach to coding, the approaches adopted by different researchers may not guarantee this. It is interesting to note, for example, that once the initial statistical comparison of researchers' coding efforts was undertaken in the Prins *et al.* (2006) study, no further verification or validation of the subsequent coding process was undertaken. Thus, nothing was done to guard against coder drift throughout the research, and there was no apparent discussion of borderline cases or important areas of disagreement.

My approach to validation involved asking a colleague to code a sample of assessor feedback statements using the coding framework that I had developed. The person concerned was employed in a teaching and research capacity in the field of medical education, but was not formally associated either with me (ie they were not a current work colleague) or the RCR. The aim of this exercise was not to calculate a reliability statistic. Instead, and in my view more importantly, it was undertaken to explore any areas of potential disagreement or uncertainty in the application of codes. The colleague in this exercise was supplied with 50 examples of assessor feedback comments and a copy of the coding framework, along with instructions for applying the the codes to the data. A meeting was also held prior to the coding exercise in which worked examples of how each of the codes should be applied were discussed. The coding exercise involved both me and my colleague coding the same set of 50 feedback comments separately, and then engaging in a further discussion in order to

understand any differences and, where necessary, refine or clarify the rules for applying certain codes. The discussion led to a number of important decisions regarding the refinement and application of codes to the assessor comments, and these are summarised below.

A key question arising out of the exercise was whether comments that were clearly intended to be developmental should also be regarded as implicitly 'negative' in regard to the performance that had been observed. In some cases, developmental comments were written such that they simultaneously implied a need for improvement and a problem with the observed performance. For example, 'Would benefit from continued experience to increase proficiency.' This led to a conversation as to whether *any* developmental comment potentially implied a need for improvement based on the the observed performance of the trainee. It was decided that, unless the assessor was explicit that their developmental comments did *not* imply criticism of the observed performance (and there were examples of this in the samples of 500 assessor statements), all developmental comments would also be coded as 'negative' feedback, as well as being coded as 'developmental'. This was done as it was felt that failing to code these comments as 'negative' would lead to them not being returned in any subsequent searching or filtering of the data (for example, in order to identify the prevalence of positive versus negative comments) which could lead to implicitly negative judgements about trainees' observed performance being systematically underrepresented. These comments would still be distinguishable from *explicitly* negative comments on the observed performance, which would be coded as either ['negative' + 'specific comment on observed performance'] or ['negative' + 'general comments on observed performance'].

Another important topic of discussion was the observation that some of the general developmental comments seemed to contain implicit assumptions about how clinical capability develops. This was typically suggested by the often-repeated assertion that 'seeing more patients' or 'doing more procedures' would result in the necessary improvements being made. Although the pithy nature of much of the data meant that the text was largely treated semantically, it seemed that this latent theme was simply too recurrent, and potentially too important in terms of revealing assessors' assumptions about skill development, to be ignored. Consequently, the code 'assumed improvement' was created in order to identify these comments.

These conversations were useful in developing the coding framework, however its application was not always straightforward, and this is illustrated below in the interests of increasing transparency of, and confidence in, the findings of the research.

#### *Applying the framework – straightforward examples*

In many cases, application of codes was straightforward. For example:

‘You communicated effectively with the patient’

is clearly a comment on observed performance, is intended to be positive, and is not a recommendation for any improvement or development. This comment would also have been coded as general, as the assessor has not identified to what aspect of communication they are referring. There are several dimensions of communication to which the assessor might have been referring here, for example: taking informed consent from the patient; giving instructions to the patient (about how to position themselves during the procedure, when to breathe or hold their breath etc.); keeping the patient informed about what the doctor was doing during the procedure; responding to the patient's questions; reassuring the patient and so on. Therefore, reference to 'communication' or 'communication skills' without any further qualifying or clarifying comment was deemed to be general rather than specific. For comparison, the comment 'You gave a clear explanation of the procedure to the patient prior to commencing,' is an example of a comment on communication which was deemed to be 'specific'.

#### *Applying the framework – ambiguous comments*

There were comments encountered throughout the analysis process which tested the limits of the decision-making approach. The following comment is an example:

Communication of result to patient could have been handled better. Patient had to ask if anything had been found. This could have been pre-empted.

The comment refers to a particular aspect of the doctor-patient communication that was observed – communication of a result to a patient – and the comment is clearly intended to be negative feedback about that aspect of the procedure, and so these aspects were easily coded. It also initially seems clear that the specific problem was that the doctor did not effectively communicate the result of the investigation to the patient, and so this would be coded as being a 'specific' feedback comment.

However, the assessor introduces some doubt about the nature of the specific problem by talking about how communication of the result could have been 'handled better', when it seems from the remainder of their comment that the result was not communicated at all. This could have been the case, and is strongly implied by the assessor's assertion that the patient's question could have been pre-empted. On the other hand, it may also have been the case that the trainee communicated the result in a way that was in some way obscure to the patient, for example by using technical medical language or, conversely, ambiguous euphemism or metaphorical language, which meant that the patient had to request a more transparent statement of whether or not there was a problem. In addition, when the assessor asserts that the patient's question could have been 'pre-empted', it is not clear whether they mean that this could have been pre-empted by communicating the result earlier, or pre-empted by telling the patient at an earlier point that the investigation would *not* yield a result straight away, and that they would have to wait for the outcome to be reported to them at some point in the future. In taking a pragmatic approach to the data, it seemed likely here that the trainee had simply not communicated the result to the patient prior to the patient's question, either because they forgot to do so, or left it so late that the patient felt they could wait no longer before asking, and the assessor clearly felt that this should have happened at an earlier point in the clinical encounter. Therefore this comment was coded as being *specific*.

A much clearer example of specific positive feedback in relation to the same aspect of communication (identified in the sample from 2010-11) was: 'Explained the procedure extremely well and following the procedure explained the relevant findings to the patient in a succinct manner.' My experience of the coding process was, therefore, that 'grey' cases like the one described above, which tested the limits and parameters of the coding criteria, and potentially raised questions about whether new codes might even be needed, often consumed a lot of time, consideration and reflection. Often, a short



time later, a comment was encountered that re-confirmed and re-established the rules of the coding process as it was generally being applied, and served to validate the decision-making process that accompanied the more challenging examples of feedback.

#### *Continuous recalibration*

Comments such as the one in which the assessor fed back that the trainee 'explained the relevant findings to the patient in a succinct manner' also illustrated that my expectations regarding the specificity of feedback comments were not unreasonable or unrealistic. Further examples reported in the results chapter provide evidence that assessors, on occasion, were able to provide detailed, specific feedback. This was important, as comments that fulfilled certain criteria, in particular comments that were both specific and developmental, were encountered so infrequently during the coding process that there was a real risk that the 'standard' for applying the code was dropped, such that a degree of coding drift took place. In order to guard against this, clear examples of infrequently-encountered types of comment triggered a review of the last several times that the code had been applied in order to establish firstly whether the code had been applied consistently across all comments. The complete data set was also reviewed to check whether other examples existed that had not been coded, due to their infrequency. This process of continual re-calibration, followed by a review of previous coding, was a feature of the coding process employed across all samples of feedback statements – the original sample of 500 statements from 2010-11, and subsequent samples, the origin of which is discussed in the following section.

#### **4.9 Scaling up the research**

Having established the prevalence of a range of qualitative features of the written feedback in a sample of 500 assessments from 2010-11, I was interested in exploring the extent to which these patterns were found throughout the remainder of the data. The data at this point consisted of not only the original 2010-11 data set, but also equivalent data from 2011-12 and 2012-13. Thus, there were two main options for the further exploration of assessor comments. The first was 'horizontal' expansion of the sample, in which another sample comments from the same training year (2010-11)

could be analysed in order to confirm whether the findings of the first sample of 500 assessor statements were representative of the whole 2010-11 population. A second option was 'vertical' expansion, in which samples drawn from successive years could be analysed in order to explore whether trends and patterns found in the first year of the programme having been introduced changed as assessors and trainees became familiar with the system.

#### *4.9.1 Expanding the coding process horizontally*

In considering how the coding framework might be applied to the remainder of the 4978 assessments that comprised the 2010-11 population, it was clear that no simple method was likely to be found. However, the technological affordances of the MAXQDA™ software meant that certain approaches could be attempted. The most promising approach was one which used the search function in MAXQDA™. The method was first to identify whether certain terms or phrases were characteristic of particular types of feedback found in the sample, and as such could act as a proxy indicator for these types of feedback. Any suitably identifying and discriminating terms could then be searched for electronically in the whole population, with the resulting 'hits' being inspected manually to confirm accuracy.

Suffice it to say that, due to the essential ambiguity that surrounds the use of even very specific-seeming words and phrases, none was found which was a suitably reliable indicator of feedback type to allow it to be used to satisfactorily screen the whole 2010-11 data set for different types of feedback. Having initially abandoned the idea of trying to apply the coding framework to all assessments in the 2010-11 population, I decided that it would be more achievable, not to say more revealing, to apply the same coding framework to samples of assessments taken from successive years of the programme.

#### *4.9.2 Expanding the coding process vertically*

The idea of applying the coding framework to data from subsequent years was appealing as the research process had already been applied and refined with the 2010-11 data, and could then be applied to other years to offer a direct comparison between the findings. Doing so would therefore add a longitudinal dimension to the study. This was not to say that the study was longitudinal in the conventional sense, as participants

were not consistent from one year to the next. However, this in itself was interesting in order to establish whether the patterns identified by the initial coding process in the 2010-11 data were features of the assessment and feedback system that were conserved over time, regardless of the exact composition of the trainee and assessor population.

In attempting to make comparisons between the patterns that were found in each year of training, my preference was to have the option of doing so statistically. It became clear at this point that the statistical functions within the MAXQDA software were extremely limited, and would not allow me easily to perform the necessary calculations. This was confirmed by the technical support team at MAXQDA. By comparison, a standard spreadsheet package, such as Microsoft Excel, offered much greater scope for making this type of comparison. However, it was also clear that none of the exporting functions of MAXQDA would allow me easily to export the previously-coded sample of 2010-11 data to Excel in a format that would allow all of the codes associated with each feedback statement to be preserved. I contacted the MAXQDA developers, who confirmed that the only export function that approximated to my requirements was an option to export all of the feedback comments that had been coded with a particular individual code, and so the overlapping aspect of the coding process, not to mention the structure of the overall feedback statements, would be lost.

In resolving these problems, I decided to generate a second random sample of 500 assessments from 2010-11 and, using the same coding framework and rules already established for the coding process in MAXQDA™, code this second sample in Excel. I then did the same for a sample of 500 assessments from 2011-12 and 2012-13. The result of this was that I had three samples from successive years, all coded in Excel and therefore capable of being analysed quantitatively. In addition, I then had two samples of 500 statements drawn from the training year 2010-11 (one coded in MAXQDA™, and the other coded in Excel), and so it was possible to make comparisons between the coding frequencies in these two samples and hence determine how representative my original sample of 500 feedback phrases actually was.

## **4.10 Statistical analysis**

Having coded data from three successive training years, including two samples of data from the first year of the programme (2010-11) and one sample each from 2011-12 and 2012-13, univariate inferential statistics were most appropriate for comparing the differences in coding frequencies between years, and to compare relationships between the feedback features identified within each year of training.

### *4.10.1 Analysing relationships between years*

Chi-squared analysis was used to compare coding frequencies between the two samples of 500 feedback statements taken from the training year 2010-11. The first sample was coded in MAXQDA and the second in Microsoft Excel. The analysis was conducted via 2x2 contingency tables, and a separate Chi-squared calculation was performed for each code used. Few significant differences between coding frequencies in the two samples were found (see Chapter 5, section 5.3.1), and so samples of 500 feedback statements were taken to be likely to be representative of the populations from which it was drawn.

A second series of Chi-squared analyses was also conducted, in order to look for significant differences between coding frequencies across all four samples (two samples from 2010-11, and one sample each from 2011-12 and 2012-13). These are reported in Chapter 5, section 5.3.2.

### *4.10.2 Analysing relationships between types of feedback and other aspects of the assessments*

I was interested as to whether the occurrence of particular types of feedback was dependent on particular properties of the trainees who were being assessed, such as whether they were early stage ('junior') or late stage ('senior') trainees, or whether they were judged to have performed well or poorly in the assessment. In addition, I had observed during the coding phase of the research that it was tempting to assume that longer passages of feedback were likely to be of higher quality than more cursory feedback statements, and yet the objectively-determined quality of these longer feedback statements suggested that at times they contained little of real educational

value. Consequently, I looked for a method of exploring the coded data that would provide evidence of certain relationships.

An approach that was considered was the use of regression analysis to 'sift' the data for significant relationships. However, in discussion with my supervisor it was agreed that a more focused comparison of feedback types with purposefully chosen conditions (seniority of the trainee, assessment outcome and length of feedback) was an important first step prior to considering regression analysis. Consequently, Chi-squared analysis was used to examine whether there were any statistically significant relationships between these features of the recorded Rad-DOPS assessments and the occurrence of particular types of feedback. The relationships examined were as follows:

- (1) I was interested in whether the trainee's overall performance in the assessment impacted on the likelihood of them receiving different types of feedback comment. Consequently, I used Chi-squared to analyse the relationship between average (modal) assessment score and a range of different type of feedback comment: positive feedback; negative feedback; specific positive comments on observed behaviour; specific negative comments on observed behaviour; comments that were clearly linked to the assessment criteria; general developmental comments; specific developmental comments..
- (2) I was also interested as to whether the assessor's overall qualitative judgement had a bearing on the type of feedback that was provided. Thus, I analysed the relationship between overall competence rating and a range of different types of feedback comment.
- (3) Given the formative nature of the WBAs, I was interested as to whether trainees in the earlier stages of training were more likely to receive certain types of comment. Therefore I analysed the relationship between stage of training and different types of feedback comment.
- (4) I was interested as to whether assessors who provided lengthier feedback statements were necessarily more likely to provide the types of feedback comment that were likely to be helpful to the learner (e.g. specific rather than general). I therefore analysed the relationship between length of feedback and

a range of different types of feedback comment. In order to explore the relationship between length of feedback and type of feedback provided, a method for determining the difference between 'brief' and 'extended' feedback was sought. Details of the decision-making process for this are provided in the next section.

Results of this statistical analysis are presented in the next chapter, Chapter 5.

#### *4.10.3 Length of feedback – brief or extended?*

In order to facilitate the exploration of the relationship between length of feedback and the presence of certain types of feedback comment (relationship (4) above), it was first necessary to identify feedback statements of different lengths and develop appropriate categories for these. My first approach, and the one that I consequently used, was to try to categorise feedback statements according to two categories – brief feedback, and extended feedback. In an attempt to distinguish between cursory (or 'brief') feedback statements and more lengthy ('extended') feedback statements, assessors' comments from the original sample of 500 assessments from 2010-2011 were analysed quantitatively (see Table 4.1). In choosing a method for separating brief comments from extended comments, it was clear that this would have to be done in a subjective manner. However, it was not done arbitrarily, and the rationale is set out as follows.

It did not seem fair or legitimate to choose a very low threshold for distinguishing brief from extended comments, as comments of only a few words could not exhibit many of the educationally desirable features that were identified during the coding phase of the research. Choosing too low a word count threshold for distinguishing between 'brief' and 'extended' feedback would therefore have led to the circular logic that feedback comprised of few words contains little value. Equally, setting too high a threshold would have potentially included most of the feedback statements in the sample, thus failing to discriminate between 'brief' and 'extended' in this context. Consequently, I filtered the feedback data for comments of increasing length until I found a word count that could potentially contain educationally valid feedback comments, and yet which still 'felt' like

brief comments in the context of the length of feedback commonly provided in the course of the Rad-DOPS assessments.

In analysing feedback comments of gradually increasing length, it was found that feedback consisting of single words failed to fulfil any of the coding criteria, other than perhaps the standalone coding of 'positive' where the word used was clearly intended to convey the positive regard of the assessor. Examples of the terms used by assessors included: 'Satisfactory'; 'Good'; 'Excellent'. There were also neutral terms: 'none'; 'nil'; 'N/A'. Feedback that consisted of two words also tended to contain generally positive remarks such as 'Well done' or 'No concerns'. Terms that may have been coded as feedback in the qualitative dimension of this study included terms such as 'Good technique' or 'Good communication'. These would have been coded as 'positive valency' and 'general behavioural comment', with 'general' being a somewhat generous description of the comment. There were also very general descriptions of progress e.g. 'Satisfactory progress'; 'Good progress'. Phrases such as 'Good technique' or 'Good communication' were technically linked to the assessment criteria, and were coded accordingly, but basically repeated them and added nothing to them. There were no suggestions for improvement found within in these very brief phrases.

Feedback statements consisting of 3-5 words displayed very similar characteristics, and tended to be general comments about the procedure, or the trainee's overall progress. Comments at times invoked the assessment criteria, but again only to reiterate the criterion rather than to expand on it in any way. The first phrases that might be considered to be negative feedback, indicating a need for improvement, appeared within this range. They comprised only 14 of the 534 comments coded within this subset of the 500 assessments, and the recommendation on all 14 occasions was some variation of 'do more procedures'.

Comments that consisted of 6 – 10 words followed a similar pattern, however some comments that were 9 or 10 words in length showed evidence of being able to convey feedback which, while not particularly elaborate, satisfied aspects of the coding framework which the more cursory comments did not (ie specific comments on observed performance or specific suggestions for development).

Taking the empirical observations in summary, I decided that feedback statements that consisted of 1-10 words would be classified as 'brief' feedback. Feedback that was comprised of 11 or more words would therefore be regarded as 'extended'. It should be emphasised at this point that this brief/extended decision was not based on thorough analysis of all of the coded statements, but rather an impressionistic visual inspection of the shortest of the feedback statements and a similar inspection of several longer statements. Further statistical analysis was needed in order to truly verify the relationship between these shorter and longer feedback statements and the presence or absence of the most helpful comments. Furthermore, Chi-squared analysis was sufficient only to establish whether the presence of each type of feedback in each of the two length categories of feedback ('brief' and 'extended') was significantly different from what would be expected. However, the Chi-squared statistic alone was not able to identify if the provision of extended feedback was a necessary condition for higher quality feedback comments to be given, nor whether extended feedback was a sufficient condition for the provision of these comments. Establishing the necessity and sufficiency of particular conditions was better addressed by Ragin's approach to qualitative data analysis, which is described below. The inclusion of Ragin's approach underscores the emergent nature of the research design, which evolved as I formatively reviewed and developed a more in-depth and nuanced understanding of the data.

Table 4.1 Frequency of feedback statements featuring a range of different word counts in the total population of 4798 Rad-DOPS assessments recorded by clinical radiology trainees between August 2010-July 2011.

Length of feedback statement (word count)	0	1	2	3-5	6-10	11-20	21-30	31-50	51-100	100+
No. of assessments	91	43	262	534	984	1295	725	564	269	29
% of total	1.9	0.9	5.5	11.1	20.5	27.0	15.1	11.8	5.6	0.6
Cumulative %	1.9	2.8	8.3	19.4	39.9	66.9	82.0	93.8	99.4	100



#### 4.11 Exploring conditions for effective feedback

My use of chi-squared analysis had indicated that there were few consistently significant relationships between certain assessment conditions (stage of training, modal assessment score, overall assessment judgement) and particular types of feedback being provided by assessors (e.g. positive feedback, negative feedback, specific comments on observed performance and so on). According to Glaesser and Cooper (2012), this is not uncommon when conducting traditional statistical analysis of coded data drawn from the naturalistic settings in which most social sciences research is conducted. As these authors point out, traditional statistical analysis is often too rigid to be able to reveal the meaningful but less absolute relationships between variables that are typically encountered in naturalistic social science settings. Instead, these authors have proposed an alternative approach to qualitative comparative analysis (QCA) first described by Ragin (1987, 2000, 2008). This approach offers the possibility of exploring coded data numerically but with due regard to the more 'fuzzy' nature of relationships that exist in the field in social science research. For this reason, I chose to employ a modified version of Ragin's (*ibid.*) technique as described below.

##### 4.11.1 Ragin's approach – necessity and sufficiency

According to Glaesser and Cooper (2012), Ragin's approach to QCA offers the advantage of retaining the causal complexity of the social world, in which factors which are interrelated can be treated as interrelated and interdependent, rather than being regarded as independent (which is often a requirement for conventional statistical analysis). These authors argue that, in the world of the social sciences, a particular outcome (e.g. educational attainment) may be causally linked to a number of conditions or predicates. Particular conditions, or combinations of conditions, may function either together or in isolation to bring about the outcome. Thus, these conditions may be regarded as 'sufficient' for the generation of the outcome. However, in complex social settings, other conditions or combinations of conditions may also be sufficient to bring about the same outcome. These alternative conditions may also be regarded as sufficient, but it is clear that neither set of conditions is 'necessary' in itself – if one sufficient condition (or set of conditions) is absent, the presence of the other would still bring about the outcome.

An example offered by Glaesser and Cooper (2012), taken from the work of Mahoney and Goertz (2006), is as follows:

$$Y = A*B*c + A*B*C*D$$

In this formula, the outcome of interest is represented by the upper case Y. The asterisk (\*) represents the logical 'AND', while the plus (+) sign represents the logical 'OR' operator. In addition, in Boolean notation, capital letters signify the presence of a condition (e.g. A) whereas lower case letters represent the absence of a condition (e.g. c). The requirement for a particular condition to be absent is also referred to as a 'NOT' function. In this formula, the combination of conditions denoted by  $A*B*c$  is *sufficient* to bring about the outcome Y. However,  $A*B*c$  is not *necessary* to generate Y, as Y may also be brought about by the presence of  $A*B*C*D$ . Mahoney and Goertz's (2006) example deals with combinations of predicates, but the logic still holds true for individual conditions that are capable of giving rise to a particular outcome.

#### 4.11.2 Quasi-sufficiency and quasi-necessity

A second principle of Ragin's approach to QCA is that, in the naturalistic social setting, relationships between variables are rarely perfectly overlapping, but instead overlap somewhat imperfectly. These imperfect yet important relationships are dealt with using the concept of fuzzy logic, which was first proposed by Zadeh (1965). In traditional logic, relationships are deemed either to be true, or untrue, and are consequently assigned the value 1 or 0 respectively to indicate this. In fuzzy logic, relationships that are partially or mostly true may be represented by a number between 0 and 1, with the strength of the relationship increasing as the number approaches 1. Thus, fuzzy logic allows the researcher to analyse relationships between predicates and outcomes in terms of partial truth, and to reflect degrees of connectedness between the two, as opposed to being confined to the binary true/false decisions that are normally supported by conventional logic. This can be illustrated using Venn diagrams (see Figures 4.6 and 4.7), and the language used to describe such relationships refers to 'coverage', which, as Glaesser and Cooper (2012) identify, is analogous to the concept of variance which is found in regression analysis. Predicates and outcomes that

achieve sufficient coverage may be classed as quasi-necessary or quasi-sufficient, as long as they exceed threshold values. Typically, coverage of at least 0.8 (or 80%) is taken to be adequate for quasi-sufficiency (or quasi-necessity) to be declared, with 0.7 (or 70%) normally demarcating the lowest threshold.

These concepts of quasi-sufficiency and quasi-necessity were useful in my study, in which traditional statistical analysis had revealed few consistent relationships between particular conditions and the incidence of different types of feedback. However, in order to conduct Ragin analysis of these conditions, it was necessary to identify a suitable outcome for which the sufficiency and necessity of certain conditions could be tested.

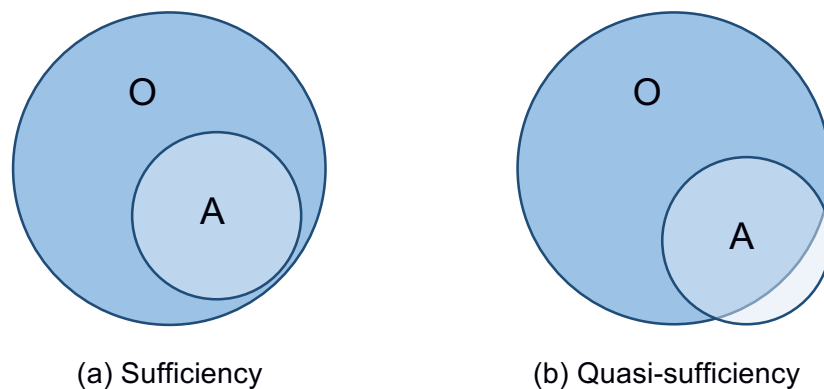


Figure 4.6 Venn diagrams illustrating the concept of sufficiency and quasi-sufficiency, after Glaesser and Cooper, 2012. O = outcome; A = condition. In (a), if A is present, then O occurs. In (b), if A is present, then O *nearly always* occurs.

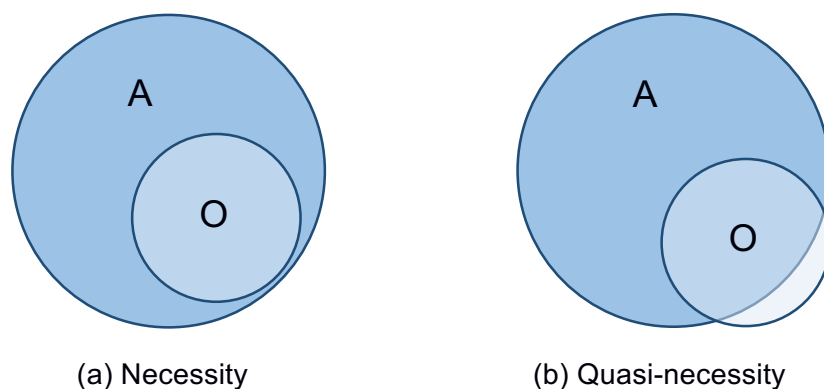


Figure 4.7 Venn diagrams illustrating the concept of necessity and quasi-necessity, after Glaesser and Cooper, 2012. O = outcome; A = condition. In (a), in order for O to occur, then A must be present. In (b), in order for O to occur, then A must *almost always* be present.

#### 4.11.3 Choosing an outcome

A difficulty that arose when utilising Ragin's approach was that there was no single outcome measure for which the potential necessary and sufficient conditions could be analysed. The outcome in which I was interested – 'good quality feedback' – was not a single entity, manifesting as the presence of a single type of feedback comment. Rather, it was a composite, determined by the analysis of the formative assessment literature, and consisted of the presence of certain types of comment (e.g. specific comments on observed behaviour), and the absence of certain other types of comment (e.g. personal comments). It was therefore possible to construct qualitatively a model of 'ideal' or 'high quality' feedback, which was comprised of a combination of features that were supported by the review of formative assessment literature as being educationally beneficial. These features were as follows:

**The presence of:**

**positive and negative comments; specific comments on observed performance; linkage to the assessment criteria; specific suggestions for development,**

**and the absence of:**

**global feedback; personal feedback; general comments on observed performance; general developmental comments.**

In using these parameters to filter the coded assessor feedback statements from all three years of assessment data, **no examples were found**. This was in itself a significant finding. It also created the further complication of identifying a level of feedback quality that would allow the database software to return some analysable results.

In attempting to find a standard that was suitably stringent and yet allowed for the analysis of assessor feedback comments, the requirement for general comments to be absent was dropped. This was due to the fact the assessment-related evidence had not shown general comments clearly to be directly linked to negative educational outcomes, to the extent that they may undo the work of specific comments. Thus, with the requirement for specific comments retained, the presence of general comments was ignored. The requirement for global comments to be absent was dropped for the

same reason. However, the requirement for personal comments to be absent was retained. This was due to the weight of evidence in the literature demonstrating the unhelpful nature of these types of comment, including their ability to divert the learner's attention away from any task-related feedback, regardless of its quality. In applying these criteria to the second set of 500 feedback statements taken from the 2010-11 data, only 23 feedback statements met the criteria. Given that the sample of 500 statements was comprised of assessments for trainees from ST1-ST5 who had achieved a range of global outcomes and numerical assessment judgements, it was felt that further analysis of this subset of feedback statements by further dividing it by stage of training, modal assessment score, overall assessment judgement and length of feedback was likely to result in subsets of very small numbers of comment, which may not be usefully or credibly analysed by Ragin's technique.

Similar numbers were found when applying the same filter to the data from 2011-12 and 2012-13, and so a still lower threshold for feedback quality was sought. Removing the requirement for *specific* (rather than general) comments on the observed performance resulted in an increased number of feedback statements which could be subjected to further analysis, and so the outcome of 'high quality feedback' was taken to be feedback which met the following criteria:

**The presence of positive and negative comments (either specific or general) on the observed performance with specific suggestions for further development, and the absence of comments at the personal level.**

Feedback statements that satisfied these criteria were found in 41/500 (8%), 34/500 (7%) and 37/500 (7%) of the Rad-DOPS assessments sampled from 2010-11, 2011-12 and 2012-13 respectively.

#### 4.11.4 Necessity or sufficiency?

Another consideration in applying Ragin's approach was whether to analyse the feedback data in terms of necessity or sufficiency. For example, Glaesser and Cooper (2012) had chosen to explore their data for sufficiency in the first instance, and presented their results accordingly. The relationship between sufficient conditions and

a given outcome is that the condition must be a subset of the outcome (as illustrated in Figure 4.6). Given the relatively small numbers of assessments that displayed the chosen outcome ('high quality feedback'), it seemed unlikely that any of the four conditions chosen for analysis would be likely to overlap with it substantially enough to be labelled as sufficient or quasi-sufficient. Instead, I chose to begin Ragin analysis by examining the *necessity* of each chosen condition for the provision of 'high quality feedback'. For necessity to occur, the outcome must be a subset of the necessary condition. Given the relatively small numbers of feedback statements that bore the chosen outcome of 'high quality feedback', it seemed most sensible to begin by examining the potential necessity, or quasi-necessity, of these conditions for the provision of 'high quality feedback'. Once the necessity analysis was completed, it was relatively straightforward to check each of the chosen conditions for sufficiency as well.

As previously mentioned, the conditions that were analysed for potential necessity and sufficiency were: modal assessment score; global assessment judgement; stage of training; length of feedback. The results of this analysis are reported in the next chapter.

#### **4.12 Analysing trainee comments**

Much of the methodology until this point has focused on the analysis of assessor feedback comments. However, as mentioned in section 4.5, the analysis of trainee comments was felt to be important in establishing the level of engagement of trainees with the written feedback process. Nicol (2010), writing about the utility of written feedback in higher education, acknowledges that 'while the quality of teacher comments is important, engagement with and use of those comments by students is equally important' (p. 503). Nicol (2010) does not specify precisely what this engagement should look like, but he does offer a number of suggestions as to how engagement might be supported, all of which require structured opportunities for face-to-face dialogue with or between learners. Thus when writing about creating more dialogic written feedback, Nicol's (*ibid.*) recommendations are actually almost wholly centred on various facets of what he terms the 'quality of teacher comments' (*ibid.*). Unlike most higher education settings, trainees in clinical radiology have the opportunity routinely to respond in writing to assessors' written comments, and so it

was possible to analyse pairs of assessor-trainee comments for signs of learner engagement, including genuine dialogical interaction.

In preparing to undertake analysis of trainee comments, I was aware that the review of literature for this research had failed to identify an existing framework for judging learner 'engagement' in written feedback. Consequently, the trainee comments in my study were coded inductively and analysed for signs of engagement with assessors' comments. More deductively, I was interested in whether or not there was any evidence of the written feedback exchanges between assessor and trainee being truly dialogical, and so pairs of assessor-trainee feedback exchanges were analysed with this in mind.

Paired assessor-trainee feedback statements were drawn from all three years of the research. Details of the inductively developed coding framework and the conditions for assigning codes to trainee statements can be found in chapter 5, section 5.6.

#### **4.13 Summary**

This chapter has outlined the complex, multi-methods approach to this research, which was undertaken in order to address the complexity of the research question and the challenges presented by both the nature and the quantity of the data. The chapter that follows presents the findings of this research. These are then drawn together in the final chapter in order to synthesise the empirical and theoretical aspects of the work and make appropriate recommendations regarding the fitness for purpose of the current WBA system in clinical radiology.

## CHAPTER 5

### 5. Results

#### 5.1 Introduction

This section presents the findings of a number of different approaches to data analysis with respect to the Rad-DOPS assessment data that were provided by the Royal College of Radiologists covering three consecutive training years from 2010-11 to 2012-13.

The first section is a presentation of relevant descriptive data, which depicts the patterns of assessment with respect to their timing within postgraduate clinical placements and the numbers of assessments conducted by both the clinical radiology trainees and their assessors. In providing a description of the assessment process *in vivo*, as opposed to official versions of what *should* be happening, or what has been found in controlled or pilot studies which offer something of an *in vitro* view of the workplace-based assessment process, these data were useful in answering key aspects of the research questions, such as whether or not trainees have been using the assessments formatively. The base-data here consisted of the total number of Rad-DOPS assessments recorded by all clinical radiology trainees in the UK in each of three consecutive training years: 2010-11, 2011-12 & 2012-13. As such, the data display a comprehensive picture of Rad-DOPS workplace-based assessment activity and outcomes nationally for the medical specialty of clinical radiology.

The second section presents the findings of content analysis of assessors' written feedback, conducted according to the approach set out previously in the methodology. The findings from the coding process are presented quantitatively, in order to display code frequencies and patterns, and qualitatively, in order to illustrate the types of feedback provided by assessors and allow the reader to make independent judgements about the validity of the coding process and, ultimately, the findings of the research. This section also includes the outcomes of statistical analysis that was



undertaken in order to gauge the representativeness of a sample of 500 assessor feedback comments.

The third section presents the results of inferential statistical analysis, specifically Chi-squared testing, which was undertaken in order to establish whether there were any significant differences between the coding frequencies in one year compared with the other years that were included in the study. Further Chi-squared analysis was undertaken in order to establish whether there were statistically significant relationships between certain assessment parameters and the provision of high quality written feedback. These parameters included the seniority of the trainee, the average score awarded in the assessment and the overall assessment judgement awarded by the assessor.

The fourth section presents the results of a particular approach to qualitative comparative analysis which has been described by Ragin (1987, 2000, 2008), and which has been adapted and applied here in order to explore relationships between certain feedback conditions and the provision of high quality written feedback. Ragin's approach aims to reflect the naturalistic context of social science research, in which relationships rarely reflect the discrete and distinct connections that might be predicted by mathematical formulae or replicated neatly in the laboratory. Rather, Ragin's approach allows the exploration of necessary and sufficient conditions for a particular outcome by allowing conditions that fall within certain parameters to be characterised as quasi-necessary or quasi-sufficient. This section identifies several quasi-necessary conditions for the provision of high quality written feedback, none of which was found to be sufficient to guarantee this particular outcome.

The final section sets out the evidence regarding the degree to which the written feedback exchanges within formative workplace-based assessments might be said to be dialogical. This was established by first selecting assessments in which the assessors' comments fulfilled the qualitative criteria for identifying high quality feedback. The associated trainee comments in these assessments were then analysed to determine whether or not they appeared to have been made in direct response to the assessors' comments, and resulted in a clear plan for further learning.

## 5.2 Descriptive statistics

The national radiology training e-portfolio is a central electronic repository of information that is updated in real time as trainees across the UK update their training record. It is therefore capable of yielding information about the system of formative workplace-based assessment that has been introduced to postgraduate clinical radiology training. The Rad-DOPS assessment outcomes for all registered UK clinical radiology trainees in each of three successive training years were downloaded from the e-portfolio system and exported to Microsoft Excel. The Rad-DOPS data included all of the written comments provided to trainees by assessors, as well as other information such as the scores awarded to the trainees across the different assessment domains and the date on which the assessment was recorded.

### 5.2.1 *Patterns of assessment across all training grades*

The first step in analysing the Rad-DOPS assessment data was to determine how many trainees at each stage of training had recorded Rad-DOPS assessments within each training year. Descriptive statistics were used to reveal the central tendency and spread of the results. This was done by calculating the mean, median and modal numbers of assessments recorded by trainees at each grade from the first year of specialty training (ST1) to the sixth year of specialty training (ST6). The range of Rad-DOPS assessment numbers recorded by trainees within each grade was also calculated. These statistics were important for revealing the general pattern of assessments being recorded by clinical radiology trainees nationally, and allow comparison with the RCR's curriculum guidance regarding the number of assessments that should be undertaken.

Table 5.1 Descriptive statistics showing the numbers of Rad-DOPS assessments completed by trainees of all grades in the three training years from 2010-11 to 2012-13.

	2010-2011						2011-12						2012-13					
	Training grade						Training grade						Training grade					
	ST1	ST2	ST3	ST4	ST5	ST6	ST1	ST2	ST3	ST4	ST5	ST6**	ST1	ST2	ST3	ST4	ST5	ST6
No. of trainees	223	136	118	113	5	0	307	275	172	170	104	1	218	236	215	151	137	17
Total assessments completed	1934	1056	877	912	19	0	2383	2131	1346	1312	835	6	1663	1624	1345	1180	897	96
Mean assessments per trainee	8.7	7.8	7.4	8.1	3.8	-	7.8	7.7	7.8	7.7	8	-	7.6	6.9	6.3	7.8	6.5	5.6
Median assessments per trainee	8	7	7	7	4	-	7	7	7	7	7	-	7	6	6	7	6	4
Modal assessments per trainee	6	7	8	6	1	-	6	6	7	1*	4	-	6	5	5	6	5	1
Range	1-30	1-24	1-32	1-23	1-8	-	1-28	1-24	1-24	1-36	1-27	-	1-19	1-24	1-21	1-40	1-29	1-13

\*The mode here appears lower than would be expected, indicating that a large number of the ST4 trainees conducted only 1 Rad-DOPS assessment. However, the next most common frequency was 7, which is more in keeping with RCR guidance regarding the number of assessments that should be recorded.

\*\*Only 1 ST6 trainee had recorded any Rad-DOPS assessments in the e-portfolio, and so descriptive statistics were not calculated for this grade.

The descriptive data for all of the Rad-DOPS assessments completed by radiology trainees are displayed in Table 5.1. As can be seen, the mean, median and modal numbers of assessments recorded by trainees were in keeping with the curriculum requirement for at least 6 Rad-DOPS assessments to be completed within the year. This is particularly the case with ST1-ST4 trainees. The numbers of assessments undertaken by ST5 and ST6 trainees show a lag. This was most likely due to senior trainees who were already following the pre-2010 training programme being reluctant to change course so close to the completion of their training.

The fact that average numbers of assessments closely align with the RCR requirements perhaps demonstrates some strategic assessment behaviour amongst trainees. The range of assessment numbers is also revealing – at times, trainees recorded very low numbers of assessments, with some trainees recording only a single Rad-DOPS assessment within a whole training year. Conversely, some trainees recorded relatively large numbers of assessments, for example in excess of 30 assessments in a given year. It would have been useful to be able to contact these trainees in order to establish their reasons for conducting numbers so well in excess of either the recommended numbers (six Rad-DOPS per year) or the numbers of assessments typically being recorded by their peers, but as previously stated this access was not possible. Inspection of the individual e-portfolio records for some of these trainees revealed that:

For one trainee (an ST3, 2010-11) who had recorded 32 Rad-DOPS assessments,

- the assessments had been conducted by 13 different assessors
- the assessments had generally been recorded in blocks of three, four or five assessments, conducted on the same day.
- no assessor had conducted more than five assessments for this trainee

For another trainee (an ST1, 2010-11) who had recorded 30 Rad-DOPS assessments,

- these had been conducted by 10 different assessors
- eight of these had been conducted by one assessor, with three of them having been conducted on the same date right at the end of the placement when the trainee had already exceeded the required number of Rad-DOPS assessments

- other assessors had also recorded blocks of assessments for this trainee on a single date – for example, two different assessors recorded four assessments each on a single day. These two blocks of assessments were recorded 36% and 41% of the way through the placement respectively.

The finding that groups of assessments have been recorded on the same day suggests that the system may not be functioning as intended. Whilst it is not impossible that a trainee be assessed three, four or five times on the same day, it is more likely that observations that were conducted over a period of time were documented at a single point in time. This may have been because the trainee sent a batch of electronic assessment requests to the assessor on the same day, sometime after the observations. It may also have been because the assessor had failed to complete the documentation in a timely manner, choosing instead to 'batch' all of the documentation rather than completing it after each observation.

Interestingly in the example of the ST1, above, the assessor who had provided the three assessments right at the end of the placement commented on all three assessment forms to the effect that the documentation should be completed at the time of the assessment. This would suggest that the trainee had sent an electronic assessment request to the assessor a considerable time after the actual clinical procedures had been observed by the assessor. This is important, as a significant delay between observation and the completion of the assessment form is likely to have an impact on the quality of the feedback provided.

In summary, it appears that while the data demonstrate signs of the training process being novel, such as the initial lack of take-up by trainees in the most senior training grades, and a delay in take-up by trainees at all grades, the general pattern amongst trainees who have opted into the system at almost every grade is stable. The pattern is characterised by: the majority of trainees recording numbers of assessments that are generally in keeping with the minimum numbers required by the curriculum; some trainees in all grades recording low numbers of assessments, and at times only a single assessment, in the course of a training year; and some trainees at all grades recording numbers of assessments that are greatly in excess of what is required of them in order to progress to the next stage of training.

### 5.2.2 Patterns of assessments conducted by assessors

Under RCR guidance, Rad-DOPS assessments may be conducted by any healthcare professional who is competent in the relevant procedure (RCR 2010). The majority of the assessors tend to be doctors, rather than allied health professionals. For example, in 2010-11, 126 of the total 1691 assessors were non-medically qualified radiographers or sonographers; the remainder were qualified medical practitioners.

As illustrated in Table 5.2, the mean number of assessments conducted by individual assessors was generally quite low (2.8-3.3 assessments) with the modal number being only one assessment. Conversely, there were some assessors who conducted as many as 35 assessments. I was interested to ascertain whether there was any relationship between the assessors who had conducted large numbers of assessments and the trainees who, as already reported, had recorded similarly large numbers. The raw data were inspected in order to establish whether such a relationship existed, and it was found that none of the assessors who had conducted 20 or more assessments had assessed any of the trainees who had conducted 20 or more assessments.

Table 5.2. Total, range and average numbers of assessments completed by assessors in the first three years since the launch of the new training programme in clinical radiology.

Training year	Number of assessors	Total assessments completed	Mean	Standard deviation	Range	Median	Mode
2010-2011	1691	4798	2.8	3.0	1-30	2	1
2011-2012	2395	8013	3.3	3.6	1-35	2	1
2012-2013	2260	6805	3.0	2.8	1-26	2	1

### *5.2.3 Timing of assessments*

The stage of the placement during which assessments were conducted provides an indication of the extent to which the WBAs were being used formatively by trainees – assessments conducted towards the end of the training post offer little scope for influencing future performance, and hence would not be particularly well suited to being genuinely formative.

The patterns that emerged are displayed in Figures 5.1-5.3. As can be seen, the general trend is one of increasing numbers of assessments being conducted throughout the training post, with the sharpest rise occurring from 0 to 50% of the way through the post, and the bulk of assessment activity occurring from 50 to 100% of the way through the post. The peak of assessment activity apparent around the mid-point of the attachment most likely coincides with the mid-point educational appraisal discussion that the RCR recommends its supervisors conduct with their trainees. Similarly, the peak toward the end of the attachment (90%) probably coincides with the formal end-of-attachment appraisal discussion.

There was evidence of a large number of assessments being recorded at the very end of placements – 527 assessments in 2010-11, 971 in 2011-12 and 746 in 2012-13, representing 11-12% of the assessments done in these years. The data also revealed evidence of retrospective assessment, with fairly substantial numbers of assessments being recorded beyond the end of the training posts – 403 in 2010-11, 1055 in 2011-12, and 669 in 2012-13, representing 8-13% of the assessments done in these years. It is important to consider the extent of very late or retrospective assessment due to the lack of opportunity for trainees to respond to feedback that is provided at the very end of the attachment, or after the attachment has ended. Retrospective assessment is also evidence of assessment documentation being completed some time after the observation of trainee performance has occurred – trainees who have finished their clinical attachment cannot be observed conducting radiological procedures. The documentation of the assessment beyond this end point clearly indicates a delay between observation and the completion of the assessment forms and begs the question not just of its accuracy but of its formative potential..

Fig 5.1. Stage within training attachment that Rad-DOPS assessments were recorded in the training e-portfolio in the training year 2010-2011. The stage is expressed as a percentage of the total time in the training attachment, rounded to the nearest 10%.

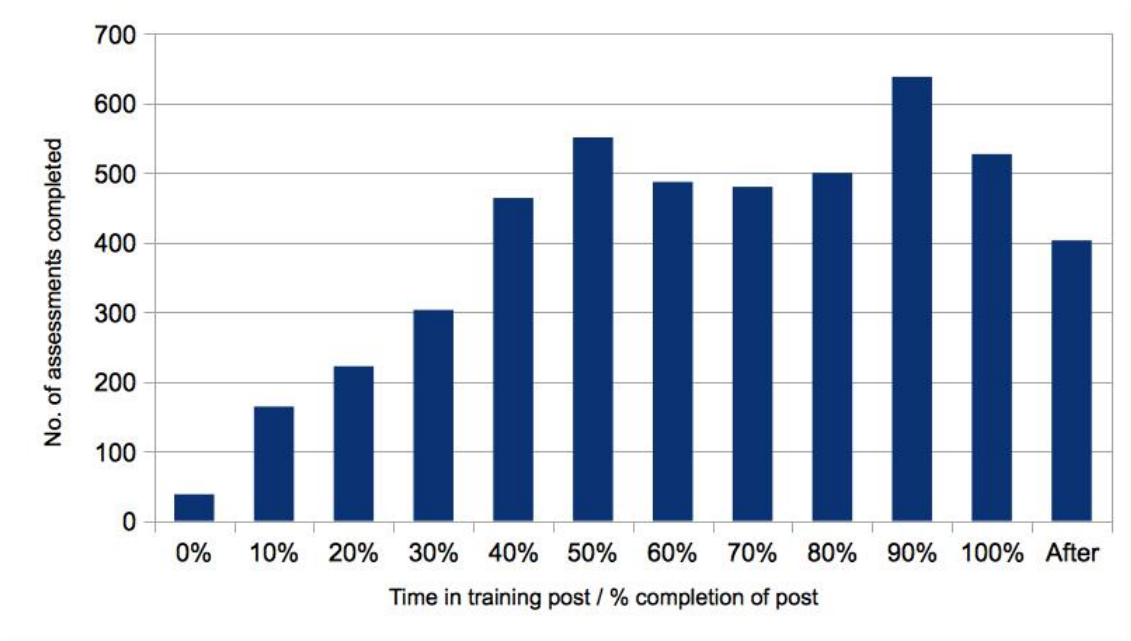


Fig 5.2. Stage within training attachment that Rad-DOPS assessments were recorded in the training e-portfolio in the training year 2011-2012. The stage is expressed as a percentage of the total time in the training attachment, rounded to the nearest 10%.

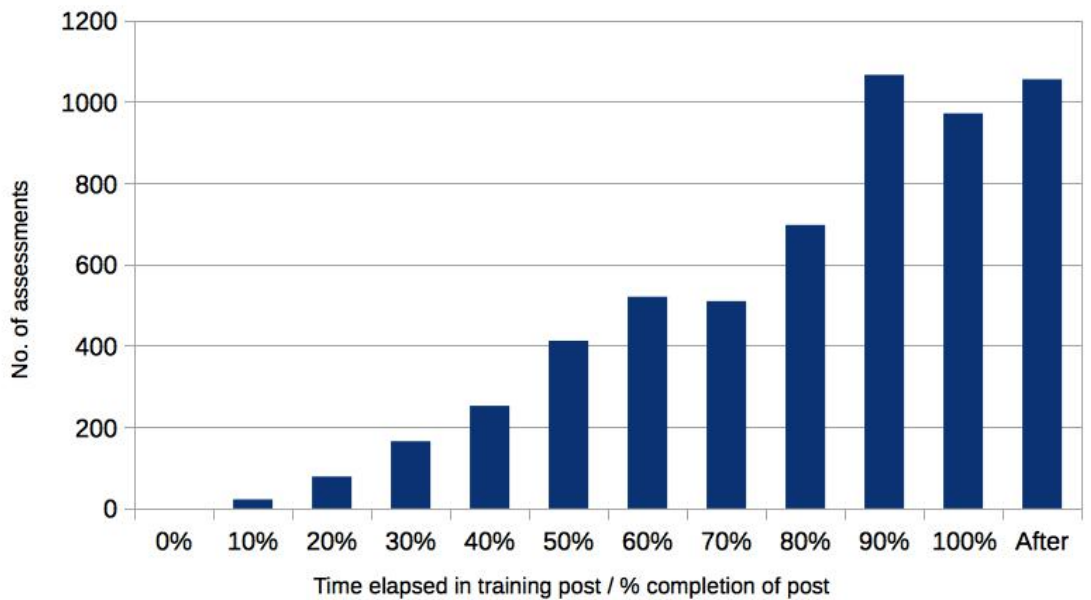
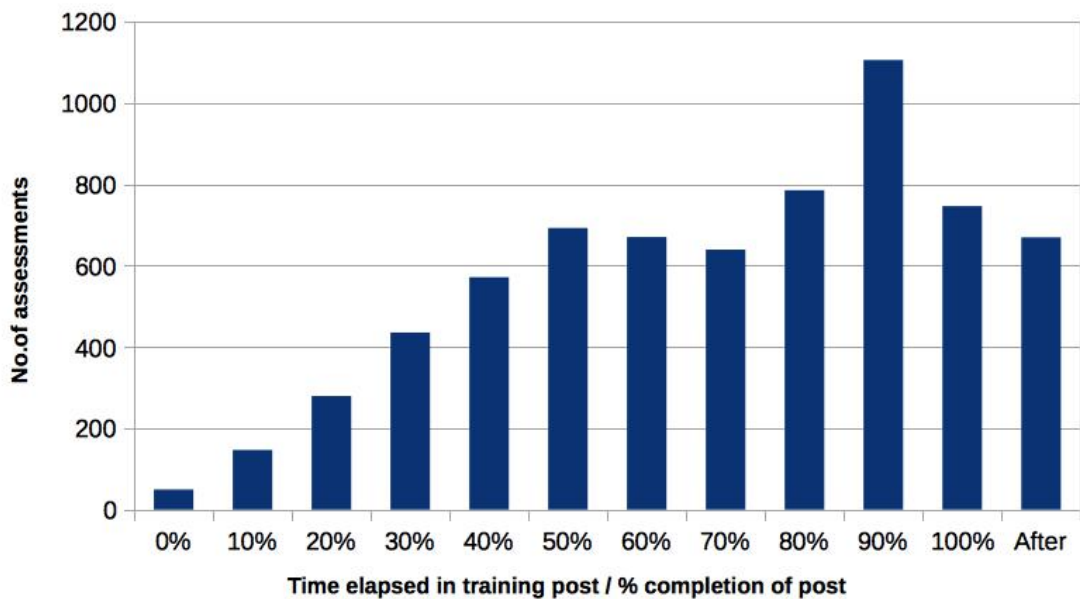




Fig 5.3. Stage within training attachment that Rad-DOPS assessments were recorded in the training e-portfolio in the training year 2012-2013. The stage is expressed as a percentage of the total time in the training attachment, rounded to the nearest 10%.



### 5.3 Content analysis of assessors' written comments

Content analysis was undertaken on a large sample (n=500) of Rad-DOPS assessments from each year of training. Table 5.3 shows the frequencies of the codes that were applied, displayed as both the total numbers of coded comments within each sample and the numbers of assessments containing each type of coded comment.

#### 5.3.1 Establishing the representativeness of the original sample of 500 assessments

The 'CHITEST' function in Excel was used to make comparisons between the coding frequencies that were found in the first and second samples of 500 feedback statements taken from the 2010-11 data. The frequencies and the associated  $p$  values that were obtained after Chi-squared analysis are displayed in Table 5.4. Given the relatively small frequencies of some codes, the level of significance was set at  $p < 0.01$ , rather than the less stringent level of  $p < 0.05$ , in order to reduce the risk of obtaining 'false positives'.

As can be seen in Table 5.4, the coding frequencies in the two samples showed significant variation in only one code:: 'assumed improvement' ( $\chi^2=7.22$ , 1, N=1000, p=0.007). The data were re-examined to ensure that inconsistent coding was not the cause of the significant variation in this case. Confirming that the 'assumed improvement' code had been applied consistently within each sample and between the two samples strongly suggests a year-on-year change. Whilst on face value this is statistically significant, the increase of 28 more comments was found in only 6% of the sample and may have no operational meaning. It is not surprising, perhaps, that assessors in the first year of a new assessment system, sampled from across the UK, showed some fluctuation in the type of feedback that they were providing. In any event, no clear reason could be identified.

Importantly, there was good agreement between the remainder of the codes that were compared. This established a sample of 500 assessments as being likely to be generally representative of the feedback found across the population of Rad-DOPS assessments recorded in the training year 2010-11. This in turn made an arguable case for the findings from subsequent years' 500 assessment samples – 2011-12 and 2012-13 – being likely to be generalisable to the respective populations. The finding that two independently-coded samples of 500 feedback statements shared significant similarities also established the second sample, which was coded in Excel, as being a legitimate surrogate for the first sample in order to conduct further statistical analysis, which was undertaken in Excel due to the previously-mentioned limitations of the MAXQDA software.

Table 5.3 Frequencies of each type of assessor comment, displayed as total frequencies within each sample and frequencies of assessments containing each type of comment.

Main codes and sub-codes	Training year 2010-11 1 <sup>st</sup> sample		Training year 2010-11 2 <sup>nd</sup> sample		Training year 2011-12		Training year 2012-13	
	No. of comments coded	No. of WBAs coded (%)	No. of comments coded	No. of WBAs coded (%)	No. of comments coded	No. of WBAs coded (%)	No. of comments coded	No. of WBAs coded (%)
Valency								
Positive	552	452 (90)	556	447 (89)	532	462 (92)	574	446 (89)
Negative	155	143 (29)	176	152 (30)	172	140 (28)	183	142 (28)
Observed performance								
Specific comment	92	76 (15)	108	73 (15)	119	78 (16)	91	67 (13)
General comment	426	349 (70)	476	366 (73)	441	371 (74)	424	362 (72)
Linked to assessment criteria	266	222 (44)	298	255 (51)	276	248 (50)	306	255 (51)
Global comment	104	104 (21)	88	83 (17)	59	52 (10)	107	97 (19)
Descriptive	27	24 (5)	14	13(3)	7	7(1)	29	25(5)
Developmental								
Specific developmental comment	37	37 (7)	43	41 (8)	37	34 (7)	41	37 (7)
Generaldevelopmental comment	101	100 (20)	110	104 (21)	97	95 (19)	107	105 (21)
Personal	58	58 (12)	62	62 (12)	81	76 (15)	70	66 (13)
Assumed improvement	48	48 (10)	79	76 (15)	79	74 (15)	53	52 (10)
Absent	14	14 (3)	14	14(3)	20	20	34	34

Table 5.4. Frequency of assessments containing different types of feedback comment found in two samples of 500 assessments taken from the overall total of 4798 Rad-DOPS assessments in 2010-11. Chi-squared analysis of the coding frequencies was used to generate *p* values.

	Code	Frequency in 1 <sup>st</sup> sample from 2010-11	Frequency in 2 <sup>nd</sup> sample from 2010-11	<i>p</i> value
Individual codes	Positive comment	452	447	0.60
	Negative comment	143	152	0.53
	Specific comment on observed performance	76	73	0.79
	General comment on observed performance	349	366	0.23
	Specific developmental comment	37	41	0.64
	General developmental comment	100	104	0.75
	Linked to assessment criteria	222	255	0.04
	Global comment	104	83	0.09
	Descriptive	24	13	0.065
	Personal comment	58	62	0.70
	Assumed improvement	48	76	<b>0.007</b>
	Absent	14	14	1.00
Combinations of codes	Specific <i>positive</i> comment on observed performance	52	41	0.23
	Specific <i>negative</i> comment on observed performance	30	40	0.22
	General <i>positive</i> comment on observed performance	331	350	0.20
	General <i>negative</i> comment on observed performance	73	83	0.38

### 5.3.2 Comparing feedback characteristics across all three training years.

Having coded samples of feedback from 2010-11, and in order to make comparisons of the feedback content across all three training years, random samples of 500 assessments from the 2011-12 and 2012-13 populations were generated and exported to Excel spreadsheets for coding. The established coding framework was used to code the assessors' feedback statements and compared across all four samples (including 2x500 from 2010-11)– in order to establish whether any significant differences between the coding frequencies existed. The results of the Chi-square analyses are displayed in Table 5.5.

In the majority of cases, variation in the frequencies of assessments containing each type of feedback comment was found to be non-significant. However, the variation within four codes – 'Assumed improvement', 'Global comment', 'Descriptive comment' and 'Absent' – was found to be significant ( $\chi^2=11.61$ , 3, N=2000,  $p=0.009$ ,  $\chi^2=23.02$ , 3, N=2000,  $p<0.001$ ,  $\chi^2=13.73$ , 3, N=2000,  $p=0.003$  and  $\chi^2=13.58$ , 3, N=2000,  $p=0.004$  respectively). The data were re-examined to ensure that inconsistent coding was not the cause of the variation. To assist in interpreting these findings, the definitions of these codings are restated here:

- 'Assumed improvement' referred to comments in which the assessor asserted that more activity on the part of the trainee, or more time, would lead to improvement;
- 'Global assessment' referred to comments that were made about the trainee's overall progress within the placement rather than focusing on their performance in a particular procedure;
- 'Descriptive comment' meant that the assessor had made a comment that was purely a non-evaluative description of the assessment encounter, without offering a judgement on the trainee's performance or providing any developmental comment;
- 'Absent' meant that no feedback had been provided by the assessor.

Table 5.5. Frequencies of assessments containing different types of feedback comment, as identified in samples of 500 Rad-DOPS assessments from 2010-11 to 2012-13. Chi-squared tests of independence between frequencies were conducted and probabilities ( $p$ ) of variation being due to chance are reported.

	Code	Frequency 2010-11 (1 <sup>st</sup> sample)	Frequency 2010-11 (2 <sup>nd</sup> sample)	Frequency 2011-12	Frequency 2012-13	$p$
Individual codes	Positive	452	447	462	446	0.30
	Negative	143	152	140	142	0.84
	<i>Specific</i> comment on observed performance	76	73	78	67	0.78
	<i>General</i> comment on observed performance	349	366	371	362	0.45
	<i>Specific</i> developmental comment	37	41	34	37	0.87
	<i>General</i> developmental comment	100	104	95	105	0.86
	Linked to assessment criteria	222	255	248	255	0.12
	Global comment	104	83	52	97	<b>&lt;0.001</b>
	Descriptive comment	24	13	7	25	<b>0.003</b>
	Personal comment	58	62	76	66	0.37
Combinations of codes	Assumed improvement	48	76	74	52	<b>0.009</b>
	Absent	14	14	20	34	<b>0.004</b>
	Specific <i>positive</i> comment on observed performance	52	41	45	44	0.67
	Specific <i>negative</i> comment on observed performance	30	40	46	28	0.09
	General <i>positive</i> comment on observed performance	331	350	337	333	0.57
	General <i>negative</i> comment on observed performance	73	83	101	80	0.11

The codes were confirmed to have been applied consistently within each sample and between the samples, thus the significant variation that existed was taken to be genuine. No clear reason for the differences observed in the first three of these particular types of comment could be identified but with no trend data year-on-year, the

sample fluctuation (assessor idiosyncrasy etc) and small cell sizes may offer simple explanations. However, the fourth code, 'Absent', followed an increasing trend throughout the three years that were analysed. This is important as it may reflect increasing disengagement on the part of the assessors and may therefore be an indication of purposeful engagement with the system being in decline.

### *5.3.3 Quality of written feedback - examples of assessors' comments across all three training years.*

As Table 5.3 shows, assessors' comments were found to differ with respect to their valency (i.e. whether they were positive or negative), their focus, their specificity and their relevance to the assessment criteria, and these aspects are explored further in the following sections. Assessor and trainee identification codes are those that were assigned by the RCR during the anonymisation process. Given the similarity of the patterns found in 2010-11, 2011-12 and 2012-13, results are reported for all three years of training together.

#### *Valency of feedback*

Of the 500 assessments analysed in each sample from 2010-11, 2011-12 and 2012-13, 446-462 (89-92%) assessment contained positive feedback comments, compared with 140-152 assessments (28-30%) which contained negative comments. Both positive and negative comments were seen to differ with respect to their focus, specificity and linkage to the assessment criteria.

#### *Focus of feedback*

The focus of feedback comments was found to vary quite substantially. The Rad-DOPS assessment is a real-time assessment of a trainee's performance in a particular clinical procedure, and of the 500 feedback statements coded, 369-391 (74-78%) statements included comments that were focused on the *observed performance* of the trainee in the particular procedure being assessed. These comments on observed performance reflected both positive and negative observations made by assessors, with the majority of these comments (95-96% in all three years) expressing positive verdicts. However, differences were observed in the specificity of these comments, as described later.

At times, assessors' comments focused on the personal attributes of the trainee, rather than the technical aspects of their clinical performance *per se*. This type of comment was found in 58-70 (12-14%) of the assessments sampled, and examples of such comments included:

Very enthusiastic and friendly trainee.

[Assessor id: 4078, 2010-11]

[Name] has showed a large interest and has had an enthusiastic approach to learning and performing ultrasounds.

[Assessor id: 0019, 2011-12]

Very professional, easy to work with, quick to learn and enthusiastic.

[Assessor id: 11784190, 2012-13]

Exceptional trainee.

[Assessor id: 10881741, 2012-13]

The 'real-time' nature of the assessment was not always accurately reflected in assessors' comments. Depending on the year of training analysed, 10-21% of assessments contained feedback comments that were categorised as *global*; in other words the assessor's comments made reference to the trainee's general progress within the attachment, rather than talking about their performance in the particular assessment event at hand. Comments that were typical of this category included:

[Name] is progressing well as a trainee in interventional radiology.

[Assessor id: 0758, 2010-11]

[Name] has made excellent progress in US guided intervention and especially in challenging cases such as the axillary US guided procedures.

[Assessor id: 4403, 2010-11]

Almost at the stage of needing indirect supervision only for an average difficulty case. Doing very well considering this is the start of your fourth year.

[Assessor id: 0075, 2011-12]

[Name] has made good progress over the three months [*sic*] time in the BID [Breast Imaging Department].

[Assessor id: 11363536, 2012-13]



Almost all of these global feedback comments were positive, with only very small numbers of assessors (none in 2010-11 or 2011-12 and only one in 2012-13) expressing a negative view of the trainee's overall progress during the attachment. The example from 2012-13 is cited below:

This assessment is based on my weekly ultrasound lists. [Name] appears to be generally competent at basic paediatric ultrasound. He however remains quite slow and struggles with acute or more complicated paediatric USS [ultrasound scanning].

[Assessor id: 11948203, 2012-13]

Another focus for assessors' comments was the future development of the trainee. Whilst WBAs are intended to be formative, developmental comments were found in only a relative minority of Rad-DOPS assessments – this type of comment was only identified in 124 to 140 (25-28%) of the assessments sampled. These developmental comments differed from the comments on the observed performance in that they were apparently intended to support the improvement of some aspect of the trainee's capability, rather than simply recording an observation about the trainee's performance. However, like the performance-focused comments, some of these developmental comments were judged to be potentially more helpful than others in terms of the specificity of the feedback given.

#### *Specificity of feedback*

As previously stated, around three quarters of the assessments sampled (74-78% depending on the year analysed) contained comments on the observed performance of the trainee. The large majority of these comments were general rather than specific in nature. General comments on the trainee's performance were found in 70-74% of assessments sampled, whereas specific comments were found in only 16-19% of assessments. The content of these general comments was such that the trainee may have required additional information or clarification in order to fully understand the meaning of the assessor. The following are examples of this type of comment:

Has a good theoretical grounding, and shows considerable maturity in his use of ultrasound and other imaging in the planning of the procedure.

[Assessor id: enMo, 2010-11]

Showed good skill in his use of ultrasound and biopsy materials during the procedure.

[Assessor id: oh435uc, 2010-11]

Has grasped the nuances of stent deployment.

[Assessor id: 11696857, 2012-13]

Phrases such as ‘good theoretical grounding’, or ‘shows considerable maturity...in the planning of the procedure’ may require further clarification in order to be genuinely meaningful. In this case, the trainee is likely to have perceived the positive regard of the assessor, but trainer and trainee may not have had a shared understanding of the phrases used – the trainee may not have been able to articulate what it was about his or her practice that the assessor perceived to be ‘mature’, for example. In a different scenario, with a less capable trainee, the same assessor might have offered little of any real educational benefit in suggesting that they develop more ‘maturity’ in planning the procedure.

As previously mentioned, a small proportion of assessments (16-19%) did contain more specific comments on the trainee’s performance:

Good technical manipulation of fluoroscopy machine with good coning and minimal exposure to obtain good quality images in patient with severe chest pain.

[Assessor id. 2028, 2010-11]

Understands the principles of CT coronary angiography and is able to correctly prescribe and administer the appropriate medication for the control of heart rate prior to the examination.

[Assessor id. 0944, 2010-11]

Has a good understanding of techniques for identifying optimal position for femoral arterial puncture (point of maximal pulsation, using anatomical landmarks inguinal ligament [and understands unreliability of anatomical crease in larger patients] & using fluoroscopy prior to puncture).

[Assessor id. 3120, 2010-11]

The needle tip was not visualised at all times and I believe he went through the collection, but managed to aspirate it fully on pulling the One Step needle back.

[Assessor id. an593ng, 2012-13]

Able to obtain good quality images and correctly excluded any significant pelvic pathology. Unfortunately, did not notice / comment on bilaterally enlarged kidneys with appearances suggestive of duplex anatomy, and although this was an incidental finding, I would not expect a post FRCR trainee to miss this.

[Assessor *id. nd101 B, 2012-13*]

In contrast to the more general comments previously cited, the assessors here have clearly identified the aspects of the trainees' performance, and any relevant aspects of the equipment, to which they are referring. In another example, the assessor [*id. 3318, 2010-11*] had provided the following specific comments: 'Radiation protection – good coning, appropriate positioning of patient and centering of image, good use of lead shield,' rather than a more general version of the same feedback such as: 'Good radiation safety awareness.'

Analysis of assessors' developmental (i.e. improvement-orientated) comments evidenced a similar dominance of general, rather than specific, comments. General developmental comments were identified in around a fifth of assessments – 19-21% depending on the year being analysed. Examples of these unspecific recommendations included:

Have patter to be able to distract patient.

[Assessor *id. 1644, 2010-11*]

Aim to feel confident in imaging supraspinatus by end of three months.

[Assessor *id. 0333, 2010-11*]

Needs more experience to learn tips/tricks for overcoming occasional difficulties (difficult puncture, wire control, etc). Think about technical factors of procedure in context of patient symptoms and desired clinical outcome.

[Assessor *id. 3342, 2010-11*]

Whilst these comments may have been meaningful to the assessor, a trainee who had developmental needs in these areas may have struggled to identify what the assessor meant by 'patter', how to 'feel confident' and which 'tips' or 'tricks' they should be learning. A learning objective that is phrased as to 'think about' is also not particularly helpful, and recommendations to gain 'more experience' assume that there is an obvious link between experience and useful learning, which may or may not exist.

Specific developmental comments were identified in fewer than 10% of assessments. In contrast to the general comments previously cited, these specific comments arguably offered a clear indication of what was required to improve performance:

Would suggest when ovaries are difficult to see, that he palpates the lower abdomen and check that ovaries are mobile with slight palpation of the transducer.

[Assessor id. 0455, 2010-11]

Must routinely scan patient in left lateral decubitus position as well as supine. Main learning: press on puncture site as the catheter is being introduced, don't do a one handed catheter removal with only limited wire in place, fully deflate balloons before reintroduction.

[Assessor id. 0729, 2010-11]

Ideally should keep sight of end of g/w and when stenting using this particular device, hold onto low-friction black 'sheath'.

[Assessor id. 0213, 2011-12]

He has a good understanding of the ultrasound equipment but needs to utilise the controls more ie. use of sector width, focal zones, TGC etc.

[Assessor id. 0295, 2011-12]

In each of these cases, the advice being offered by the assessor is clear even at some remove from the assessment encounter, and needs little decoding other than having an awareness of some of the technical terminology of the field.

In some cases, assessors offered no particular advice for improvement, instead expressing the view that continued practice would inevitably lead to improvement. Comments in this category typically stated that factors such as 'time', 'experience' or 'more procedures', would lead to improved performance. This type of comment was referred to as 'assumed improvement' in the coding framework, and comments of this nature were identified in 48-76 assessments (10-15% of the sample).

### *Linkage to assessment criteria*

Analysis revealed that a sizeable number of assessments – 44-51% of the sample – contained feedback that was linked to the assessment criteria. Interestingly, only around a fifth to a quarter of these comments (21-25%) were simultaneously coded as being specific in nature. Linkage of written feedback to the assessment criteria was therefore clearly not a guarantee of specificity.

## **5.4 Inferential statistical analysis**

### *5.4.1 The relationship between overall competence rating and the type of feedback received*

The Rad-DOPS assessment form asks assessors to provide an overall rating of the trainee's performance in the observed procedure, selected from four possibilities. These are listed below (see Figure 5.4), alongside the abbreviated term that was used during analysis in this study, and which will be used for the sake of brevity in presenting the results of this analysis. The coded data from 2010-11, 2011-12 and 2012-13 were analysed using Chi-squared in order to determine if there were significant differences between the types of feedback provided to trainees depending on the overall competence rating allocated to them by their assessors.

The number of trainees awarded the lowest rating – 'additional supervision' – in each year was too low to be validly included in statistical analysis. Therefore, comparisons were made between the second lowest rating – 'direct supervision' – and the two highest ratings – 'indirect supervision' and 'independent practice' – in turn. The results of Chi-squared comparisons between the comments provided to trainees rated as requiring direct supervision and indirect supervision, and between trainees rated as requiring direct supervision or being ready for independent practice, are displayed in Table 5.6. In each case the tables display the *p* value yielded by Chi-squared analysis performed in Excel. Bold type has been used to indicate whether each *p* value was significant at the level of  $p < 0.01$ .

Figure 5.4 Overall competence ratings that were assigned to trainees at the end of each Rad-DOPS assessment, and the abbreviated terms that were used for these in the course of the research process.

<b>Overall competence rating</b>	<b>Abbreviated term</b>
Trainee requires additional support and supervision <i>(Demonstrates basic radiological procedural skills resulting in incomplete examination findings. Shows limited clinical judgement following encounter)</i>	Additional supervision
Trainee requires direct supervision <i>(Demonstrates sound radiological procedural skills resulting in adequate examination findings. Shows basic clinical judgement following encounter)</i>	Direct supervision
Trainee requires minimal/indirect supervision <i>(Demonstrates good radiological procedural skills resulting in sound examination findings. Shows good clinical judgement following encounter)</i>	Indirect supervision
Trainee requires very little/no senior input and able to practise independently <i>(Demonstrates excellent and timely radiological procedural skills resulting in a comprehensive examination. Shows good clinical judgement following encounter)</i>	Independent practice

As can be seen from the table, the most common finding was that there was no significant difference between the frequency of most types of feedback comment provided to trainees who received different types of overall competence rating. In fact, in 2010-11 and 2011-12, 12 of the 16 types of feedback comment that were compared showed no significant variation. Other types of feedback showed significant variation, but not according to any particular pattern. Throughout all three years there were only two coding frequencies that showed consistently significant differences between trainees with different overall competence ratings: 'general developmental comment' and 'negative comment'.

Table 5.6 Results of Chi-squared analysis of comments received by trainees with different overall competence ratings in the three training years from 2010-11 to 2012-13. Results are displayed numerically as *p* values. **Bold type** indicates significant difference ( $p < 0.01$ ).

Year	Overall competence rating	No. of trainees	Positive comments	Negative comments	Specific comments on observed performance	Specific positive comments on observed performance	Specific negative comments on observed performance	General comments on observed performance	General positive comments on observed performance	General negative comments on observed performance	Specific developmental comments	General developmental comments	Global comment	Personal comment	Linked to assessment criteria	Assumed improvement	Descriptive	Absent
2010-11	Direct supervision required <i>versus</i> Indirect supervision required	214	0.165	<b>&lt;0.001</b>	0.125	0.558	0.148	0.095	0.077	0.037	0.558	<b>0.005</b>	0.105	0.047	0.126	<b>&lt;0.001</b>	0.300	0.446
		231																
	Direct supervision required <i>versus</i> Independent practice	214	0.543	<b>&lt;0.001</b>	0.118	0.255	0.196	<b>0.001</b>	0.425	0.026	0.255	<b>0.006</b>	0.219	0.785	0.309	0.041	*	*
		47																
2011-12	Direct supervision required <i>versus</i> Indirect supervision required	202	0.23	<b>&lt;0.001</b>	0.85	0.40	0.74	0.89	0.20	<b>&lt;0.001</b>	0.52	<b>&lt;0.001</b>	0.762	0.195	0.694	<b>&lt;0.001</b>	0.751	0.899
		212																
	Direct supervision required <i>versus</i> Independent practice	202	0.10	<b>&lt;0.001</b>	0.31	0.42	0.25	0.20	0.01	<b>&lt;0.001</b>	0.89	<b>&lt;0.001</b>	0.375	0.905	0.876	<b>&lt;0.001</b>	0.295	*
		73																
2012-13	Direct supervision required <i>versus</i> Indirect supervision required	162	0.045	<b>&lt;0.001</b>	<b>0.007</b>	0.328	0.015	0.365	<b>&lt;0.001</b>	<b>0.002</b>	<b>0.008</b>	<b>&lt;0.001</b>	0.702	0.391	0.993	0.067	0.935	0.660
		252																
	Direct supervision required <i>versus</i> Independent practice	162	0.929	<b>&lt;0.001</b>	0.213	0.645	0.013	<b>0.002</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.012	<b>&lt;0.001</b>	0.483	0.733	0.012	<b>&lt;0.001</b>	0.408	0.070
		76																

\*Not calculated as expected values too low for Chi-squared to be valid.

The first of these, 'general developmental comment', was significantly more likely to occur in assessments rated as 'Direct supervision' than in assessments rated as 'Indirect supervision' ( $\chi^2=12.84$ , 3, N=2000,  $p=0.005$  in 2010-11;  $\chi^2=18.15$ , 3, N=2000,  $p<0.001$  in 2011-12 and  $\chi^2=19.19$ , 3, N=2000,  $p<0.001$  in 2012-13). The same code was also found to be significantly more likely to occur in assessments rated as 'Direct supervision' than in assessments rated as 'Independent practice' ( $\chi^2=12.45$ , 3, N=2000,  $p=0.006$  in 2010-11;  $\chi^2>25.90$ , 3, N=2000,  $p<0.001$  in 2011-12 and  $\chi^2=20.40$ , 3, N=2000,  $p<0.001$  in 2012-13).

The second code, 'negative comment', was significantly more likely to occur in assessments rated as 'Direct supervision' than in assessments rated as 'Indirect supervision' ( $\chi^2=22.17$ , 3, N=2000,  $p<0.001$  in 2010-11;  $\chi^2=16.54$ , 3, N=2000,  $p<0.001$  in 2011-12; and  $\chi^2>25.90$ , 3, N=2000,  $p<0.001$  in 2012-13). The code was also found to be significantly more likely to occur in assessments rated as 'Direct supervision' than in assessments rated as 'Independent practice' ( $\chi^2=19.27$ , 3, N=2000,  $p<0.001$  in 2010-11;  $\chi^2>25.90$ , 3, N=2000,  $p<0.001$  in 2011-12 and  $\chi^2>25.90$ , 3, N=2000,  $p<0.001$  in 2012-13).

Although it is arguably predictable that negative or developmental commentary might be associated with a judgement that a trainee needs direct supervision, a degree of qualification is necessary. Firstly, the frequencies of comments that were being compared in each case were low – in fact, the more common outcome for trainees of any level of performance was that they received no negative feedback on their performance and no suggestions for improvement. Secondly, the significance of the differences observed in 'negative' feedback disappeared when looking at the frequency of *specific* (as opposed to general) negative comments awarded to each group. Thirdly, these general negative comments were often *very* general, including phrases such as 'technique is currently a bit unrefined'.

Similarly, the general developmental comments provided by assessors were often very general, often being limited to comments such as 'get more practice', 'see more patients', 'learn tips and tricks' and so on, and so were unlikely to provide much helpful information for these trainees. Only in the most recent training year analysed – 2012-13



– were trainees rated ‘Direct supervision’ significantly more likely than their colleagues who were rated ‘Indirect supervision’ to receive specific guidance on how to improve ( $\chi^2=11.83$ , 3, N=2000, p=0.008). It may be the case, therefore, that as the WBA system becomes established assessors are beginning to recognise the importance of providing specific developmental comments to the trainees who are most in need of their support. That said, it is important to beware over-interpreting this result. It is an isolated finding which was not replicated when comparing trainees in need of direct supervision with trainees who were capable of independent practice.

#### *5.4.2 The relationship between modal score and the type of feedback received*

The sample data were analysed to test for significant differences between the types of feedback provided to trainees versus the modal score for each assessment, which was calculated from the scores allocated to each assessment criterion by assessors. The modal score for each assessment was calculated as the assessor’s overall competence rating, phrased as a comment about readiness for independent practice, may not relate well to the scores given for individual assessment criteria. This is because these criteria are scored with reference to stage of training, rather than by comparison with readiness for independent practice. For example, a very junior trainee may be scored highly for their performance relative to their stage of training, whilst being some way off readiness for independent practice.

An unexpected finding that emerged through these calculations was that very few trainees received modal scores of three or lower. Only 15 assessments, recorded by 14 different trainees, were found to have a modal score of three in 2010-11 (N=4798 assessments). None had a modal score of less than three. In 2011-12, 50 assessments, recorded by 35 different trainees, had a modal score of three (N=13). Only seven assessments, recorded by seven different trainees, had a modal score of two. One trainee had a modal score of one, but the very positive written feedback, the overall competence rating of ‘minimal supervision’ and the trainee’s relative seniority (ST4) suggested that the assessor had misinterpreted the scoring system. In 2012-13, 23 assessments, undertaken by 20 different trainees, had a modal score of three (N=5). Seven assessments, recorded by four different trainees, had a modal score of two. Three trainees had each recorded an assessment in which the modal score

was 1. However, from the feedback provided by the assessors only one of these appeared to be genuine, with the other two appearing to be misinterpretations of the scoring scale.

These results suggest that despite the potential of the assessments to be used to block trainee progression, most assessors display leniency when scoring trainees. Even a relatively low score of three out of six represents the fairly moderate judgement of 'borderline for stage of training', with scores of two or one being equivalent to 'below expectation' or 'well below expectation' respectively. As a result of their very low numbers, scores of one to three out of six were grouped together with scores of four out of six for further analysis.

Comparisons were made between trainees who had the lowest modal scores ( $\leq 4$ ) and those who had the highest possible modal score (6), and between trainees who had the lowest scores and those who had the second highest possible modal score (5). The results of these calculations can be seen in Table 5.7. The picture that emerged through this analysis was much less consistent than that observed when comparing trainees with different overall competence ratings. In fact, few patterns either within years or between years were apparent. That said, one relatively consistent finding was that significant differences between the frequencies of negative comments were apparent when comparing the lowest scoring assessments (those with a modal score  $\leq 4$ ) with the highest scoring assessments (those with a modal score =6):

$\chi^2=17.38$ , 3, N=2000,  $p<0.001$  in 2010-11;

$\chi^2=23.02$ , 3, N=2000,  $p<0.001$  in 2011-12;

$\chi^2=12.11$ , 3, N=2000,  $p=0.007$  in 2012-13.

These low scoring trainees were more likely to receive negative feedback on their observed performance than their high scoring counterparts. It is arguably predictable that this would be the case, but inspection of the raw data showed the more frequent outcome for these trainees was actually that they received no negative comments on their performance. For example, of the 262 trainees in 2010-11 who had a modal score of  $\leq 4$ , 91 received negative feedback comments compared with 171 who did not.

Table 5.7 Results of Chi-squared analysis of comments received by trainees with different modal scores in the three training years from 2010-11 to 2012-13. Results are displayed numerically as *p* values. **Bold type** indicates significant difference ( $p < 0.01$ ).

Year	Modal score	No. of trainees	Positive comments	Negative comments	Specific comments on observed performance	Specific positive comments on observed performance	Specific negative comments on observed performance	General comments on observed performance	General positive comments on observed performance	General negative comments on observed performance	Specific developmental comments	General developmental comments	Global comment	Personal comment	Linked to assessment criteria	Assumed improvement	Descriptive	Absent
2010-11	Modal score $\leq 4$ versus Modal score =5	262	0.435	0.141	0.115	0.275	0.154	0.015	<b>0.003</b>	0.197	0.222	0.892	0.355	0.011	0.228	0.546	0.258	0.670
		199																
2010-11	Modal score $\leq 4$ versus Modal score =6	262	0.775	<b>&lt;0.001</b>	0.014	0.133	0.036	<b>&lt;0.001</b>	0.217	<b>0.005</b>	0.036	0.036	0.779	0.677	0.509	0.578	0.607	*
		39																
2011-12	Modal score $\leq 4$ versus Modal score =5	313	0.079	0.014	0.309	0.916	0.225	0.571	0.204	0.017	0.114	0.087	0.833	<b>&lt;0.001</b>	0.015	0.027	*	0.109
		146																
2011-12	Modal score $\leq 4$ versus Modal score =6	313	0.323	<b>&lt;0.001</b>	<b>0.003</b>	0.035	0.024	<b>&lt;0.001</b>	0.160	<b>0.001</b>	0.050	<b>0.002</b>	0.438	*	0.360	<b>0.009</b>	*	*
		41																
2012-13	Modal score $\leq 4$ versus Modal score =5	279	<b>0.006</b>	<b>&lt;0.001</b>	0.021	0.997	<b>&lt;0.001</b>	0.403	<b>&lt;0.001</b>	<b>0.005</b>	<b>0.003</b>	<b>0.002</b>	0.885	<b>0.008</b>	0.770	0.168	0.731	0.831
		170																
2012-13	Modal score $\leq 4$ versus Modal score =6	279	0.034	<b>0.007</b>	0.932	0.193	0.073	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.074	0.138	0.124	*	<b>0.008</b>	0.178	0.447	*	*
		50																

\*Not calculated as the numbers were too low for Chi-squared to be valid.

#### *5.4.3 The relationship between stage of training and the type of feedback received*

The frequencies of different types of feedback comment provided to trainees at different stages of training were also analysed to look for significant differences based on seniority. The results are displayed in Table 5.8. The pattern in this case was one of generally non-significant differences between core trainees – i.e. those in Specialty Training years one three (ST1 to ST3) – and those who were in Higher Specialty training (ST4 to ST6). Only one code – ‘linked to the assessment criteria’ – showed a statistically significant difference in one year, indicating that assessors were more likely to link their comments to the published assessment criteria if the trainee being assessed was in core training rather than higher training:  $\chi^2=7.03, 1, N=500, p=0.008$  in 2010-11.

This finding was not replicated in any of the subsequent years. In searching for a reason, the raw coding frequencies of the ‘linked to assessment criteria’ code in 2011-12 and 2012-13 were inspected. The data revealed that in these latter two years the number of comments linked to the assessment provided to core trainees decreased, while the number of these comments provided to higher trainees increased. It is tempting to interpret the increase in ‘linked’ comments being provided to senior trainees as being a sign of assessors getting to grips with the formative function of the WBAs, however this does not explain the decline in this type of comments being provided to junior trainees.

#### *5.4.4 The relationship between length of feedback and the type of feedback received*

Finally, the frequencies of different types of comments were analysed with respect to the length of the overall feedback statement provided by the assessor. Results are displayed in Table 5.9. On this occasion, significant differences were found in most cases, demonstrating formally what was apparent via informal inspection of the data – that very brief comments (less than 10 words) were unlikely to contain feedback that met the more exacting standards of feedback quality e.g. ‘specific developmental’ comments or ‘specific comments on observed performance’.

Table 5.8 Results of Chi-squared analysis of comments received by trainees in core training and higher training. Results are displayed numerically as *p* values. **Bold type** indicates significant difference ( $p < 0.01$ ).

Year	Stage of training	No. of trainees	Positive comments	Negative comments	Specific comments on observed performance	Specific positive comments on observed performance	Specific negative comments on observed performance	General comments on observed performance	General positive comments on observed performance	General negative comments on observed performance	Specific developmental comments	General developmental comments	Global comment	Personal comment	Linked to assessment criteria	Assumed improvement	Descriptive	Absent
2010-11	Core (ST1-3)	390	0.967	0.081	0.122	0.235	0.265	0.067	0.034	0.431	0.427	0.302	0.881	0.030	<b>0.009</b>	0.562	*	*
	Higher (ST4-6)	110																
2011-12	Core (ST1-3)	377	0.893	0.881	0.931	0.439	0.890	0.145	0.311	0.999	0.864	0.892	0.903	0.417	0.083	0.677	*	0.302
	Higher (ST4-6)	124																
2012-13	Core (ST1-3)	324	0.809	0.079	0.201	0.153	0.660	0.828	0.492	0.094	0.199	0.559	0.292	0.787	0.014	0.016	0.576	0.419
	Higher (ST4-6)	173																

*\*Not calculated as the expected values involved were too low for Chi-squared to be valid.*

Table 5.9. Results of Chi-squared analysis of types of comment present in feedback of different lengths. Results are displayed numerically as *p* values. **Bold type** indicates significant difference ( $p < 0.01$ ).

Year	Length of feedback	No. of trainees	Positive comments	Negative comments	Specific comments on observed performance	Specific positive comments on observed performance	Specific negative comments on observed performance	General comments on observed performance	General positive comments on observed performance	General negative comments on observed performance	Specific developmental comments	General developmental comments	Global comment	Personal comment	Linked to assessment criteria	Assumed improvement	Descriptive
2010-11	Brief ( $\leq 10$ words)	187	<b>0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.037	0.240	<b>&lt;0.001</b>	<b>&lt;0.001</b>	*
	Extended ( $> 10$ words)	313															
2011-12	Brief ( $\leq 10$ words)	198	0.12	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.173	0.802	0.011	0.060	*
	Extended ( $> 10$ words)	302															
2012-13	Brief ( $\leq 10$ words)	181	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.812	0.619	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>0.003</b>
	Extended ( $> 10$ words)	319															

\* Not calculated as the expected values involved were too low for Chi-squared to be valid.

## 5.5 Qualitative comparative analysis – Ragin's approach

As discussed in Chapter 4, it is quite challenging to meet the demands of conventional statistical analysis when analysing naturalistic data sets. Consequently, I decided to use a more flexible approach to statistical analysis to explore similar relationships to those which had previously been analysed using Chi-squared analysis. However, rather than looking at the relationships between certain conditions (such as modal score or stage of training) and the incidence of different types of individual *comment*, I was able to look at the conditions that would lead to high quality feedback *statements* – i.e. feedback that was composed of a combination of comments that had been theoretically judged to comprise high quality formative feedback.

The analysis chosen was that described by Ragin (1987, 2000, 2008), qualitative comparative analysis (QCA), which is useful for determining the sufficient and necessary conditions required to produce a given outcome. This outcome has to be selected, or constructed, prior to commencing analysis. For example, if one were interested in exploring the conditions that might contribute to 'educational success', an indicator of educational success must first be chosen in order to commence the investigation. One might choose, say, 'achievement of an undergraduate university degree', as an indicator of a certain type or level of educational success. This then becomes the 'outcome', with a number of potential contributing factors, known as 'conditions' or 'predicates', being considered mathematically in order to determine if they are connected with the chosen outcome. Analysis focuses on determining whether particular conditions appear to be necessary, or sufficient, in order to achieve the outcome of interest. In this hypothetical example, potential conditions for the achievement of a university degree might include demographic factors (e.g. ethnicity, gender), socioeconomic factors (e.g. employment status of parents, eligibility for free school meals) and educational factors (e.g. attendance at either a selective or non-selective school).

In my study, the outcome of interest was 'high quality feedback'. This is less easy to specify objectively than an outcome such as 'possession of a university degree', however a tentative model of 'high quality feedback' was determined qualitatively from the review of literature as described in Chapter 4.

### 5.5.1 Exploring average assessment scores as a condition for high quality feedback.

As used in the Chi-squared analysis above, an average (modal) score was calculated for each assessment undertaken by trainees in each of the three training years. Ragin's QCA was used to analyse the relationship between the different levels of modal assessment score in individual assessments and the provision of high quality feedback as criterially defined above (see Table 5.10). As explained in Chapter 4, the threshold for declaring necessity or sufficiency is 80% coverage, or overlap, between the condition and the outcome. The threshold for declaring quasi-necessity or quasi-sufficiency is 70%. My analysis revealed that low modal score on an individual assessment – i.e. a modal score of  $\leq 4$  out of 6 – clearly exceeded the 70% threshold for declaring quasi-necessity in two out of three training years and came close to this threshold (66%) in the remaining year. The numbers involved are small, but there is evidence here for low modal assessment score functioning as a necessary condition for high quality feedback.

However, low modal score was not found to be a sufficient condition for high quality feedback. As shown in Table 5.11, only a small proportion of the assessments that had a modal score of  $\leq 4$  also contained high quality feedback. The proportion fell well below the 70% threshold for quasi-sufficiency, and this pattern was consistent across all three years of the study.

Table 5.10 Proportion of assessments adjudged to contain high quality feedback, distributed across assessments of different modal assessment score ( $\leq 4$ , 5 or 6). These proportions were used to determine necessity, or quasi-necessity.

Modal assessment score	High quality feedback					
	2010-11 N=41		2011-12 N=34		2012-13 N=37	
	n	%	n	%	n	%
All modal scores	41	100	34	100	37	100
$\leq 4$	27	66*	27	79**	30	83***
5	14	34	7	21	5	14
6	0	0	0	0	2	5

\* Falls short of threshold for quasi-necessity (70%)

\*\* Exceeds threshold for quasi-necessity, and falls just short of threshold for necessity (80%)

\*\*\* Exceeds the threshold for necessity (80%)



Table 5.11 Proportion of assessments with a modal score of  $\leq 4$  which contained high quality feedback. Proportions are used to judge sufficiency of low modal assessment score for the provision of high quality feedback.

Year	2010-11	2011-12	2012-13
Assessments with a modal score of $\leq 4$ (N)	262	309	279
Number containing high quality feedback (n)	27	27	30
Proportion containing high quality feedback (%)	10	9	11

A further finding was that high quality feedback comments seemed to be less likely to occur when high numerical scores are awarded and, in particular, were almost never found assessments with a modal score of 6 (see Table 5.10). In the language of Ragin's approach to QCA, a high modal assessment score could be deemed to be a NOT function. Thus, Ragin's logic supports the idea that the awarding of high assessment scores was in some way working against the provision of high quality feedback.

### 5.5.2 Exploring stage of training as a necessary condition for high quality feedback

Feedback statements from all three years were analysed in order to determine if stage of training could be deemed to be necessary for the provision of high quality feedback, based on the postulation that inexperienced trainees may be more likely than their more experienced colleagues to be given specific developmental feedback by their assessors, regardless of their performance in a particular assessment.

In Table 5.12 none of the proportions was close enough to the 70% threshold for any one stage of training to be deemed necessary for the occurrence of high quality feedback but combining the figures for ST1-ST3 resulted in the 70% threshold for quasi-necessity being met. This combination is more than simply mathematically

convenient – these three training years together comprise ‘core radiology training’ as defined by the RCR (2010, p. 173). Therefore, a trainee being in core radiology training (ST1-ST3) was a quasi-necessary condition for the provision of high quality feedback.

So few examples of high quality feedback were found in the assessments of the most senior trainees (ST5 and ST6) that seniority might be considered to be a NOT condition for the provision of high quality feedback. Consequently, it may be argued that senior trainees are systematically being excluded from the formal process of formative workplace assessment.

Table 5.12 Proportion of assessments adjudged to contain high quality feedback that were present at each stage of training (ST1-ST6). Shaded cells indicate core training.

Year of training	High quality feedback					
	2010-11 N=41		2011-12 N=34		2012-13 N=37	
	n	%	n	%	n	%
All years	41	100	34	100	37	100
ST1	14	34	18	53	12	32
ST2	12	29	7	21	10	27
ST3	9	22	1	3	5	14
ST4	6	15	4	12	8	22
ST5	0	0	4	12	2	5
ST6	-	-	0	0	0	0

Analysis of the data shown in Table 5.13 revealed that being in core radiology training was not a sufficient condition for the provision of high quality feedback, as the proportion of assessments undertaken by core trainees that also contained high quality feedback fell well short of the 70% threshold for quasi-sufficiency.

Table 5.13 The number of assessments recorded by core trainees (ST1-ST3), and the proportion of these (number and percentage) which contained high quality feedback.

	Training year 2010-11	Training year 2011-12	Training year 2012-13
Number of assessments recorded by core radiology trainees (ST1-ST3) (N)	390	377	324
Number containing high quality feedback (n)	35	26	27
Proportion containing high quality feedback (%)	9	7	8

### 5.5.3 Exploring overall competence ratings as necessary conditions for high quality feedback

Analysis of the data suggested that none of the overall competence ratings was likely to be a necessary condition for the provision of high quality feedback (see Table 5.14). There was evidence that the highest overall rating – ‘independent practice’ – was a NOT condition.

Table 5.14 Proportion of assessments adjudged to contain high quality feedback across three levels of overall competence rating.

Overall competence rating	High quality feedback					
	2010-11 N=41		2011-12 N=34		2012-13 N=37	
	n	%	n	%	n	%
All	41	100	34	100	37	100
Direct supervision	21	51	14	41	21	57
Indirect supervision	18	44	16	47	14	38
Independent practice	2	5	4	12	2	5

As with the two previously-analysed conditions (modal assessment score and stage of training), none of the overall ratings analysed was found to be sufficient for the occurrence of high quality feedback.

The number of trainees in each of the three years who received the lowest overall rating – ‘trainee requires additional support and supervision’ – was small. Analysis of samples of 500 assessments from all three years revealed that 8, 13 and 8 assessments were found to contain these judgements in 2010-11, 2011-12 and 2012-13 respectively. These numbers were too small to be suitable for further analysis using QCA methodology, hence they do not feature in Table 5.14. However, it is interesting to note that only one of these assessments in 2010-11 contained specific developmental comments as to how the trainee could improve. Furthermore, none of the assessments contained feedback that satisfied all of the criteria for ‘high quality feedback’, in spite of the overall rating indicating that additional support was deemed necessary. Of the 13 assessments in 2011-12 that identified the trainee as needing ‘additional support and supervision’, only two contained specific developmental comments, with none containing feedback that satisfied all of the criteria for high quality feedback. In 2012-13, eight of the 500 assessments were found to have a global judgement of ‘trainee requires additional support and supervision’. None of these contained specific developmental comments, and thus none contained feedback that satisfied the criteria for high quality feedback.

This failure to provide explicit developmental feedback to the trainees who were in greatest need of support is easily overlooked, given the relatively small numbers of assessments involved. Yet it is an important finding in the analysis of a formative assessment system, the backdrop of which includes the RCR’s declaration that ‘feedback on performance is essential for successful work-based experiential learning’ (RCR, 2010, p. 156).

#### *5.5.4 Exploring length of feedback as a necessary condition for high quality feedback*

Analysis of the data for length of feedback suggested that ‘brevity’ – as indicated by feedback of 10 words or fewer – may be a NOT condition for the provision of high quality feedback. Conversely, ‘extended feedback’ (which contained 11 or more words)

was found to be a necessary condition for the provision of specific developmental comments (see Table 5.15).

Table 5.15 Proportion of assessments adjudged to contain high quality feedback that were present in brief and extended feedback statements.

Length of feedback	High quality feedback					
	2010-11 N=41		2011-12 N=34		2012-13 N=37	
	n	%	n	%	n	%
All	41	100	34	100	37	100
Brief feedback	0	0	0	0	1	3
Extended feedback	41	100	34	100	36	97

Table 5.16 The number of assessments containing extended feedback (>10 words), and the proportion of these (number and percentage) which contained high quality feedback.

	Training year 2010-11	Training year 2011-12	Training year 2012-13
Number of assessments containing extended feedback (N)	313	303	319
Number of these assessments containing high quality feedback (n)	41	34	36
Proportion containing high quality feedback (%)	13	11	11

Further analysis of the assessments containing extended feedback demonstrated that extended feedback fell some way short of qualifying as a sufficient condition for the provision of high quality feedback (see Table 5.16). In other words, a large proportion of the assessments that contained extended passages of feedback failed to deliver feedback of the highest quality.

### 5.5.5 Summary of Ragin analysis results

Ragin's approach to identifying necessary and sufficient conditions for the achievement of a particular outcome allowed for the exploration of relationships which may not meet the strict demands of traditional statistical analysis. In my research, the following were established as being necessary or quasi-necessary conditions for the provision of high quality written feedback:

- Low modal assessment score ( $\leq 4$ )
- Low overall competence rating (as assigned by assessors at the end of each assessment)
- The trainee being in core radiology training (ST1-ST3)
- Extended feedback (ie that which was in excess of 10 words)

Conversely, some conditions appeared to positively work against the provision of high quality feedback. These were tentatively proposed to be NOT conditions, and included:

- High modal assessment score (6/6)
- High overall competence rating ('Independent practice')
- Seniority – trainees at ST5 & ST6 levels received little or no high quality feedback
- Brief feedback (10 words or fewer)

Further analysis of the necessary, or quasi-necessary, conditions highlighted above demonstrated that none was sufficient to generate the desired outcome of high quality formative feedback. For example, whilst it was necessary for trainees to receive a low overall competence rating in order to receive high quality feedback, it is clear that the former does not guarantee the latter. Indeed, a large number of trainees who received low overall competence ratings failed to receive formative feedback that satisfied the criteria for high quality. The same was true for trainees who received a low modal assessment score. It was also the case that whilst assessors needed to provide reasonably lengthy feedback in order for the feedback to be adjudged to be 'high quality', a large proportion of the lengthier feedback statements failed to meet the criteria for high quality feedback. For example, upon first inspection the following statement appears useful:

Very enthusiastic and friendly trainee. Showing promise in his ability to perform fluroscopy [*sic*]. He is still early in his GI block with limited experience so far; I am sure he will gain a high standard by the end of June. Areas to improve are control of C-arm and try to concentrate on clinical question being asked and apply this to your study technique. I have little doubt that these minor faults will be ironed out over the coming weeks.

[Assessor id: 4078, 2010-11]

However, closer scrutiny reveals that it begins with a comment that directs the trainee's attention towards themselves as a person ('personal feedback'), and continues with a comment about the trainee's progress to date ('global feedback'). There is then a 'general developmental' comment about needing to improve control of the equipment, with no indication of how this should be done. The statement also includes two statements of assumed improvement: 'I am sure he will gain a high standard by the end of June' and 'I have little doubt that these minor faults will be ironed out over the coming weeks'. The statement is perhaps intended to be an expression of confidence in the trainee, but there is very little of any real instructional value here: there are no specific comments, either positive or negative, on the trainee's performance in the procedure that was being observed, and there are no specific developmental comments setting out how further improvement could be achieved.

## **5.6 Analysing trainee comments**

Trainee comments were analysed in order to provide an insight into the nature of the written feedback conversation and hence support an objective consideration of the degree to which the written feedback process in Rad-DOPS assessment was dialogical. According to Nicol (2009), 'while the quality of teacher comments is important, engagement with and use of those comments by students is equally important.' Nicol (2009) does not specify precisely what this engagement should look like, but he does offer a number of suggestions as to how engagement might be supported, all of which require face-to-face dialogue with or between learners. Thus when writing about creating more dialogic written feedback (*ibid.*), his recommendations regarding the written feedback are actually almost wholly centred on various facets of what he terms the 'quality of teacher comments' (Nicol, 2009, p. 1). No existing framework for judging 'engagement' in written feedback through the analysis of learners' written comments was found during the review of literature for this research. Therefore, the trainee

comments in my study were coded inductively and subsequently analysed for signs of engagement with assessors' comments.

#### *5.6.1 Themes emerging from the analysis of trainee comments*

The trainee responses fell into a number of themes, which are displayed in Figure 5.5. These themes were identified inductively, and were used to develop a coding framework which is shown in Table 5.17.

##### Non-engagement and minimal engagement

The most notable feature of the data in all three years was the high incidence of minimal engagement with assessors' feedback, which included trainees making no comment at all. Minimal engagement, in the form of trainees making cursory comments such as 'thanks', 'agree', or 'as above', plus non-engagement, in which trainees made no comment, accounted for 189/500, 299/500 and 354/500 of the trainee responses sampled from 2010-11, 2011-12 and 2012-13 respectively. Interestingly, statistical analysis (Chi-squared) confirmed that the differences in these frequencies was highly significant ( $\chi^2=114.66$ , 2, N=1500,  $p<0.001$ ). This represents a substantial decline in trainee engagement over the first three years of the programme.

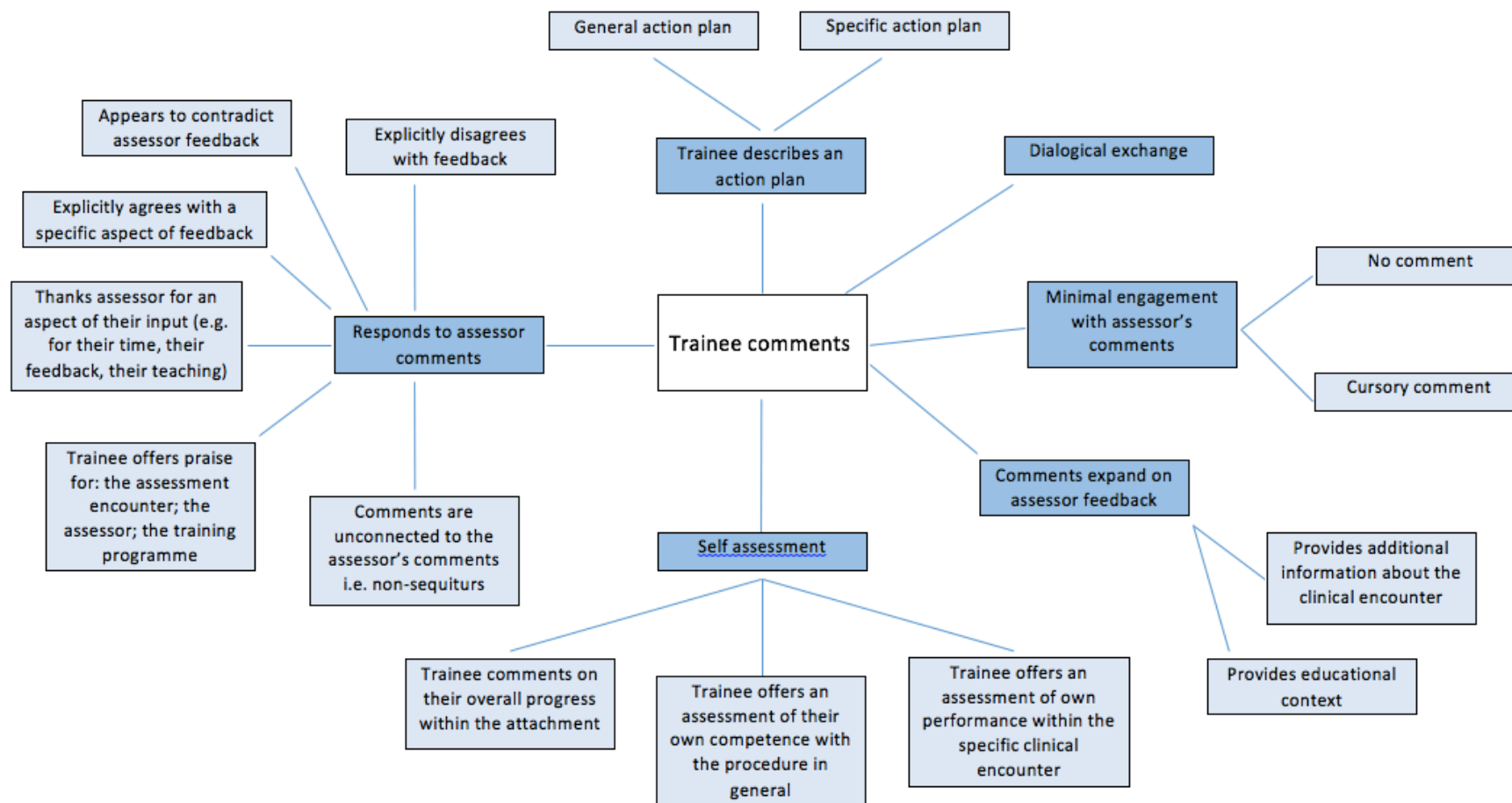
When trainees did respond at greater length they did so in a number of different ways. A number of trainees in every year responded largely to express their appreciation for some aspect of the educational process (60/500 in 2010-11, 42/500 on 2011-12 and 15/500 in 2012-13). This tended to involve praising the assessor for the quality of their instruction, praising the utility of the specific assessment encounter or praising the educational value of the placement. Again, Chi-squared analysis revealed that the declining trend was significant ( $\chi^2=28.53$ , 2, N=1500,  $p<0.001$ ).



Table 5.17 Coding framework for the analysis of trainee comments, which was developed inductively and then applied to the samples of 500 assessments from 2010-11, 2011-12 and 2012-13.

<b>Code</b>	<b>Conditions for applying the code</b>
<b>No comment</b>	Trainee makes no comment, usually indicated by the typed addition of a placeholder such as '.' or 'x'.
<b>Cursory comment</b>	Comments that were limited to single words or very short phrases e.g. 'thanks', 'agree', or 'as above'.
Limited to 'as above' or equivalent	The trainee responds with nothing more than this phrase, or its equivalent.
Limited to 'agree' or equivalent	The trainee responds with nothing more than this phrase, or its equivalent.
Limited to 'thanks' or equivalent	The trainee responds with nothing more than this phrase, or its equivalent.
<b>Trainee response more detailed than the assessor's comments</b>	
Provides additional information	Makes a comment that goes beyond the comments made by the assessor in some way.
Provides context	Sets the assessment in context that was not clear from the assessor's comments e.g. 'this was my first attempt at this type of procedure'.
<b>Self assessment</b>	
Comments on own perceived level of competence	Makes a comment on how proficient they are at this type of procedure in general.
Comments specifically on own performance in the assessment	Makes a comment on how they think they performed in the particular procedure that was being assessed.
Comments on own progress within the attachment	Makes a global comment about progress to date
<b>Connection between trainee's comment and the assessor's comment</b>	
Trainee thanks the assessor	Trainee thanks the assessor for their input, whether for the feedback, conducting the assessment, or conducting clinical teaching
Trainee explicitly agrees	Trainee makes a brief or extended comment expressing agreement with the assessor's feedback.
No obvious connection/ non-sequitur	The trainee's comment does not appear to be aligned with the apparent subject or focus of the assessor's feedback.
Contradictory	The trainee's comments appear to contradict the assessor's feedback in some regard.
Explicitly disagrees	The trainee states candidly that they disagree with the assessor's feedback.
<b>Action plan</b>	
General action plan	Trainee states what they will do in order to improve, but in general terms e.g. doing more procedures, doing more reading, getting more experience.
Specific action plan	Trainee describes specific actions that they will take in order to improve their performance e.g. I will practise following the track of the needle after the biopsy gun has fired to ensure an adequate sample has been taken from the region of interest.
<b>Dialogical feedback exchange</b>	The application of this code requires the assessor's comments to satisfy the 'high quality feedback' criteria. The trainee's comment must then be an explicit response to this, which goes beyond cursory thanks or agreement.

Figure 5.5 Themes identified in trainee responses to assessors' comments.



The following are some examples of this type of feedback, with my own identification codes applied:

Excellent teacher. Always available to help, and keen to teach.

*[Trainee id. ei388up, 2010-11]*

I have learned loads and really enjoyed these sessions with Dr [Name]. Thank you!

*[Trainee id. ar517ar, 2010-11]*

Good learning experience and trainer - supportive but allows for a degree of autonomy.

*[Trainee id. Da500l, 2011-12]*

Very good attachment – good exposure to cases.

*[Trainee id. ae588Ja, 2011-12]*

These responses suggest that trainees perceived the educational input they had received to be valuable. However, the comments are not focused on any learning points arising out of the assessment encounter, and in each of these examples the entire trainee response was limited to the comments reported here.

### Self-assessment

A number of trainees in each year (94/500 in 2010-11, 52/500 in 2011-12 and 44/500 in 2012-13) proffered an assessment of their own performance or capabilities. This included articulating their view of how they had performed in the particular clinical encounter that had been the focus of the Rad-DOPS assessment. The following examples illustrate this category of trainee comment:

Happy with how case went, need to regularly perform a greater volume of cases.

*[Trainee id: av674Ed, 2012-13]*

Tricky case, found it tough to get in into a previous endarterectomy site. Getting over the bifurcation was challenging, but satisfying.

[Trainee id: uy285ck, 2012-13]

At other times, trainees volunteered an assessment of their competence in performing the procedure in general, for example:

My confidence for performing ultrasound guided tunneled [sic] chest drain has improved significantly. My technique will become better with more practice.

[Trainee id: um663Gu, 2010-11]

I have taken better images with more experience, and am developing a systematic way of carrying out the procedure.

[Trainee id: ur142Na, 2011-12]

I feel relatively confident in performing barium swallows.

[Trainee id: ob393 T, 2012-13]

Some trainees gave an indication of how they felt they had been progressing overall throughout the placement:

Happy with progress to date, increasing confidence with general and paediatric ultrasound.

[Trainee id: ar102t, 2010-11]

Have learnt a lot and really enjoyed the placement so far. Need further practice remembering to collamate [sic] and change machine settings prior to positioning patients before imaging.

[Trainee id: im782LI, 2010-11]

The decline in the numbers of trainees who offered a self-assessment over the first three years of the programme was significant ( $\chi^2=26.08$ , 2, N=1500,  $p<0.001$ ) and is arguably another indication of diminishing trainee engagement over time.

### Reflective comments

A small number of trainees in 2010-11 and 2011-12 (12/500 and 16/500 respectively) had reflected on their learning from the specific case at the heart of the Rad-DOPS, and at times appeared to have made some broader meaning of the encounter. The following examples serve to illustrate this type of reflective response:

Ultrasound has a crucial role in imaging the paediatric population. I am also able to appreciate the complexities and challenges of imaging children, particularly in dynamic studies and in the future I look to improve my technique and approach.

*[Trainee id: an324ha, 2011-12]*

I have a greater appreciation of the complexity of such work - not all procedures are straight forward.

*[Trainee id: nd668 I, 2011-12]*

This kind of reflective comment was not seen at all in the 500 assessments sampled from the 2012-13 population and, taken together with the previously mentioned declining frequency of self-assessment comments and increasing incidence of cursory or absent trainee comments, arguably validates the notion that trainees are disengaging with the formative aspects of the WBA system.

### Action planning

A number of assessments in every year showed evidence of trainees making an action plan for further learning as a result of the assessment (128/500 in 2010-11, 96/500 in 2011-12 and 88/500 in 2012-13). In the majority of cases (94/128 in 2010-11, 80/96 in 2011-12 and 69/88 in 2012-13), this was expressed in very general terms, with common intentions being to 'see more patients', 'get more experience', 'increase numbers of procedures' and so on. Less often, trainees made specific action plans, such as:

[I need to] try and mention current findings in comparison with previous imaging. If images unclear try and scan in different planes.

*[Trainee id: he401j, 2010-11]*

I feel that I have learnt the basic sequences required, but aim to concentrate more on dynamic assessment of images in order to produce accurate pictures.

[Trainee id: a1175 P, 2010-11]

The decrease in the number of trainee self-assessments was significant ( $\chi^2=10.88$ , 2, N=1500,  $p=0.004$ ). Analysis of a subset of the self-assessment data – the numbers of trainees making *specific* action plans – revealed a noticeable decrease after the first year of the programme (34/500 in 2010-11, 16/500 in 2011-12 and 19/500 in 2012-13), although the probability yielded by the Chi-squared test ( $\chi^2=8.48$ , 2, N=1500,  $p=0.014$ ) fell short of significance at the  $p<0.01$  level.

#### Alignment between assessors' and trainees' comments

In analysing pairs of assessor statements and corresponding trainee responses, it was clear that there was sometimes a lack of alignment between the two. At times (92/500 in 2010-11, 78/500 in 2011-12 and 58/500 in 2012-13) this manifested as comments that appeared to be non-sequiturs. The following examples are taken from 2012-13:

#### Example 1:

Assessor:  
[10974937] Performs appropriate checks but needs to remember to check breast feeding for herself rather than rely on Tech or referrer. Competant [*sic*] and can now administer Mag 3 in the department under my certificate without direct supervision provided the renogram has been correctly authorised.

Trainee (ST4):  
[ia710Ro] Need to think about the different bungs that may be attached for a needle free injection and how that will impact dose and recount values.

Example 2:

Assessor: [11534360] The case mix included difficult 'characters' which prove challenging at the best of times but you coped well with moving targets.

Trainee (ST3): [uy285ck] Need to focus on image optimisation starting with a larger field of view then honing in.

Separate codes were created in order to identify trainee responses that were either apparently contradictory to the assessor's view, or in which the trainee explicitly rejected the assessor's feedback. In both cases, the numbers of comments that satisfied these criteria were very low. For example, in 2010-11 and 2012-13 only 4/500 trainee responses appeared to contain an element of contradiction, while explicit rejection of the assessor's views was not encountered in any of the three years analysed. The following are examples of the apparently contradictory comments that were identified:

*From 2010-11*

Assessor: [797] Can work on reducing the patient's dose.

Trainee: [as695I] Good experience, have learnt how to perform this procedure

*From 2012-13*

Assessor: [11093935] [Trainee] is clearly able to do US guided procedures. He needs to remember the importance of patient positioning although for this patient it was not possible to get her in an optimum position due to space/bed size etc. I have no issues with [him] performing drainages independently.

Trainee: [ob360 M] I am very happy with this feedback but feel that some supervision would be needed still.

In fact, only a small proportion of assessor-trainee pairs made comments that were adjudged to be clearly aligned to each other: 124/500 in 2010-11, 76/500 in 2011-12 and 69/500 in 2012-13. However, it was apparent that alignment alone was no guarantee of a truly formative feedback exchange due to the brevity of, and lack of information within, some assessors' comments. Consequently, when analysing paired comments in order to determine the degree of dialogic feedback that had taken place, the assessor's comments were first selected using quality criteria derived from the literature which have already been set out in section 5.5.

### *5.6.2 Identifying dialogical feedback exchanges*

In order for any of the written feedback exchanges in this study to be deemed to be dialogical, the first condition applied was that the assessor's comments were of sufficient quality to warrant engagement on the part of trainee. Quality was determined by applying the qualitatively-derived algorithm alluded to in section 5.5 to assessors' comments, before analysing the associated trainee responses.

High quality in assessors' comments was indicated by:

- presence of positive or negative comments on the observed performance,
- presence of specific suggestions for improvement,
- absence of personal comments.

As mentioned previously in 5.5.1, applying these selection criteria to samples of 500 assessor feedback statements from 2010-11, 2011-12 and 2012-13 reduced the number of statements to 41/500 (8%), 34/500 (7%) and 37/500 (7%) respectively. The trainee responses relating to these feedback statements were then checked in order to determine the nature and extent of the engagement demonstrated by the trainee. Specifically, I was looking for direct responses to the assessor's comments that included a specific plan for further development. Only a very small number of these feedback exchanges were identified in each year (4/500 in 2010-11, 2/500 in 2011-12, 3/500 in 2012-13). Given their



limited number, these are presented below in their entirety to give the reader a sense of the nature of these feedback exchanges.

### Training year 2010-11

Assessor: [id. 2912] Has a good grasp of the need for pre-procedural planning and non-invasive imaging review - aim to do this independently and present to supervisor [sic]. Needs confirmation about patient approach to consent although no concern is expressed and all required information is discussed. Learn to improve direction of support staff. Demonstrates improving catheter and wire technique; needs to improve upon ultrasound-guided approach.

Trainee: [id. at898in] Agree with above comments and plan focus on ultrasound technique background reading on catheters and planning techniques.

Assessor: [id. 3074] Good recall for image sequence, use of equipment with dose reduction. Advise [sic] given: ensure entire swallowing sequence is recorded.

Trainee: [id. al175 P] I feel that I have learnt the basic sequences required, but aim to concentrate more on dynamic assessment of images in order to produce accurate pictures.

Assessor: [id. 2704] Daniel is competent at performing barium swallows. Couple of points.  
1. Remember to remove relevant jewellery before starting the examination!  
2. When counting down to exposure, count more slowly  
3. Watch collimation is not too tight - beware cutting off areas of interest!

Trainee: [id. an753 F] Good feedback after performing the swallows and I appreciate the above comments and understand the areas of improvement.

Assessor: Good images obtained. Whole of the colon assessed.  
[*id. 1063*] Compassionate towards patient. Good use of the C-arm equipment.  
As [trainee] gets more experience needs to try to reduce radiation dose by pulse screening and taking fewer spot images.

Trainee: I need more exposure to make good utilisation of the time, reduce  
[*id. an281 V*] the radiation dose and make my technique better. Thank you for your time for this assessment.

### Training year 2011-12

Assessor: Good communication with the patient and good initial observation of  
[*id. 1080*] function utilising fluoroscopy. Could improve understanding of pathophysiological processes.

Trainee: Will need to do some more reading around the theory behind the  
[*id. I ir*] reasons for carrying out the procedure. Will need to improve basic knowledge on pathophysiology.

Assessor: [Trainee] has demonstrated good awareness of the 3D perspective  
[*id. 1920*] required for barium enema examinations. He has with support produced good double contrast images. He needs to work on his manipulation of the image intensifier and speed of examination but he is well on the right course to produce good examinations with practise [*sic*].

Trainee: I feel that i have learnt a lot about the technique of DCBE but still  
[*id. mi662ar*] need to work on skills of moving the patient to get the best images. I will continue to work on this in the future.

## Training year 2012-13

Assessor: [id. 10251993] Need to revise ultrasound anatomy of the neck: namely the floor of mouth (very good radiographics article on this). Went through nodal stations in a systematic way. Familiar with normal appearance of cervical lymph nodes.

Trainee: [id. ta924le] I will revise the anatomy as advised.

Assessor: [id. 10409910] Did well for early attempt at pleural drainage. Advise sitting whilst consenting the patient and advise reading the NPSA report on chest drain insertion.

Trainee: [id. mr555Ka] Agree with above and will sit down for consent and ask in [sic] any more questions.

Assessor: [id. 10808798] [The trainee] approached the procedure with a methodical and diligent manner, exhibiting good communication skills with both the patient and other staff members. Good quality report and understanding of the procedure. The only area to improve on is the use of cones to limited radiation exposure and the use of the foot pedal to distance herself from the radiation source. Keep up the good work.

Trainee (ST1): [id. ho367Ge] Many thanks for feedback! Very helpful advice on reducing patient dose as well as my own with collimation and by utilising the inverse square law with the aid of the foot pedal and extension kit. Will definately [sic] use these tips in future fluoroscopic procedures.

Of the three examples from 2012-13, it might be said that this last example (Assessor 10808798 and Trainee ho367Ge) is the only one that is truly dialogical, in that it appears to

be a genuine conversation between assessor and trainee about next steps in educational development. The observation that this is the only truly dialogical feedback exchange that was found in a random sample of 500 assessments raises an important question regarding the fitness for purpose of WBA as a formative assessment system in clinical radiology.

### *5.6.3 Missed opportunities*

As alluded to in section 5.6.2 above, it appeared at times that very concise trainee responses represented something of a missed opportunity in terms of the assessor-trainee feedback exchanges being truly dialogical. This was apparent when the trainee failed to comment at any length on the detailed comments of their assessor. For example, in the following case, also taken from the training year 2012-13, the assessor provides detailed feedback on the trainee's performance, whereas the trainee's comment is much more cursory:

Assessor: This was a very difficult ultrasound examination in an obese patient  
[*id. 10079024*] however [the trainee] showed good understanding of probe positioning and is beginning to learn about parameter optimisation to improve image quality. Good concise report pointing out the limitations of the study.

Trainee: Will continue to practise ultrasound skills.  
[*id. AC401M*]

The trainee's response on this occasion appears prosaic compared with the detailed, contextualized comments on performance offered by the assessor. This may have been due to the fact that the assessor's feedback did not include any negative comments or specific suggestions for improvement, and so the trainee may have felt that there was little of any substance to which they could respond. However, the following example illustrates how similar missed opportunities existed even when the assessor did offer comments on future improvement:

Assessor: Familiarise with ultrasound parameters such as frequency/  
[*id. 10005611*] penetration to optimise images. Advise to interact more with  
patient to get relevant clinical info and get them to warm up.  
Especially when patients present at [name of hospital] with lumps  
and bumps and are understandably worried about cancer. Faster  
report dictation.

Trainee: I will address the above comments.  
[*id. ba358e*]

Again, the assessor's comments conflict with recommendations from the literature as to the importance of including both positive and negative comments when providing feedback, with the balance on this occasion being entirely on the 'negative' side. Regardless, the comments offered have not led to a specific action plan having been generated.

There were also examples of trainees' comments being noticeably more detailed than those of their assessor. The following example is again taken from the training year 2012-13:

Assessor: See above.  
[*id. 11363765*]

Trainee (ST2): I have not performed many of these procedures but find that the  
[*id. it278Ra*] technique used is similar to other drainage procedures. Would like to  
do more to enhance my skills.

On this occasion, the assessor's cryptic comment, 'See above', refers to a string of top (i.e. 6/6) scores for each of the assessment criteria within the Rad-DOPS form. The descriptors for each of these criteria are couched with reference to 'expectation for stage of training', and so the relative inexperience of the trainee (ST2) would not necessarily militate against high scores. However, assessors are reminded within the structure of the Rad-DOPS form to comment on 'areas of good practice and areas for development' – no comment had been offered in this assessment on specific aspects of good practice, and it is difficult to

believe that even a talented ST2 trainee would have no areas for development at that point in training – ST2 is only the second year of a six-year training programme. In fact, by the trainee's own admission, they were inexperienced in the procedure and identified a need to improve upon their current level of capability.

#### *5.6.4 Summary of analysis of trainee comments*

The evidence is that genuinely dialogical feedback in Rad-DOPS assessments is rare, and is decreasing year on year. When it does exist, it is arguably something of a hit-and-miss affair. Analysis of pairs of assessor-trainee comments across the first three years of the clinical radiology training programme has demonstrated that very few dialogical feedback exchanges have been recorded. It could be argued that, in general, assessors rarely provide feedback of sufficient quality to warrant engagement on the part of the trainee. However, it is also the case that when high quality feedback is provided, it is at times responded to in the most cursory fashion by the trainee; a missed opportunity. Equally, there is evidence of earnest engagement on the part trainees when the assessor has provided very little in the way of helpful feedback – a missed opportunity on the part of the assessor.

The data also reveal that the style of many of the assessors' written comments is that of reportage rather than dialogue, despite the fact that trainee responses (when they existed) were often addressed to the assessor e.g. thanking them for their time or for the feedback provided. This suggests that the assessors' comments are written for the benefit of another audience, such as the trainee's educational supervisor or the annually-convened ARCP panel, as well as the trainee, and may be further evidence of a tension at the heart of these formative assessments which could be serving to limit their validity.

## 5.7 Summary of key findings

### *Descriptive statistics*

Descriptive statistical analysis of a number of assessment-related parameters revealed several important patterns. These were:

- trainees generally undertaking numbers of assessments which closely matched the minimum numbers required by the curriculum
- the majority of assessments being recorded in the latter half of each placement, with peaks of assessment activity apparent around the time of supervisory meetings
- large numbers of assessments being completed at the very end of placements or after the end of placements
- large numbers of assessors conducting very few assessments in any one year.

These findings raise concerns about trainee and assessor engagement with the formative process. The evidence suggests that there may be a degree of instrumentalism on the part of trainees, for whom recording the correct number of assessments is arguably the motivation behind these patterns. The RCR might also be concerned about the benefit to the assessment system of training consultants who complete very low numbers of assessments each year.

### *Content analysis of assessor comments*

Qualitative analysis of assessors' feedback statements demonstrated that:

- assessors generally provided feedback of very low quality
- comments on trainees' performance tended to be expressed in very general terms, with a clear emphasis on positive rather than negative comments
- assessors' feedback often failed to articulate any need for further development or learning

- when the need for further development was identified, it was often explicitly assumed that continued practice would lead to the necessary improvements
- suggestions as to how to effect further improvement were rarely provided, and tended to be expressed in general rather than specific terms
- specific comments on how to improve performance in a given procedure were rarely provided.

These patterns were maintained throughout three consecutive years of the programme. This suggests that they were not due to ‘teething’ difficulties associated with the introduction of a novel assessment system.

#### *Statistical analysis of assessors’ coded comments*

The coding frequencies generated via qualitative content analysis were subjected to statistical tests of independence using the Chi-squared formula. The analysis revealed that:

- very few significant relationships existed between the type of comment provided and important underlying factors such as the stage of training of the trainee being assessed, their average score in the assessment, and the overall competence rating assigned to them by their assessor
- trainees who were at an earlier point in their training, who had received a lower average assessment score, or who had received a low overall assessment judgement were significantly more likely to receive negative feedback comments than their more senior or more highly scoring colleagues, however they seemed no more likely to receive specific comments on how to effect further improvement.

#### *Ragin analysis of assessors’ feedback statements.*

An approach to qualitative comparative analysis described by Ragin (1987, 2000, 2008) was used to explore the conditions that gave rise to whole feedback statements (as opposed to individual feedback comments) that were judged to be of relatively high quality. Analysis revealed that:



- early stage of training, low average (modal) assessment score and low overall competence rating were necessary conditions for the provision of high quality feedback
- none of the conditions tested was found to be sufficient for the provision of high quality feedback, with the majority of trainees receiving feedback that fell short of the qualitatively derived 'high quality feedback' threshold
- some conditions – late stage of training, high average assessment score and high overall assessment judgement – functioned as NOT conditions, in that they effectively nullified the possibility of trainees receiving high quality feedback.

These findings call into question the formative function of these assessments, as the majority of trainees of all abilities and levels of seniority fail to receive high quality written feedback.

#### *Analysis of paired assessor-trainee statements*

Pairs of assessor-trainee statements were analysed to determine the degree to which the feedback exchanges within the Rad-DOPS assessments could be said to be dialogical. In all three years of the study, fewer than 1% of the feedback exchanges were found to be dialogical, and there was a pattern of statistically significant declining trainee engagement over the first three years of the programme. This pattern was exemplified by:

- increasing numbers of trainee feedback fields containing no comments, or cursory comments
- declining numbers of trainee self-assessments
- declining numbers of reflective comments
- declining numbers of trainee action plans

The next chapter, Chapter 6, provides a discussion of the implications of these findings, placing them in the context of what is currently known about best practice in formative

assessment and considering the extent to which the assessment system that has been introduced to clinical radiology in the UK is fit for purpose.

## CHAPTER 6

### 6. Discussion

#### 6.1 Introduction

In asking whether or not the workplace-based assessment system introduced by the Royal College of Radiologists (RCR) in 2010 is fit for purpose, my research has been primarily concerned with establishing the validity of these assessment arrangements in practice. To this end, a validation argument was constructed in chapters 2 and 3, in which the theoretical and evidential case was made for formative assessment and feedback being linked to learning. With this as the backdrop to my study, my research set out to analyse the empirical evidence of the extent to which WBA in clinical radiology is fulfilling its formative purpose. This chapter is focused on a discussion of the key findings of my research, and a consideration of the range of contextual factors which are arguably most likely to be responsible for the current usage of WBA in clinical radiology. The chapter concludes with a reflection on the implications of my findings for policy makers, practising clinicians, and patients, and with recommendations for the RCR as it continues to develop its approach to the postgraduate training of radiologists.

#### **6.2 Do trainee doctors and their assessors appear to use workplace-based assessments formatively?**

Gardner (2012), writing about quality in assessment practice, underlines as an important principle of assessment that ‘assessment of any kind should ultimately improve learning’ (p. 106). This is particularly the case in formative assessment, in which it might be argued that validity rests almost wholly on the capacity for the assessment to enhance learning.

However given the unpredictability of learning, in no small part due to its contingent and often contested nature, my research did not set out to measure directly whether or not improved learning occurred *per se*. Rather, drawing on ideas from quality improvement methodology, I analysed a number of evidentially-related process measures, including the number and timing of the assessments, the feedback that trainees received and the nature of their responses, as a gauge of the usefulness of the clinical radiology assessment system for supporting learning.

### 6.2.1 *Timing of the assessments*

My examination of assessment literature had revealed that, at times, researchers and educators alike use the term 'formative' to describe assessment which, in reality, struggles to fulfil the description in any meaningful way. For example, in Sinclair and Cleland's 2007 enquiry as to which university students sought formative feedback, they were referring to written feedback provided to students *after* they had undertaken a high-stakes, end of module assessment, in a module that bore little resemblance to forthcoming modules that the students were likely to encounter. Thus the timing of this feedback rendered it of little practical use to the students, other than perhaps as an explanation of the mark they had been awarded. It was reported that few students, and especially among the poorest performing students, bothered to collect their feedback. Carless (2013), writing about assessment in higher education settings, also highlights the limited effect of feedback which comes too late in the learning process to be of value to learners. Consequently, I felt that it was important to establish whether or not radiology trainees were undertaking formative assessments at a point within each training placement at which they had a realistic chance of acting on the assessor's feedback.

The data in my study suggest that, after an initial lag period, assessments were distributed throughout the training placements to some extent, rather than all being requested at the end as might be expected if trainees viewed the assessments as being purely summative. It looks, therefore, as if some trainees and their assessors have been using the assessments from a relatively early stage within placements. However, peaks of assessment activity were apparent around 50% and 90% of the way through placements, at the points where formal discussions about progress with educational supervisors

typically occur. There was also clear evidence of a large number of assessments being uploaded at the very end of placements when the opportunities for continued development are likely to have been limited – some 527 to 971 assessments depending on the year being analysed. Of more concern from the point of view of providing formative feedback, there was evidence of large numbers of assessments being recorded beyond the end of the training post – 403 to 1055 assessments depending on the year of the programme that was analysed. Thus, it appears that many assessments designed to be used in ‘real time’ were being completed retrospectively. Taken together, these very late and retrospective assessment accounted for 19-25% of assessments recorded, depending on the year being analysed.

Retrospective assessment was also documented by Rees *et al.* (2014) in their evaluation of workplace-based assessment use amongst doctors in foundation training. In their study of 70 UK foundation trainees, they found that:

[assessments] were sometimes initiated retrospectively, at times long after the event, particularly when trainees had completed insufficient tools during their placements. (*ibid.*, p. 7)

This may suggest that for some trainees, the reality is not so much that assessment is driving learning than supervision and curriculum requirements are driving assessment. If so, the primary motivation of these trainees is more likely to be the fulfilment of their training obligations rather than the pursuit of useful learning experiences. The former does not necessarily preclude the latter – assessment encounters, however motivated, may contain within them the potential for genuinely helpful observation and developmental feedback. However, undertaking formative assessments late in the clinical placement offers limited scope for formative development.

The patterns of assessment revealed in the study 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 pattern of late assessment may indicate that trainees are imposing traditional concepts of assessment *of* learning onto the workplace-based assessment programme – concepts such as ‘passing’ or ‘failing’ – which cause

them to delay assessment until such times as they believe themselves to be competent. This approach may even be self-affirming – late assessment requests from trainees create pressure on assessors to ‘sign off’ trainees as competent, given the limited scope for further development and the imperative for the trainee to provide an appropriate number of satisfactory assessments in order to ensure progression. Thus, in *perceiving* the assessments to be high-stakes, trainees may in turn be inadvertently *creating* high stakes assessment events by leaving them to a point when the assessor has, in effect, a pass/fail-type decision to make.

### 6.2.2 Frequency of the assessments

In writing about the conditions required for productive formative assessment and feedback, Carless (2013) also identifies the importance of well-timed formative assessment, but similarly emphasises the significance of frequent assessments:

The possibilities for productive feedback provision are deeply affected by the number, timing and sequence of assessment tasks which students undertake (*ibid.*, p. 93).

For Carless, frequent formative activity is important for the development of trust, which he argues is central to the subsequent development of a ‘transformative, dialogic learning environment’ (*ibid.*, p. 91). This idea of frequent assessment contrasts with my finding that most clinical radiology trainees only undertake numbers of Rad-DOPS assessments that are close to curriculum requirement of having six within a 12-month period. If evenly spaced throughout the training year, the frequency of these six assessments (one assessment every two months) cannot realistically be regarded as ‘frequent’. This number would equate to two assessments within a typical four-month radiology placement – arguably not the basis for developing a trusting feedback relationship. Furthermore, rather than the assessments being evenly spaced, my research revealed evidence of some trainees recording several assessments on a single day. This is similarly antithetical to the notion of developing a trusting feedback relationship over time.

Thus, it appears that, regardless of the quality of assessors' formative input, the patterns of WBA use that are discernible from the analysis of the national e-portfolio record suggest that the system is generally not functioning in a manner that is conducive to the conduct of 'transformative, dialogic' feedback exchanges (Carless, 2013, p. 91).

### *6.2.3 Does written feedback provided to clinical radiology trainees support the development of competence?*

Analysis of samples of 500 written feedback statements from across three training years demonstrated that there were differences in the feedback provided by individual assessors with respect to valency (i.e. whether the feedback is positive or negative), focus and specificity; and that these differences were conserved over time.

#### *The predominance of positive general feedback*

Feedback samples from all three training years contained a large proportion of positive comments, most of which were phrased in general rather than specific terms. This is generally in keeping with the findings of other researchers within medical education. For example, Canavan *et al.* (2010) coded multisource feedback comments linked to professionalism and found that 90% contained positive feedback. Fernando *et al.* (2008) found that 77% of assessors in undergraduate workplace-based assessments provided positive comments. Educationally, this degree of positivity may be cause for optimism regarding the WBA process. In his meta-analysis of educational interventions, Hattie (1999) identified 'feedback', and in particular positive 'reinforcement' (p. 9), as being key educational activities linked to improved performance, and so it would appear to be important that assessors do take the trouble to comment positively on aspects of performance that they believe to be of a high standard. However, positivity in itself is not necessarily prized by trainees. Participants in one study described overly positive comments as being a negative property of some feedback, believing it to be 'educationally unhelpful' (Rees *et al.*, 2014, p. 7). Medical trainees in the Rees *et al.* study declared a preference for what the authors termed 'feedforward' (*ibid.*, p. 7) i.e. feedback that was clearly intended to improve their performance. Their stated preference for what they

termed 'personalised' feedback, contrasted with 'non-specific' feedback in the Rees *et al.* (2014, p. 7) study, is also confirmation of trainees' preference for specific feedback. In other words, positive general feedback, of the type that was found in the large majority of assessments in the current study, is not necessarily highly valued by trainees.

According to Nicol (2010), one of the things that assessors can usefully do to clarify their feedback is to link their comments to published assessment criteria. As noted by Torrance (2007), there is a risk that doing so in a focused manner may create a convergent assessment environment – one which is focused narrowly on published criteria such that other non-prescribed learning is ignored. However, an arguably greater risk is that trainees' confusion may lead them either to dismiss the comments or to exhibit the attribution error and self-handicapping behaviours described by Berglas and Jones (1978) and Thompson and Richardson (2001). The evidence in my research was that 44-51% of assessors' formal feedback statements did contain reference to the assessment criteria. It is worth noting that the remainder – around half of all feedback statements provided – therefore did not refer to the assessment criteria, and that the majority of these (91-94%) were also phrased in general terms.

#### *Assessors fail to provide negative feedback*

One of the recurrent themes in the feedback data was the lack of negative feedback provided by assessors. The educational implications of a lack of negative feedback may be a cause for some concern, given the difficulties that learners can experience with accurate self-assessment. Davis *et al.* (2006), in their systematic review of the evidence regarding physicians' ability to self-assess, found that doctors of all levels of seniority struggle to identify accurately their level of capability across a range of professional competency domains:

In the studies indicating poor or limited accuracy of self-assessment, this finding was independent of level of training, specialty, the domain of self-assessment, or manner of comparison (p. 1100).



Clinical educators should therefore be aware that a failure to deliver feedback to trainees on the aspects of their performance they could improve may be to deprive these learners of an important supply of senior clinician input regarding the standard of their performance. Certainly, if appropriately formulated negative feedback is a driver for learning, trainees in clinical radiology appear to be missing out.

#### *A difficulty with giving negative written feedback?*

In considering why clinical radiology assessors fail to provide negative written comments in WBA, it may be the case that negative feedback is simply unwarranted, due to the assessor having observed the trainee performing to a high standard. Certainly, the finding that the award of a high modal assessment score, or a high overall competence rating, rendered trainees significantly less likely than their low scoring peers to receive negative comments would suggest that assessors reserved their negative comments for less well-performing trainees. The same was true for trainees in the later stages of training, who were significantly less likely than their more junior colleagues to receive negative comments. However, it was also the case that the vast majority of trainees who had low assessment scores, or who were at an early stage in their training, did not receive negative feedback. The data unequivocally show that assessors are not providing this type of feedback when it might be most expected i.e. when trainees' assessment outcomes are poor, or when they are at the earliest stage of their training.

#### *Failure to fail*

Whilst individual WBAs in clinical radiology are not pass/fail assessments the competitive nature of medical trainees referred to by the GMC (2010), and the implications of accumulating a number of unsatisfactory assessment outcomes for a trainee's progression, mean that negative feedback is likely to be perceived in a similar manner to a 'fail' in a single high stakes encounter. Thus, the published evidence from the likes of Ingram (2013), Rees *et al.* (2008), Cleland *et al.* (2008) and Dudek *et al.* (2005) regarding assessors' unwillingness to fail medical students whose performance is clearly unsatisfactory is potentially illuminating in the clinical radiology WBA context.

In exploring the reasons that assessors in medical education may 'fail to fail' underperforming students or trainees, Dudek *et al.* (2005) found that a number of factors influenced the thinking of assessors. These included: insufficient knowledge of what to document when perceiving a student to be failing; lack of corroborating evidence from other sources, including colleagues, leading the assessor to doubt the veracity or defensibility of their own judgement; and anticipating an appeal process that would be time consuming and/or stressful. These findings indicate that assessors often conduct a personal risk/benefit analysis when deciding whether or not to fail an underperforming trainee. In other words, the issue for the assessor is not simply whether they could identify a failing trainee – participants in Dudek *et al.*'s study (*ibid.*) generally felt that they could do so with confidence – but their confidence was offset by an awareness of the challenges to their judgement that may be precipitated by failing a trainee, and the need to identify and record specific, objective and well-documented evidence of underperformance. Assessors, therefore, often erred on the side of leniency rather than awarding a grade that would be challenged.

The reasons revealed by Dudek *et al.* (2005) may arguably apply in the field of clinical radiology, resulting in an observed reluctance to award low scores and a disinclination, even if low scores have been awarded, to write anything negative about the trainee in an official (albeit educational) document. As one assessor in Rees *et al.*'s (2008) study put it, 'You don't want to sort of be the one who sticks the knife in them' (p. 5). It is likely that the duality of purpose in WBA – simultaneously being intended to provide developmental feedback while also feeding into high stakes decisions about progression – is serving to corrupt their declared primary formative function, with trainees and assessors alike being aware of the potential consequences for a trainee of having negative comments recorded in their official e-portfolio. Stobart's (2008) argument about the constitutive nature of assessment is apposite here, and will be considered in more depth below. Suffice it to say at this point that the Annual Review of Competency Progression (ARCP) panel, who decide whether trainees progress to the next year of their training (or indeed whether they are deemed to have completed training, and are suitable to be appointed as a consultant in the UK), do so based solely on the 'picture' of the trainee painted by their WBA e-portfolio record. It is not difficult to appreciate why trainees in this context would be reluctant to receive negative feedback, and why assessors may be unwilling to provide it.

### *A lack of high quality feedback*

A tentative, qualitatively-generated formula for identifying high quality feedback was synthesized from the review of the evidence on formative assessment and feedback. The standard described by this framework proved to be unattainable for all clinical radiology assessors in all three years of the programme that were analysed, and the framework was subsequently revised to reduce its stringency. Even then, only a small number of assessor feedback statements in each year were found to qualify as 'high quality feedback'. Whether this was due to what Prins *et al.* (2006) term 'production deficiency' (p. 300), in which assessors possessed the ability to provide high quality feedback but failed to deploy their skills, or 'availability deficiency' (*ibid.*, p. 300), in which the essential skills did not exist, cannot be determined from my data. Thus, in spite of the assessor training provided by the RCR, the quality of formal feedback within clinical radiology training does not compare favourably with the type of feedback that has been demonstrated within the education literature to be associated with enhanced learning. Pedder and James (2012, p. 36) note that staff development for the delivery of effective assessment for learning must lead educators to 'go beyond changes in surface behaviours' by applying techniques or adopting certain practices. The evidence is that even surface behaviour change has thus far not been achieved amongst clinical radiology assessors. The majority of the assessors have done little more than engage in a token manner with the process, and have not even adopted the 'surface behaviours' of feedback provision. There was also evidence of a trend of decreasing assessor engagement over time, with numbers of assessments containing no feedback comments increasing significantly over the first three years of the programme. When high quality feedback was provided, Ragin analysis demonstrated that it was confined to certain groups of trainees: those who were in the early ('core') phase of radiology training, who had a low modal assessment score, or who had a low overall competence rating. However, it was also found that the majority of trainees in these categories did not receive high quality feedback. Furthermore, trainees who might be deemed to be in the greatest need of high quality feedback – those whose overall competence was rated as needing 'additional support' – generally did not receive this input. Trainees who fell outside these categories – those in the latter ('advanced') phase of training, those who had high modal assessment scores and those who had high overall competence ratings – were also systematically excluded from receiving high quality

formative written feedback on their performance. The evidence suggests that, for these trainees, the system is effectively not formative and is instead almost solely a collection of evidence for presentation to the ARCP panel.

### *Trainee engagement*

The findings regarding trainee engagement are even more stark. The majority of trainees in each of the years analysed simply did not comment on their assessor's feedback, or did so in the most cursory fashion. Furthermore, the proportion of trainee comment fields which lacked a response or contained only a cursory response increased significantly over the first three years of the programme. This was accompanied by a marked decrease in trainees providing reflective comments or self-assessments. Taken together, these findings suggest that the WBA system entered a period of rapid decline after its introduction. If this was maintained it is difficult to be optimistic about the state of the system in 2015-16.

It might be argued that poor trainee engagement was predicated on poor quality feedback having been provided by assessors, but as reported in section 5.6.3 trainees often did not respond in any meaningful way even when their assessors had provided relatively high quality feedback. When trainees did respond to assessors' high quality comments, the trainees' comments were often not aligned to the assessor's comments. Consequently, despite the RCR's provision for a written feedback exchange within the documentation, it is not an exaggeration to say that dialogical feedback almost never occurred.

The nature of the traditional mass higher education setting presents obvious difficulties in achieving dialogical written feedback exchanges between teachers and students, and is evident from the references to these challenges within the higher education literature. Not least, there is the difficulty of the traditional 'arm's length' educational relationship between lecturers and their students, particularly in subjects that attract large cohorts of students. By contrast, the professional clinical setting typically provides the scope for close working and educational relationships between senior and junior colleagues, and so the lack of a close relationship arguably does not provide a reason for the lack of dialogical feedback exchanges within the formal WBA documentation in clinical radiology. The answer may lie

in the goodness of fit, or rather the lack of it, between the long-established approach to education in clinical radiology and the formal system of formative assessment that has been introduced by the RCR. Given the strong oral tradition of teaching and learning in clinical medicine, there is evidence that a bureaucratic system of workplace-based assessment and feedback may be superimposed onto a training process that has existed for many years prior to its introduction and which is not well matched by the new electronic, document-based approach. The question therefore arises: has the investment in doctors' time in implementing the WBA process contributed to identifiable educational value? My research certainly indicates a conclusive 'No'.

### **6.3 The broader picture of formative assessment in clinical radiology**

The many hours of doctors' time given over to conducting WBA should surely demand that the WBA system delivers something of educational worth. In seeking to quantify just how much time is involved in conducting WBA, Watson *et al.*'s (2014) pilot study of the use of DOPS assessments for ultrasound-guided procedures found that the mean time taken to complete an assessment was 6min 35s. Taking this as an estimate of the mean time taken to complete the Rad-DOPS assessments at the centre of my study, it would equate to over 1000 hours of the participating doctors' time in the first year of their use, and an even greater number of hours in subsequent years. Yet, the findings of my research suggest that there are fundamental problems with the conception and implementation of the WBA system in clinical radiology, such that much of this time may have been effectively wasted. In the section that follows, I place these problems in the context of larger educational concepts, before making a number of recommendations for the RCR in considering how WBA should be developed in clinical radiology.

#### *6.3.1 Instrumentalism and assessment as learning*

My research has revealed strong evidence that assessors and trainees in clinical radiology have thus far taken a largely instrumental approach to WBA. Ecclestone (2012) defines instrumental assessment as occurring when 'instruments and methods become ends in themselves and develop a life of their own' (p. 142). There is evidence that this has

occurred within the UK foundation training programme for newly-qualified doctors – Rees *et al.* (2014) found that trainees were motivated to record retrospective assessments as a consequence of having ‘completed insufficient tools’ (p. 7) within their placements. In other words, trainees requested WBAs from their assessors long after a particular clinical encounter had occurred to ensure that the numbers of assessments they had in their portfolios matched the number that they had been told would be required. My research also identified a pattern of deadline-driven approaches to WBA in clinical radiology, and retrospective assessment, both of which are likely to be attributable to the same need to complete a pre-determined number of assessments. The finding that most trainees recorded numbers of assessments that were equal to, or very close to, the six Rad-DOPS per year required by the RCR’s curriculum is a further indication of instrumentalism in WBA.

Torrance (2007), writing about workplace assessment in the vocational, post-compulsory sector, uses the phrase ‘assessment as learning’ (p. 291, original emphasis) to refer to this type of approach. In considering its origins, he has noted that, despite the best intentions of educators, ‘assessment as learning’ may have come about as a result of a particular approach to the implementation of formative assessment. Formative assessment that is conducted convergently – in which teachers report on achievement measured against defined curriculum objectives (one might say ‘competencies’) – is likely, he argues, to give rise to a culture of box-ticking in which learners use assessments to demonstrate achievement. This is in contrast to Torrance and Pryor’s (1998) divergent approach, in which formative assessments are used in a much more open-ended manner to explore learners’ current capabilities. Torrance (2007) further points out that in professional learning the segue from assessment *of* learning to assessment *as* learning may have been so swift that, in effect, assessment *for* learning has never existed. Furthermore, the transition may have been made even harder to identify as instrumental assessment has donned the apparel of genuine formative assessment:

...it might be argued that in post-compulsory education and training, practice has moved directly from assessment *of* learning to assessment *as* learning, but this is justified and explained in the language of assessment *for* learning: providing feedback, communicating criteria to the learner, and so forth. Thus the supposedly educative formative process of identifying and sharing

assessment criteria, and providing feedback on strengths and weaknesses to learners, is ensconced at the heart of the learning experience in the post-secondary sector, infusing every aspect of the learner experience. But it is a process which appears to weaken rather than strengthen the development of learner autonomy (*ibid.*, p. 291).

It is interesting that Torrance (2007) does not view this corruption of formative assessment as benign – it is not just that educators and learners are wasting their time, but rather that the learning process has been actively damaged. In the context of clinical radiology, it is not difficult to appreciate the impact on learner autonomy when considering how trainee doctors – however knowledgeable and highly skilled – may react upon discovering that too few WBAs being recorded in their e-portfolio is sufficient for them to be labelled as showing ‘lack of engagement with educational processes’ (NACT, 2013, p. 5) and potentially being viewed as a ‘trainee in difficulty’.

It might be argued that if the learning process has been damaged the results should be apparent, perhaps via declining standards in the practice of trainee radiologists or fewer trainees progressing successfully to the completion of training. However, that would be to assume that learning in clinical radiology training *depended* on the existence of a functioning WBA system. On the contrary, it is my contention that rather than being a core educational process, the WBA system in clinical radiology is in fact a parallel process, which co-exists with established approaches to teaching and learning in clinical radiology. Thus it has little to do with how radiology is actually learned by trainees in the authentic clinical environment. This separation of WBA from pedagogy is one of the fundamental ways in which assessment for learning in radiology differs from assessment for learning in classrooms, despite the similarity of some of the rhetoric. This is, I argue, one of the ways in which the system is flawed.

### 6.3.2 *Assessment as pedagogy*

The difficulty of conceiving the WBA system in clinical radiology separately from established approaches to professional learning within the discipline may arguably stem from a greater problem – that there is no evidence that WBA has been conceived of as


pedagogy at all. This contradicts James's (1998) assertion that 'assessment should become fully integrated with teaching and learning, and therefore part of the educational process rather than a "bolt-on" activity' (p. 172). Seeing it as such, argue Pedder and James (2012), reduces the likelihood that formative assessment will be dispensed with when the pressures of the educational environment become overwhelming. As McIntyre (2000) identifies, there are features of the classroom environment that may well overwhelm teachers, and thus act to limit the implementation of high quality oral and written formative interactions. McIntyre (*ibid.*) identifies these as: multi-dimensionality, simultaneity, immediacy, unpredictability, publicness and historical embeddedness. Each of these, in combination or in isolation, can act to limit the apparent practicability of high quality assessment for learning practices such that they are simply dispensed with by teachers. It is not difficult to appreciate how each of these features is also likely to be present in the topography of the clinical environment, with the presence of patients adding a particular piquancy to the idea of publicness.

The answer, according to Pedder and James (2012), is to integrate assessment for learning within 'routine classroom practices' (p. 37), on the basis that everything learners do within the classroom context carries the potential for yielding information for teachers and students about their current capability. The substitution of the word 'clinical' for 'classroom' in the Pedder and James (2012) quotation, above, may similarly unlock the ability of consultants and trainees in radiology to engage in genuinely formative exchanges. It is surely the case that, in the professional learning context, trainee and consultant radiologists' routine *clinical* practices would provide their supervisor with ample informal opportunities and information to gauge trainees' current level of capability and inform decisions about the next steps in their learning. Formative assessment cast in this light is likely to align much better with the ways that teaching and learning normally proceed in clinical radiology, and suggests a somewhat different approach to the sporadic, formal assessment of individual clinical encounters that has thus far been adopted.

### 6.3.3 Peer assessment and self-assessment

The WBA system in clinical radiology does not currently provide a role for peer assessment, and provides only limited scope for self-assessment through the trainee



comment box at the end of the WBA form. This is a significant departure from established school-based approaches to formative assessment, and one which arguably serves to limit the gains that may otherwise be achieved by embracing a broader concept of assessment for learning. For example, as established by Moss and McManus (1992) in their interview study with UK medical students, the hierarchical nature of medicine can provoke anxiety in learners that limits the usefulness of the feedback they receive. According to Ende *et al.* (1995), attempts by senior doctors to effectively flatten this hierarchy by softening their feedback interactions with junior colleagues can at times cause them to deliver messages that are so imprecise as to be confusing to the trainees concerned. In any case, senior doctors may not always be the best people to offer feedback. As Haber and gard (2001) argue, their professional knowledge is deeply embedded and tacit, and therefore difficult for them to recognise and articulate. Instead, near peers may be better placed to communicate the recently-learned nuances of clinical practice. That is not to say that senior doctors cannot provide effective feedback to juniors, but it requires these senior doctors to have developed a metacognitive ability that they may not naturally possess. Interestingly, according to Carless (2013) the development of this metacognitive awareness, along with the potential to learn directly from the peers whose work they are assessing, is a potential advantage of involving peers in formative assessment:

A useful strategy in the pedagogy of dialogic feedback is to involve students as assessors so that they develop an awareness of making judgements about quality, deepening their understanding of alternative ways of tackling a task, developing a more critical perspective on their own work and potentially learning from the work of their peers (*ibid.*, p. 93).

Of course, the risk is that the feedback offered by peers may be less expert than the perspective offered by senior colleagues, but by way of a pay-off Carless (2013) offers the hope of establishing a truly dialogic feedback environment, in which learners are less concerned with presenting themselves as competent, and more willing to admit to weaknesses and misunderstandings. Is it clear from my research that the written feedback exchanges that occur between assessors and trainees in clinical radiology are far from dialogic, and that trainees rarely proffer a view on their limitations. Consequently, establishing a role for peer assessment and feedback may release much more of the potential of formative assessment in clinical radiology.

#### 6.3.4 *The role of the teacher and learner in formative assessment*

At present, the approach to WBA in clinical radiology is typified by senior doctors conducting infrequent, set-piece observations of individual clinical encounters and recording outcomes and feedback. The evidence suggests that rather than addressing a genuinely formative purpose, the process facilitates trainees being able to present evidence of engagement with the assessment system to an annually-convened ARCP panel. There is no doubt that the introduction of WBA into clinical radiology has therefore changed the educational role of senior doctors, which was formerly largely limited to instructing trainees within the clinical environment, by requiring them to function as formal assessors and documenters of their trainees' progress. However, as has already been described, these doctors appear to have done so in a utilitarian fashion, which limits the impact of formative assessment. It might be said that the new approaches adopted by assessors in clinical radiology align with traditional roles and practices of teachers in school-based settings. It is useful, therefore, to draw on important learning from school-based research, which tends to suggest that for assessment for learning to function as pedagogy, and thus truly impact formatively on learning, a transformation is required in the traditional role of teachers.

This transformation has often been difficult: Marshall and Drummond (2006) found that the classroom practices exhibited by teachers participating in a Learning How To Learn (LHTL) project often failed to align with the teachers' espoused values. The majority of teachers in their study failed to promote learner autonomy, or to make learning explicit, to the degree that they claimed to value these dimensions of classroom practice. Conversely, they overemphasized a performance orientation within their teaching and learning practices despite claiming not to value this approach. Pedder and James (2012) contend that teachers who fail to transform their practice according to formative assessment principles fail to mobilise both themselves and their learners as agents within the educational process. It is reasonable to propose that a continuation of didactic approaches to teaching by definition limits the potential for dialogic, formative exchanges. As Sadler (1998) observes, non-convergent environments are required in order to allow teachers to explore how their learners utilise what Bloom *et al.* (1956) originally characterised as higher order cognitive functions e.g. constructing arguments, synthesising information and solving

problems creatively. One might expect that encouraging this higher level thinking would be a priority for educators in a sophisticated professional domain such as medical education, and that increasing the agency of both teachers and learners would be the key to achieving this in practice. Yet the agency afforded to teachers and learners, essential for the creation of a truly dialogic learning environment, is not necessarily unfettered, and may not function in the educationally 'pure' manner envisaged by Pedder and James (2012). As Black and William (2012) point out:

Within the classroom, the actors or agents involved are, of course, the teacher and the students, all of whom exercise agency to a greater or lesser extent, *within the constraints and affordances they perceive to be present*. This means that their actions are to be interpreted in terms of their perceptions of the structure in which they have to operate, in particular the significance they attach to beliefs or actions through which they engage...with the other agents and forces. These ways [of engaging] may inhibit or encourage any changes, notably those required for successful formative assessment, in which case both teachers and students would have to change the roles that they have adopted. (*ibid.*, p. 207, my emphasis).

In other words, the system within which teachers and learners operate, or believe themselves to be operating, is likely to have an impact on how they function. In effect, their agency is not deployed purely for the creation of a dialogic formative environment, but is used instead to make decisions or take actions which are influenced by the other factors that impinge on them. As Holland *et al.* (1998, p. 52) put it, teachers and learners operate in 'a socially and culturally constructed realm of interpretation in which...significance is attached to certain acts, and particular outcomes are valued over others.' Thus in radiology training, formative assessment may be undertaken convergently if this is perceived to be the most expedient way to demonstrate that trainees have achieved competence, or negative feedback may be withheld or substantially tempered by an assessor in order to protect a generally competent trainee from the close attention of an ARCP panel.

Another challenge to the creation of a divergent formative assessment culture is the complexity of doing so for the teacher. Even if senior doctors in clinical radiology do wish to operate according to what might be termed the 'true spirit' of assessment for learning,

they need to have the opportunity to develop a sophisticated and, most likely for the majority of them, novel concept of the roles of both the teacher and the learner in a genuinely formative environment. As Black and Wiliam (2012) highlight, even classroom teachers can struggle with some of the transitions required, for example: from regulation of activity to regulation of learning; from evaluating what learners can do to understanding the thinking that lies behind the learner's performance; and from assessing learners to promoting and supporting their efforts at self-assessment. Each of these transitions represents and requires a greater, overarching transition, which may be described as a handing over of responsibility for learning from teachers to learners, and it may be the case that this clashes with the views of both groups as to what education actually is.

In order for assessment for learning to enjoy its fullest expression, learners, not just teachers, must assume responsibility for the three processes that Wiliam and Thompson (2007) describe as comprising effective formative classroom practice: establishing where they are in their own learning through self-assessment; ascertaining where they need to go; and determining how to get there. It is clear from my research that, despite the RCR's (2010) assertion that training in clinical radiology should be trainee-led: almost none of the trainees in my study offered a view of their current level of capability; few had articulated where they were trying to get to; and fewer still had formulated anything that would pass for a clear plan of action as to how they might move forward in their learning. It seems, therefore, that trainees may require similar educational input to the assessors in order to support their engagement with the spirit of formative assessment in clinical radiology. However, as with the assessors, trainees may perceive the WBA system as it currently operates to be too high stakes to risk an honest self-assessment of their current levels of competence. The recent reports of a trainee doctor having their educational reflections used against them in legal proceedings (Furmedge, 2016) will have done little to encourage the belief that the WBA process is low stakes. Once again, it seems that the current system, and the negative washback effects that it creates, is likely to confound the best efforts of the RCR to introduce genuinely formative assessment into clinical radiology training.

### 6.3.5 Constitutive assessment and the long shadow of the ARCP

In presenting his argument regarding the constitutive nature of assessment, Stobart (2008) cites the poignant story of 'Hannah the nothing' (p. 2). Hannah is a Year 6 school pupil who is apprehensive about the forthcoming National Curriculum tests, having been told that failure to achieve Level 4 in the test would result in her not being allocated a level – in her language, 'being a nothing' (Stobart 2008, p.2). For Stobart, this is one illustration of how assessment 'does not objectively measure what is already there, but rather creates and shapes what is measured – it is capable of 'making up people' (*ibid.*, p. 1).

For clinical radiology trainees, an equivalent high-stakes experience is the annual review of competency progression (ARCP), in which a panel of senior doctors meet to decide whether or not trainees progress to the next year of training, or for those in their final year, qualify as specialists in clinical radiology. The ARCP process is explicitly intended to be a review of evidence (Gold Guide, 2016), and panels may not interview or conduct a 'viva' with trainees as a means to arriving at a decision. Rather, ARCP panels must rely wholly on the trainee's e-portfolio record as a basis for making their judgement. Consequently, there is a very real sense in which trainees do not just *have* an e-portfolio – to the ARCP panel, trainees *are* their e-portfolio.

It is therefore in the interest of trainees who wish to progress smoothly through training to carefully curate an e-portfolio record that presents them in the best possible light to the ARCP panel. This is because while the guidance issued by the four UK departments of health emphasises workplace-based assessments as a 'key element' of a trainee's evidence (Gold Guide, 2016, p. 49), there is no guidance as to how the panel should evaluate WBAs. For example, the criteria could include: engagement with the WBA system, as measured by numbers of assessments done; satisfactory outcomes, as measured by the numerical assessment scores or the overall competence ratings being above a certain standard; or the demonstration of educational progression, with a spread of assessments showing improvement over time. Here, Carless's (2013) emphasis on the importance of trust for effective formative assessment is germane – with little in the way of conventional mechanisms for the development of trust (such as transparency and frequent interactions), trainees must hope that ARCP panels take a positive view of their educational record, and

are left guessing as to what will pass muster. For radiology trainees, whose educational history as doctors will necessarily consist of repeated high performance in challenging summative assessments, it is reasonable to predict that they would resort to a proven strategy by ensuring their WBAs demonstrate nothing but continued high achievement. This would explain why so few trainees risk undertaking more than the required numbers of assessments, and leave them until close to deadlines to ensure that they have had time to develop sufficient competency to 'perform well' in the assessments. My finding, reported in 5.6.1, that trainees made few reflective comments (12/500 in 2010-11, 16/500 in 2011-12 and 0/500 in 2012-13) is also understandable in this context – reflective comments generally referred to a previous or ongoing learning need, which trainees may be unwilling to share with a panel of senior colleagues making summative progression decisions.

#### *6.3.6 Summary*

In drawing together the aspects of the discussion above, it seems that there are several fundamental difficulties with the WBA system in clinical radiology. These have functioned either individually or, more likely, in combination to ensure that the engagement of assessors and trainees alike has largely been piecemeal and valueless. In considering all of the available evidence, it is difficult to escape the conclusion that the WBA system in clinical radiology at present is not fit for purpose.

## CHAPTER 7

### 7. Concluding remarks

#### 7.1 Introduction

In addressing the question of whether WBA in clinical radiology is fit for purpose, Chapter 6 presented a discussion of the results and their implications for formative assessment in this medical specialty. My research identified that there was strong evidence of tokenistic engagement on the part of assessors and trainees, which manifested as poor quality feedback on the part of assessors, minimal or even declining engagement on the part of trainees, and patterns of assessment behaviour that could arguably be seen as fulfilling instrumental ends. Tensions were identified in the role of the assessments linked to the multiple purposes to which the assessment outcomes may be put. These purposes included providing formative feedback, informing decisions about progression, identifying trainees in difficulty and protecting patient safety. This chapter places these findings in the context of their implications for WBA practice, and makes recommendations for improvement.

#### 7.2 Wider implications

##### *7.2.1 Fragmentation of professional competence*

The decision taken by the RCR to articulate curriculum content in terms of lists of behavioural statements has arguably had the effect of fragmenting professional competence in a manner that, as previously discussed, resists the determination of a

straightforward way to produce overall competence, and may also serve to limit the formative nature of WBAs. This approach to curriculum design is one that has been adopted by other UK medical specialties – for example, the syllabus content for all 30 specialties overseen by the Royal College of Physicians (RCP, 2016) is also expressed as detailed lists of behavioural statements. This approach has been almost universally embraced by medical specialties in the UK, but there are exceptions. The Royal College of Psychiatrists (RCPsych), for example, has adopted an approach to competency-based education originally developed in Canada – known as the CanMEDS framework (Royal College of Psychiatrists 2016, p. 5) – in which the competencies are less granular and are aimed at describing more overarching elements of psychiatrists’ capability. As the RCPsych describe it, ‘[the] curriculum is based on meta-competencies and does not set out to define the psychiatrist’s progress and attainment at the micro-competency level’ (*ibid.*, p. 6). Moreover, the GMC guidance on the acceptable design of postgraduate medical curricula does not overtly require lists of detailed behavioural statements:

The curriculum must set out the general, professional, and specialty-specific content to be mastered, including...the acquisition of knowledge, skills, and attitudes demonstrated through behaviours (GMC, 2008, p. 7).

Thus, medical royal colleges are free to propose alternative approaches to organising their curricular content. In recent times, the notion of entrustable professional activities (EPAs) has been proposed by Ten Cate (2013, p.6) as an antidote to the use of an ‘elaborate framework of competencies, sub-competencies, and milestones while observing trainees’. According to Ten Cate (2013, p. 6), EPAs are ‘tasks or responsibilities that can be entrusted to a trainee once sufficient, specific competence is reached to allow for unsupervised execution’. The difference between competences and EPAs as expressed here is arguably at risk of appearing semantic. However, Ten Cate (2013) contrasts EPAs with competences by declaring that the former should be ‘a more limited number of comprehensive, critical tasks that should apply over multiple patient conditions’ (p. 6). The EPA concept has not yet been implemented in the UK, although pilots of EPAs are due to commence in June 2016 at the RCP. The format that EPAs take when eventually implemented remains to be seen, and it will be interesting to observe whether their introduction will have any impact on the current approach to WBA. Perhaps more holistic



statements of competence will lead to more holistic assessment criteria. Whether this would be to the benefit or detriment of formative feedback is difficult to predict, as the role of assessors in the process is also of vital importance.

### *7.2.2 Distortion of formative assessment*

The evidence from my study strongly supports the notion that the description of WBAs as formative is primarily an indication that individual assessments are not intended to facilitate pass/fail judgements. There is also a clear expectation that the assessments should lead to formative feedback. However, the WBA process described by the RCR lacks some of the key features of the fullest conception of formative assessment, not least the facility for peer assessment and a prominent role for self-assessment, and my analysis of the formally recorded e-portfolio data suggests that the hierarchical assessor-trainee interaction is failing to deliver feedback of any real quality for trainees. Further research may illuminate whether WBA encounters in the clinical environment support educationally beneficial dialogic interactions between assessors and trainees, over and above any interactional feedback that may naturally occur in the workplace setting. However, even if found to be present, high quality verbal exchanges between assessor and trainee would only fulfil one aspect of formative practice in its most complete sense.

### *7.2.3 A system in decline*

An unexpected finding from my study was that WBA in clinical radiology seems to be a system in decline. The data revealed a significant rise in absent assessor feedback over the first three years of the programme, with concomitant indicators of declining trainee engagement. These included a significant increase in trainees failing to comment and significant decreases in trainee self-assessments and reflective comments. Given the novelty of the process, one might expect that the data would show signs of the system maturing over time, with assessors and trainees alike becoming more proficient in their formative interactions, but if this rapid rate of decline is maintained it does not bode well for the effective functioning of the system now in 2015-16.

### 7.3 Recommendations

It is my contention that my research has revealed evidence of a need for improvement in the current system of WBA in clinical radiology if it is to deliver on its primary, formative aim. Consequently, there is arguably a number of small operational recommendations that could be made e.g. around the timing of assessments within placements, requirements to specify development actions, numbers of assessments undertaken within placements and throughout the year and so on. However, at the root of seeking improvement there are four overarching recommendations for improvement. My recommendations derive directly from the empirical evidence and include: broadening the concept of formative assessment in clinical radiology; conceptualising appropriately the learning that takes place in postgraduate radiology training; and developing a robust concept of what it is to be a clinical radiologist in the UK. Further research may be needed to underpin the latter two of these recommendations, to ensure that conceptions of radiological practice and of learning in postgraduate radiology training are based on robust empirical evidence.

#### *7.3.1 Broadening the concept of formative assessment in clinical radiology*

The evidence in my study suggests that formative assessment in clinical radiology has been very narrowly conceived. The approach to assessment is deliberately hierarchical – assessors must be senior doctors, or other clinicians with substantially more experience than the trainee being assessed (RCR, 2010). In a marked digression from well-established formative practice in classroom settings, peer assessment is completely absent from the WBA endeavour. In addition, scant provision has been made for self-assessment, other than supplying trainees with a right of reply within the Rad-DOPS forms. Certainly, no training was provided for radiology trainees regarding self-assessment prior to the implementation of the programme, and the training delivered to assessors made no mention of supporting learners' efforts at self-assessment. Assessment has arguably been conceived of as something that is 'done to' learners rather than 'done with', 'done among' or 'done by' learners. In fact, there is no clearly stated ambition that the WBA system should support the development of learner autonomy which, as Willis (2011) points out, is a central goal of formative assessment in most school or classroom-based educational

settings. Furthermore, formative assessment has not been embedded in an overall approach to teaching and learning – official accounts of WBA offered by bodies such as the GMC (2010) and the RCR (2010) contain no discussion of formative assessment as pedagogy. It is my recommendation, based on the evidence in my research, that the WBA system in radiology would benefit from the RCR taking a broader, more ambitious view of assessment for learning, and revising its documentation, its assessment guidance and its approach to professional development for assessors in order to create a system that is capable of releasing the full potential of formative assessment through enhancements such as peer and self-assessment. This latter point – professional development of assessors – is of particular importance. As the evidence in my study has demonstrated, genuine engagement of senior radiologists with the assessment system is essential if the process is to become more than a bolt-on educational activity. This recommendation is timely, given the GMC's publication of the GMC's standards senior doctors with educational roles (GMC, 2016).

### *7.3.2 Limiting negative washback*

Even if the RCR were to take a more comprehensive approach to developing the pedagogy and practice of formative assessment, the evidence in my study suggests that there are systemic difficulties with the functioning of the WBA system in its current state. The most obvious is the negative washback that is arguably generated by the ability of powerful and influential Annual Review of Competency Progression (ARCP) panels to inspect trainees' individual WBA outcomes, including assessors' and trainees' comments. This feature of the WBA system alone appears to be working against the substantial potential for positive washback that exists within the radiology WBA context. As discussed in Chapter 2, the use of these assessments in the naturalistic clinical setting should have the effect of increasing positive washback by maximising what Messick (1996, p. 2) refers to as the 'authenticity' of the assessment. The naturalistic setting also offers the potential for increasing assessment directness i.e. limiting the introduction of assessment-related artefacts (such as simulated patients, artificially limited timeframes within which to demonstrate clinical competence and so on) which might serve to introduce construct-irrelevant variation into the assessment process.

In considering how to limit negative washback, it is difficult to see how trainees could come to trust an assessment system that is intended to be formative whilst their WBA outcomes are systematically reported to a progression panel. I contend that this development/decision duality of purpose cannot continue if the RCR are to overcome the competitive instincts of trainee doctors, and the survival instincts that are arguably common to all, and on the basis of my research I would recommend that steps are taken to address this tension.

### *7.3.3 Conceptualising 'learning' in assessment for learning*

In developing an approach to formative assessment that more closely resembles the strategies that have been demonstrated in the education literature to be effective, another important undertaking for the RCR would be to consider what their theory of learning actually is. At present, the approach to WBA in clinical radiology has been largely adopted from other medical specialties, rather than devised or created specifically for radiology training. The Rad-DOPS, for example, is merely a radiology-specific adaptation of the more generic DOPS assessment form that is found in the Foundation phase of postgraduate training, and within higher specialty training for the medical specialties i.e. the 30 specialties that belong to the Royal College of Physicians (RCP, 2016). There is no sense that the assessment system has been designed for the particular learning needs of radiologists, other than assessments being adapted to fit the particular technical requirements of the specialty.

Another key element of the educational approach that was adopted by the RCR, along with the WBA system, was the so-called competency-based design of the radiology curriculum. It is beyond the scope of this thesis to embark on an in-depth analysis of competency-based education and training, and its appropriateness for clinical radiology in the UK. It was also not my intention to conduct a detailed analysis of the radiology curriculum. However, it is important to note that a competency-based approach is an articulation of a theory of learning, whether or not the RCR and GMC have recognised this to be the case. The curriculum in radiology consists of a large number of 'competences' – individual statements of capability that have been categorised as 'knowledge', 'skills' or

'behaviours' and organised under different radiological domains. Thus the competency-based approach implies that professional capability can be resolved into component parts, and that acquisition of the component parts will therefore aggregate to produce overall competence in an individual. The stating of multiple, independently observable units of performance also implies a behaviourist epistemology, which does not necessarily align well with the concept of a doctor as a thinking, reasoning professional who employs discernment in the judicious use of a complex knowledge base. There are also important questions of whether competence resides in individuals as opposed to teams or systems, and whether competence is fixed and generic rather than being fluid and contingent. It is likely that more contemporary descriptions of the professional learning context, such as the socio-cultural perspective offered by the likes of Lave and Wenger (1991), would offer a more satisfactory account of what it is to become a radiologist.

Regardless of any broader question regarding the suitability of a competency-based curriculum for the education of professionals, competency-based approaches have implications for formative assessment. As observed by Torrance (2007), assessors in a competency-based system often perceive their feedback to be most helpful when it refers the learner to the assessment criteria. This is not necessarily 'poor practice', and in fact has been promoted by authors such as Nicol (2010) as an antidote to vague, generalised feedback, but again it is important to be aware of the consequences for learning. Such assessment practice is often described as 'convergent' (e.g. Torrance and Pryor, 2001, p. 616), and can signal the apparent redundancy of any learning that is not aligned to the assessment criteria:

Competence-based assessment is a particularly strong form of criterion referencing practised in vocational and especially work-based learning environments. What the learner can do, and can be seen to do, in relation to the tasks required of them for competent practice, are paramount. It is of little interest to the learner or assessor to identify what else the learner can do (i.e. engage in divergent assessment) although this may be of considerable importance to their longer-term development. (Torrance 2007, p 292).

In other words, a determination to reduce professional capability to large numbers of granular behavioural statements may in fact draw teachers' and learners' attention away

from the very essence of professional practice which, according to Polanyi (1967), is comprised of the tacit elements of knowledge which are resistant to being readily specified or explicitly taught. A case in point may be the OSCE exam. Hodges *et al.* (1999) found that despite the now almost unassailable high status of the OSCE as an assessment of clinical capability in medical education, senior clinicians in their study scored significantly better than junior doctors on global scales, but scored significantly more poorly than the juniors on the more detailed checklists. Thus, there may be something about the nature of expert professional performance that is not easily deconstructed and specified as 'competences'. Nonetheless, the RCR along with most other medical specialties in the UK have adopted this checklist format for the design of their curriculum and, consequently, their workplace-based assessments.

It is therefore my recommendation that, rather than simply emulating other UK medical royal colleges, the RCR should create its own concept of learning in clinical radiology, underpinned both by empirical research in clinical radiology departments and by recourse to fully developed theories of professional learning. This should then be used to develop an approach to formative assessment that is well aligned to the nature of learning in the radiology context.

#### *7.3.4 Conceptualising clinical radiology*

If it is the case, as argued above, that the professional whole is typically greater than the sum of competency-based curricular parts, an important question arises for the RCR as to how the whole may be best understood and described. In other words, as well as asking how one *becomes* a clinical radiologist, it is also necessary to ask, 'What *is* a clinical radiologist?' It may be the case that the answer genuinely is 'a person who has achieved all of the radiology curriculum competences'. Yet a socio-cultural perspective would suggest that the answer is more complex.

A parallel case might be said to be provided by the work of Oliver (2013), who observes that in the world of learning technology it is common to theorise 'learning' in sophisticated and complex ways, and yet simply to regard 'technology' as a given. As Oliver (2013)

observes, if technology is theorised at all, it tends to be done in positivistic terms, with its capabilities being characterised as ‘affordances’ – actions that it allows the learner to perform or not to perform. Thus, technology is viewed as deterministic, permitting or limiting what the user can achieve. Instead, argues Oliver, a social account of technology ‘is consistent with the constructivist, learner-centred accounts currently favoured in educational technology research...and recognises individual agency in a way that the deterministic perspective does not’ (*ibid.*, p. 35). In the case of clinical radiology, it might be argued that to define its practice positivistically, using lists of behavioural statements, is to imbue radiology with a rigidity, a determinism and a lack of reciprocity that fails to capture its true nature. A socio-cultural account, on the other hand, would acknowledge the complex, evolving professional clinical contexts in which radiology is practised and therefore allow the RCR to cast clinical radiology as ‘part of broader systems of relations, social structures, in which [it has] meaning’ (Murphy, Sharp, & Whitelegg, 2006, p. 5).

In recommending that the RCR take a socio-cultural approach in order to produce a more satisfactory account of radiological practice, it is interesting to observe the transformative effect of a socio-cultural perspective on assessment. Rather than being ‘of learning’, ‘for learning’ or even ‘as learning’, the socio-cultural viewpoint instead casts assessment as ‘practices which develop patterns of participation and subsequently contribute to [trainees’] identities as learners and knowers’ (Cowie, 2005, p. 139). It is not clear that the current radiology system does this to any significant degree, and in fact the ‘bolt-on’ feeling of the whole WBA arrangement may be detracting from radiology trainees’ experience of legitimate participation.

If the goal of assessment as a practice that supports participation were to be realised, the system that currently exists would need to be radically redesigned in light of what it means to be a clinical radiologist, how clinical radiology tends to be learned in reality, and what might best support the increasing participation of trainees in the radiological community of practice. As Willis (2011) observes, this concept of workplace-based assessment is a challenging one, and the RCR would need to be bold in breaking with established practice in UK postgraduate medical education in order to make such a change. Whether the RCR decide to take such a radical step, or to make less ambitious changes to their WBA system, it is clear that the current system is not fit for purpose and is in urgent need of reform.

## CHAPTER 8

### 8. Reflections on the study

#### 8.1 Limitations of the study

In choosing the overall approach to this study, I was aware that my research question suggests a number of possible research strategies. For example, in asking the question about the fitness for purpose of formative assessment within the clinical context, observation of the assessment encounter was one of the first possibilities that was considered. It quickly became clear that this was not going to be possible due to the sensitive nature of the setting and the number of organisations and individuals who would have to give consent in order to allow the study to proceed. A second option was an interview study, designed to explore the experiences of both the assessors and the trainees involved in the process. Unfortunately, this access could not be granted by the RCR, and so an interview study was not practicable. Consequently, the formal national e-portfolio record was the only data to which access was available and I chose to pursue approaches to data analysis that were suitable for the analysis of large scale data sets.

Having ruled out automated approaches to data analysis on the grounds that they were unable to distinguish between the types of feedback, the analysis required the reduction of the data set for the purposes of manageability. Accordingly, I decided that a sampling approach was the best way to reduce the data while retaining the integrity of the original pairs of assessor and trainee feedback comments. An attendant risk of sampling is that the sample is not representative of the population, and so a comparison of two coded samples of 500 assessments was undertaken in order to establish their representativeness, and found good agreement between the coding frequencies. Thus the reader may have a degree of confidence that the findings do generalise to the populations from which they



were taken, given that the three populations in 2010-11, 2011-12 and 2012-13 were of a similar order of magnitude.

In conducting research into the formally recorded elements of WBA in clinical radiology in the UK, the criticism may be made that my research fails to capture the informal, interactional feedback that might be expected to be present in the naturalistic clinical radiology setting. In acknowledging this limitation, I would argue that the assessment data at the heart of my study are not intended to be a surrogate for any other pedagogical interaction, which may or may not be occurring in radiology departments throughout the UK. In fact, it was into this context that the formal process of WBA was introduced in by the RCR in 2010, and it is therefore the value of this formal process that I was keen to establish. Furthermore, I would argue that it is not sufficient to merely suppose that any particular pedagogical interactions do occur, and so any claims that might be made regarding the nature of the teaching and learning environment in clinical radiology that are not themselves supported by research evidence are potentially invalid, and are not a challenge to my findings.

## **8.2 Personal reflections**

Having moved from a teaching role in school-based education to postgraduate medical education in 2009, I was intrigued to find that the approaches with which I was familiar in the school setting had recently begun to make their way into medical education. In particular, references within medical education to formative assessment and assessment for learning resonated with my previous experience and so I felt well positioned to be involved in the introduction of this approach to assessment into clinical radiology training. However, I became aware that the formative assessment process that was being introduced to medical education differed from the approaches with which I was familiar in schools. I also became aware that the consultants who were being trained to undertake the assessments with trainees appeared to be starting from quite a low base in terms of their own educational knowledge. Thus I determined to undertake empirical research in order to establish whether the new approach could be said to be working.

In starting out on my research journey, I found myself initially underestimating the complexity of the research task I had set myself. Consequently, my ideas about how the research might proceed developed from planning to undertake a fairly straightforward analysis of assessor's written feedback to include the analysis of a number of different assessment-related domains. It was also interesting to me that, having started out by intending to avoid my scientific instincts and conduct a purely qualitative study, it became clear that the best way to establish particular findings was to undertake descriptive and inferential statistical analysis of a number of relevant domains. I had not initially intended to undertake multi-methods research, and as a result of my experience I now find myself at something of a crossroads at this point in my research career: I am currently unsure whether I wish to develop further as a qualitative researcher, as a quantitative researcher, or to develop expertise in true mixed-methods research methodology. My experience through this study has heightened my awareness of each of these approaches, and in my view I have developed a range of research skills as a consequence.

Furthermore, through insights gained in this study, I feel that I have also developed in my role as an educationalist. I have extended my understanding of important principles of assessment validity, which has allowed me to bring a new critique to ongoing discussions about the future of curriculum design and workplace-based assessment in radiology and physician education in the UK. For example, I have become aware that the prevailing concept of validity in medical education, which might be termed 'validity as assessment accuracy' is somewhat narrow in comparison with validity concepts that are commonly held among school- and higher education-based researchers. Through my current role at the Royal College of Physicians, and my ongoing contact with the Royal College of Radiologists, it is my hope that these personal insights might have some impact on the future of workplace-based assessment design in medicine in the UK.

## REFERENCES

- Adler, P.A. and Adler, P. (1994) 'Observational Techniques.' In Denzin, Norman K. and Yvonna S. Lincoln, (eds.). *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications, 1994.
- Alderson, J.C. and Wall, D. (1993) 'Does washback exist?', *Applied Linguistics*, 14(20), pp. 115-129.
- Archer, J., McGraw, M. and Davies, H. (2010) 'Assuring validity of multisource feedback in a national programme', *Archives of Disease in Childhood*, 95(5), pp. 330-335.
- Assessment Reform Group (2002) *Assessment for Learning: 10 principles. Research-based principles to guide classroom practice*, Assessment Reform Group, London, United Kingdom.  
<http://webarchive.nationalarchives.gov.uk/20101021152907/http://www.ttrb.ac.uk/ViewArticle.aspx?ContentId=15313> (Last accessed August 2016)
- Augustine, K., McCoubrie, P., Wilkinson, J.R. and McKnight, L. (2010) 'Workplace-based assessment in radiology – where to now?' *Clinical Radiology*, 65(4), pp. 325-332.
- Baumeister, R.F. and Leary, M.R. (1997) 'Writing Narrative Literature Reviews', *Review of General Psychology*, 1(3), pp. 311-320.
- Baumeister, R.F., Vohs, K.D. and Tice, D.M. (1990) 'The Strength Model of Self-Control', *Current Directions in Psychological Science*, 16(6), pp. 351-355.
- BBC (2013) Foreign-trained doctors 'face GP exam discrimination', Online:  
<http://www.bbc.co.uk/news/health-23245607> (Accessed September 2015)
- Berglas, S. and Jones, E. (1978) 'Drug choice as a self-handicapping strategy in response to non-contingent success', *Journal of Personality and Social Psychology*, 36(4), pp. 405–417.
- Bernard, A.W., Kman, N.E. and Khandelwal, S. (2011) 'Feedback in the emergency medicine clerkship', *Western Journal of Emergency Medicine*, 12(4), pp. 537–542
- Bindal, T., Wall, D. and Goodyear, H.M. (2011) 'Trainee doctors' views on workplace-based assessments: Are they just a tick-box exercise?' *Medical Teacher*, 33(11) pp. 919-927

- Bing-You, RG, Trowbridge, RL (2009) 'Why medical educators may be failing at feedback', *Journal of the American Medical Association*, 302(12), pp. 1330-1331.
- Black, P. and Wiliam, D. (1998) 'Assessment and Classroom Learning', *Assessment in Education: Principles, Policy & Practice*, 5(1), pp. 7-74.
- Black, P. and Wiliam, D. (2003) "In Praise of Educational Research': Formative Assessment', *British Educational Research Journal*, 29(5), pp. 623-637.
- Black, P. and Wiliam, D. (2009) 'Developing the Theory of Formative Assessment', *Educational Assessment, Evaluation and Accountability*, 21(1), pp. 5-31.
- Black, P. and Wiliam, D. (2012) 'Developing a Theory of Formative Assessment' in *Assessment and Learning*, ed. J. Gardner, 2nd edn, London: Sage Publications Ltd., pp. 206-229.
- Bloom, B., Englehart, M., Furst, E., Hill, W. and Krathwohl, D. (eds) (1956) *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain*, New York: David McKay Co. Inc.
- Bloom, B. S., Hastings, J. T., and Madaus, G. (1971) *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.
- Bloxham, S. and Campbell, L. (2010) 'Generating dialogue in assessment feedback: exploring the use of interactive cover sheets', *Assessment and Evaluation in Higher Education*, 35(3), pp. 291-300.
- Blundell, A., Harrison, R., Turney, B.W. (2011) *The essential guide to becoming a doctor (third edition)*, West Sussex: Wiley-Blackwell.
- Boud, D. (1995) 'Assessment and Learning: Contradictory or Complementary?' in Knight, P. (ed.) *Assessment for Learning in Higher Education*, London: Kogan-Page.
- Boud, D. (2000) 'Sustainable Assessment: Rethinking Assessment for the Learning Society', *Studies in Continuing Education*, 22(2), pp. 151-167.
- Boud, D. and Molloy, E. (2013) 'Changing Conceptions of Feedback', in Boud, D. and Molloy, E. (Eds.), *Feedack in Higher and Professional Education: Understanding and doing it well*, London: Routledge, p. 11-33.

- Brann, M. and Mattson, M. (2004) 'Chapter 8: Reframing Communication during Gynaecological Exams: A Feminist Virtue Ethic of Care Perspective', in *Gender in Applied Communication Contexts*, ed. P.M. Buzzanell, H. Sterk, L.H. Turner, London: Sage Publications Ltd. pp. 147-168.
- Brannick, M.T., Erol-Korkmaz, H.T., Prewett, M. (2011) 'A systematic review of the reliability of objective structured clinical examination scores', *Medical Education*, 45(12), pp. 1181-1189
- Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77-101
- British Association of Dermatologists (2015) 'Workplace Based Assessments of Competence'. Online <http://www.bad.org.uk/healthcare-professionals/education/dermatology-specialty-trainees/workplace-based-assessments-of-competence#> (Accessed January 2016). ☑
- Buck, G. (1988) 'Testing listening comprehension in Japanese university entrance exams', *JALT Journal*, 10(1&2), pp. 15-42, in Bailey, K.M. (1996) 'Working for washback: A review of the washback concept in language testing', *Language Testing*, 13(3), pp. 257-279. ☑
- Bullock, A.D., Hassell, A., Markham, W.A., Wall, D.W. and Whitehouse, A.B. (2009) 'How ratings vary by staff group in multi-source feedback assessment of junior doctors', *Medical Education*, 43(6), pp516-520
- Burford, B., Illing, J., Kergon, C., Morrow, G. and Livingston, M. (2010) 'User perceptions of multi-source feedback tools for junior doctors', *Medical Education*, 44(2), pp165-167
- Byrne, G., Hill, J., Dornan, T., O'Neill, P. (2007) *Core clinical skills for OSCEs in surgery*. Philadelphia: Elsevier.
- Canavan, C., Holtman, M.C., Richmond, M. and Katsufrakis, P.J. (2010) 'The quality of written comments on professional behaviours in a developmental multisource feedback program', *Academic Medicine*, 85(10 Suppl), pp. S106-S109
- Carless, D. (2006) 'Different perceptions in feedback process', *Studies in Higher Education*, 31(2), pp. 219-233.
- Carless, D. (2013) 'Trust and its role in facilitating dialogic feedback', in Boud, D. and Molloy, E. (Eds.), *Feedack in Higher and Professional Education: Understanding and doing it well*, London: Routledge, pp. 90-103.

- Carracio, C., Wolfsthal, S.D., Englander, R., Ferentz, K Martin, C. (2002) 'Shifting paradigms: From Flexner to Competencies', *Academic Medicine*, 77(5), pp. 361-367.
- Case, B.J., Jorgensen, M.A., Zucker, S (2004) *Assessment report - Alignment in Educational Assessment*, San Antonio: Pearson Education.  
[http://images.pearsonassessments.com/images/tmrs/tmrs\\_rg/AlignEdAss.pdf?WT.mc\\_id=TMR S Alignment in Educational Assessment](http://images.pearsonassessments.com/images/tmrs/tmrs_rg/AlignEdAss.pdf?WT.mc_id=TMR_S Alignment in Educational Assessment) (Last accessed August 2016) ☑
- Cleland, J.A., Knight, L., Rees, C., *et al.* (2008) 'Is it me or is it them? Factors influencing assessors' failure to report underperformance in medical students', *Medical Education*, 42(8), pp. 800–809.
- Cohen, S.N., Farrant, P.B. and Taibjee, S.M. (2009) 'Assessing the assessments: U.K. dermatology trainees' views of the workplace-based assessment tools', *British Journal of Dermatology*, 16(1), pp34-39
- Cohen, L., Mannion, L. and Morrison, K. (2007) *Research methods in education* (sixth edition) London: Routledge
- Cohen, L., Mannion, L. and Morrison, K. (2011) *Research methods in education* (seventh edition) London: Routledge
- Cohen, L., Mannion, L. and Morrison, K. (2013) *Research methods in education* (seventh edition) London: Routledge
- Cook, D.A, Dupras, D.M., Beckman, T.J., Thomas, K.G. and Pankratz, V.S. (2009) 'Effect of Rater Training on Reliability and Accuracy of Mini-CEX Scores: A Randomized, Controlled Trial', *Journal of General Internal Medicine*, 24(1), pp. 74-79.
- Cowie, B. (2005) 'Pupil commentary on assessment for learning', *The Curriculum Journal*, 16(2), pp. 137-151.
- Creswell, J.W. and Plano Clark, V.L. (2007) *Designing and conducting mixed methods research*, London: Sage Publications Limited.
- Creswell, J. W. and Plano Clark, V. L. (2011) *Designing and Conducting Mixed Methods Research* (2<sup>nd</sup> edition), London: Sage Publications Limited.
- Crisp, B. (2007) 'Is it worth the effort? How feedback influences students' subsequent submission of assessable work', *Assessment & Evaluation in Higher Education*, 32(5), pp. 571-581.

- Crossley, J., Johnson, G., Booth, J. and Wade, W. (2011) 'Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales', *Medical Education*, (45)6, pp. 560-569.
- Dannefer, E.F. (2013) 'Beyond assessment of learning toward assessment for learning: educating tomorrow's physicians', *Medical Teacher*, 35(7), pp. 560-563.
- Dath, D. and Iobst, W. (2010) 'The importance of faculty development in the transition to competency-based medical education', *Medical Teacher*, 32(8), pp. 683-686.
- Davies, H., Archer, J., Southgate, L. and Norcini, J. (2009) 'Initial evaluation of the first year of the Foundation Assessment Programme', *Medical Education*, 43(1), pp. 74-81.
- Davis, D.A., Mazmanian, P.E., Fordis, M., Van Harrison, R., Thorpe, K.E. and Perrier, L. (2006) 'Accuracy of Physician Self-assessment Compared With Observed Measures of Competence: A Systematic Review', *Journal of the American Medical Association (JAMA)*, 296(9), pp. 1094-1102.
- Day, S.C., Grosso, L.G., Norcini, J.J., Blank, L.L., Swanson, D.B. and Horne, M.H. (1990) 'Residents' perceptions of evaluation procedures used by their training program', *Journal of General Internal Medicine* 5(5), pp. 421-426.
- Denzin, N.K. and Lincoln, Y.S. (2000) *The SAGE Handbook of Qualitative Research*, London: Sage Publications Ltd.
- Department of Health (1993) *'The Calman Report': Hospital doctors – Training for the future: the report of the Working Group on Specialist Medical Training*, London: Department of Health.
- Department of Health (2002) *Unfinished business: proposals for the reform of the senior house officer grade – a consultation paper*, London: Department of Health.
- Department of Health (2004) *Modernising medical careers: the next steps*, London: Department of Health. ☑
- Downe-Wambolt, B. (1992) 'Content analysis: method, applications and uses', *Healthcare for Women International*, 13(3), pp. 313-321.
- Dressel, P. (1983) 'Grades: One more tilt at the windmill', in A.W. Chickering (Ed.), *Bulletin*, Memphis: Memphis State University, Center for the Study of Higher Education.

- Dudek, N.L., Marks, M.B. and Regehr, G. (2005) 'Failure to fail: The perspectives of clinical supervisors', *Academic Medicine*, 80(10 Suppl), pp. S84-87.
- Dweck, C. S. (2000) *Self-theories: Their role in motivation, personality and development*. East Sussex: Psychology Press.
- Ecclestone, K. (2012) 'Instrumentalism and achievement: socio-cultural understandings of assessment in vocational education', in Gardner, J. (ed) *Assessment and Learning* (2<sup>nd</sup> edition), London: Sage Publications.
- Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K. and Kyngäs, H. (2014) 'Qualitative Content Analysis: A Focus on Trustworthiness', *SAGE Open*, 4, pp. 1-10.
- Ende, J. (1983) 'Feedback in Clinical Medical Education', *Journal of the American Medical Association (JAMA)*, 250(6), pp. 777-781.
- Entwistle, N. (1987) *Understanding classroom learning*, London: Hodder and Stoughton.
- Eva, K.W. and Regehr, G. (2008) "'I'll never play professional football' and other fallacies of self-assessment", *Journal of Continuing Education in the Health Professions*, 28(1), pp. 14-19.
- Fedor, D.B. (1991) 'Recipient responses to performance feedback: A proposed model and its implications', in *Research in Personnel and Human Resources Management*, in G.R. Ferris and K.M. Rowland (Eds.) Vol. 9, pp73-120.
- Fenwick, T. (2014) 'Assessment of professionals' continuous learning in practice', in S. Billett, H. Gruber & C. Harteis (Eds.), *International Handbook of Research in Professional and Practice-based Learning*. Springer
- Fernando, N., Cleland, J.A., McKenzie, H. and Cassar, K. (2008) 'Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments', *Medical Education*, 42(1), pp. 89–95. ☑
- Foundation Programme (2015) *Rough Guide to the Foundation Programme* (4<sup>th</sup> edition), London: Rough Guides Ltd.
- Frambach, J.M., van der Vleuten, C.P.M. and Durning, S.J. (2013) 'AM Last Page: Quality criteria in qualitative and quantitative research', *Academic Medicine*, 88(4), p. 552.
- Frederiksen, J. and Collins, A. (1989) 'A systems approach to educational testing', *Educational researcher*, 18(9), pp. 27-32.



- Fullan, M. (2001) *The New Meaning of Educational Change*, London: Routledge-Falmer.
- Furmedge, D. (2016) 'Written reflection is dead in the water', *BMJ Careers*, Online: [http://careers.bmj.com/careers/advice/Written\\_reflection\\_is\\_dead\\_in\\_the\\_water](http://careers.bmj.com/careers/advice/Written_reflection_is_dead_in_the_water) (Accessed September 2016)
- Gardner, J. (2012a) 'Assessment and Learning: Introduction' in *Assessment and Learning*, ed. J. Gardner, 2nd edition, London: Sage Publications Ltd., pp. 1-8.
- Gardner, J. (2012b) 'Quality assessment practice' in *Assessment and Learning*, ed. J. Gardner, 2nd edition, London: Sage Publications Ltd., pp. 103-122.
- General Medical Council (1993) *Tomorrow's Doctors: Recommendations on Undergraduate Medical Education*, London, General Medical Council. [http://www.gmc-uk.org/Tomorrows\\_Doctors\\_1993.pdf\\_25397206.pdf](http://www.gmc-uk.org/Tomorrows_Doctors_1993.pdf_25397206.pdf) (Last accessed February 2015).
- General Medical Council (2006) *Good Medical Practice: Guidance for Doctors*, London: General Medical Council. [http://www.gmc-uk.org/Good\\_Medical\\_Practice\\_Archived.pdf\\_51772200.pdf](http://www.gmc-uk.org/Good_Medical_Practice_Archived.pdf_51772200.pdf) (Last accessed August 2016)
- General Medical Council (2008) *Standards for Curricula and Assessment Systems*, London, General Medical Council. [http://www.gmc-uk.org/Standards\\_for\\_curricula\\_and\\_assessment\\_systems\\_1114.pdf\\_48904896.pdf](http://www.gmc-uk.org/Standards_for_curricula_and_assessment_systems_1114.pdf_48904896.pdf) (Last accessed August 2016)
- General Medical Council (2010) *Workplace Based Assessment: A guide for implementation*, London, General Medical Council. [http://www.gmc-uk.org/Workplace\\_Based\\_Assessment\\_A\\_guide\\_for\\_implementation\\_0410.pdf\\_48905168.pdf](http://www.gmc-uk.org/Workplace_Based_Assessment_A_guide_for_implementation_0410.pdf_48905168.pdf) (Accessed January 2016)
- General Medical Council (2013) *Good Medical Practice*, London: General Medical Council. [http://www.gmc-uk.org/static/documents/content/GMP\\_.pdf](http://www.gmc-uk.org/static/documents/content/GMP_.pdf) (Last accessed August 2016)
- General Medical Council (2016) *Promoting excellence: standards for medical education and training*, London, General Medical Council. [http://www.gmc-uk.org/Promoting\\_excellence\\_standards\\_for\\_medical\\_education\\_and\\_training\\_0715.pdf\\_61939165.pdf](http://www.gmc-uk.org/Promoting_excellence_standards_for_medical_education_and_training_0715.pdf_61939165.pdf) (Last accessed August 2016).

- Gill, D. and Griffin, A. (2010) 'Good Medical Practice: what are we trying to say? Textual analysis using tag clouds', *Medical Education*, 44(3), pp.316-322.
- Glaesser, J. and Cooper, B. (2012) 'Educational achievement in selective and comprehensive local education authorities: a configurational analysis', *British Journal of Sociology of Education*, 33(2), pp.223-244.
- Gold Guide (2016) *A Reference Guide for Postgraduate Specialty Training in the UK*, UK: COPMeD. <http://www.copmed.org.uk/publications/the-gold-guide> (Last accessed August 2016).
- Gordon, S. and Reese, M. (1997) 'High stakes testing: Worth the Price?', *Journal of School Leadership*, 7(4), pp. 345-368.
- Gormley, G.J., McCusker, D., Booley, M.A. and McNeice, A. (2011) 'The Use of Real Patients in OSCEs: A survey of medical students' predictions and opinions', *Medical Teacher*, 33(8), pp.684-687.
- Graneheim, U.H. and Lundman, B. (2004) 'Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness', *Nurse Education Today* 24(2), pp. 105-112.
- Grant, J (1999) 'The incapacitating effects of competence: A critique', *Advances in Health Sciences Education* 4: pp. 271–277.
- Greene, J.C.C., Caracelli, V.J. and Graham, W.F. (1989) 'Toward a Conceptual Framework for Mixed-Method Evaluation Designs', *Educational Evaluation and Policy Analysis*, 11(3), pp. 255-274.
- Guilford, J.P. (1936) *Psychometric Methods*, New York: McGraw-Hill.
- Haber, R.J. and Lingard, L. (2001) 'Learning oral presentation skills: A rhetorical analysis with pedagogical and professional implications', *Journal of General Internal Medicine*, 16(5), pp. 308-314.
- Harden, R.M. and Gleeson, F.A. (1979) 'Assessment of clinical competence using an objective structured clinical examination (OSCE)', *Medical Education*, 13(1), pp. 41–54.
- Harlen, W. (2007) *Assessment of Learning*, Sage Publications Ltd., London.

- Harlen, W. (2012) 'On the relationship between assessment for formative and summative purposes' in *Assessment and Learning*, ed. J. Gardner, 2nd edn, London: Sage Publications Ltd., pp. 87-102.
- Harlen, W. and James, M. (1997) 'Assessment and Learning: differences and relationships between formative and summative assessment', *Assessment in Education: Principles, Policy & Practice*, 4(3), pp. 365-379.
- Hattie, J. A. (1999) Influences on student learning (Inaugural professorial address, University of Auckland, New Zealand).  
<http://projectlearning.org/blog/wp-content/uploads/2014/02/Influences-on-Student-Learning-John-Hattie.pdf> (Last accessed August 2016)
- Hattie, J. and Timperley, H. (2007) 'The Power of Feedback', *Review of Educational Research*, 77(1), pp. 81-112.
- Herbers, J.E. Jr, Noel, G.L., Cooper, G.S., Harvey, J., Pangaro, L.N. and Weaver, M.J. (1989) 'How accurate are faculty evaluations of clinical competence?' *Journal of General Internal Medicine*, 4(3), pp. 202-8.
- Higgins, E.T. (1997) 'Beyond pleasure and pain', *American Psychologist*, 52(12) pp. 1280-1300.
- Hodges, B. (2003) 'Validity and the OSCE', *Medical Teacher*, 25(3), pp. 250-254.
- Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., Hanson, M. (1999) 'OSCE checklists do not capture increasing levels of expertise', *Academic Medicine*, 74(10), pp. 1129-34.
- Holmboe, E.S. (2004) 'Faculty and the Observation of Trainees' Clinical Skills: Problems and Opportunities', *Academic Medicine*, 79(1), pp. 16-22.
- Holmboe, E.S., Hawkins, R.E., Huot, S.J. (2004a) 'Direct observation of competence training: A randomized controlled trial', *Annals of Internal Medicine* 140: pp. 874-881.
- Holmboe, E.S., Yepes, M., Willams, F. and Huot, S.J. (2004b) 'Feedback and the mini clinical evaluation exercise', *Journal of General Internal Medicine*, 19(5Pt2), pp558-561
- Ingram, J.R., Anderson, E.J. and Pugsley, L. (2013) 'Difficulty giving feedback on underperformance undermines the educational value of multi-source feedback', *Medical Teacher*, 35(10), pp. 838-846.

- lobst W.F., Sherbino, J., Cate, O.T., Richardson, D.L., et al. (2010) 'Competency-based medical education in postgraduate medical education', *Medical Teacher*, 32(8), pp. 651-656.
- lobst, W.F. and Holmboe, E.S. (2015) 'Building the continuum of competency-based medical education', *Perspectives in Medical Education*, 4(4), pp. 165–167.
- Jackson, and Wall, (2010) 'An evaluation of the use of the mini-CEX in the foundation programme', *British Journal of Hospital Medicine*, 71(10), pp. 584-588.
- James, M. (1998) 'Chapter 9: Assessment, learning and the involvement of students', in *Using Assessment for School Improvement*, Oxford: Heinemann, pp. 171-189.
- James, M. and Lewis (2012) 'Assessment in Harmony with our Understanding of Learning: Problems and Possibilities', in J. Gardner (ed) *Assessment and Learning*, 2nd edition. London: Sage Publications Ltd., pp. 187-205.
- Job, V., Dweck, C.S. and Walton, G.M. (2010) 'Ego depletion – Is it all in your head?: Implicit theories about willpower affect self-regulation', *Psychological Science*, 21(11), pp. 1686-1693.
- Johnson, G., Barrett, J., Jones, M., Parry, D. and Wade, W. (2008) 'Feedback from educational supervisors and trainees on the implementation of curricula and the assessment system for core medical training', *Clinical Medicine*, 8(5), pp484-489.
- Johnson, R.B., Onwuegbuzie, A.J. and Turner, L.A. (2007) 'Toward a Definition of Mixed Methods Research', *Journal of Mixed Methods Research*, 1(2), pp. 112-133. ☑
- Jolly, B. and Boud, D.B. (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 D. Boud, and E. Molloy, London: Routledge, pp. 104 –124.
- Kane, M. (2006) 'Validation' in R.L. Brennan (ed.), *Educational Measurement* (4<sup>th</sup> edition). Washington DC: American Council on Education/Praeger, pp. 17-64.
- Kluger, A.N. and DeNisi, A. (1996) 'The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory', *Psychological Bulletin*, 119(2) pp. 254-284.
- Kondracki, N.L., Wellman, N.S. and Amundson, D.R. (2002) 'Content analysis: review of methods and their applications in nutrition education', *Journal of Nutrition Education and Behavior*, 34(4), pp.224-230.

- Krippendorff, K. (2004) 'Reliability in Content Analysis: Some Common Misconceptions and Recommendations', *Human Communication Research*, 30(3), pp. 411-433.
- Kruger, J. and Dunning, D. (1999) 'Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments', *Journal of Personality and Social Psychology*, 77(6), pp. 1121-1134.
- Latham, G.P. and Locke, E.A. (1991) 'Self-regulation through goal setting', *Organizational Behaviour and Human Decision Processes* 50(2), pp. 212-247.
- Lave, J. and Wenger, E. (1991). *Situated Learning. Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.
- Leung, W.C. (2002) Competency-based medical training: review, *British Medical Journal*, 325(7366), pp. 693-696.
- Loevinger, J. (1957) 'Objective tests as instruments of psychological theory', *Psychological Reports*, 3, pp. 635-694.
- Lösel, F. and Schmucker, M. (2003) Chapter 'Assessor's Bias' in *Encyclopedia of Psychological Assessment, Vol. 1* (Ed. Rocio Fernández-Ballesteros) London: Sage.
- Lum, G. (2009) *Vocational and Professional Capability: An Epistemological and Ontological Study of Occupational Expertise*, London: Bloomsbury Publishing.
- Mahoney, J., and Goertz, G. (2006) 'A tale of two cultures: Contrasting quantitative and qualitative research', *Political Analysis* 14(3): pp.227–49.
- Mainz, J. (2003) 'Defining and classifying clinical indicators for quality improvement', *International Journal for Quality in Health Care*, 15(6), pp. 523–530.
- Manikandan, S. (2011) 'Measures of central tendency: median and mode', *Journal of Pharmacology and Pharmacotherapeutics*, 2(3), pp. 214-215.
- Mansell, W., James, M. & the Assessment Reform Group (2009) *Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme*. London: Economic and Social Research Council, Teaching and Learning Research Programme.
- Margolis, M.J., Clauser, B.E., Cuddy, M.M., Ciccone, A., *et al.* (2006) 'Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study', *Academic Medicine*, 81(10 Suppl), pp. S56–S60.

- Marshall, B. and Drummond, M. J. (2006) 'How teachers engage with Assessment for Learning: Lessons from the classroom', *Research Papers in Education*, 21(2), pp. 133–149.
- MAXQDA, software for qualitative analysis, (1989-2016) VERBI Software - Consult – Sozialforschung GmbH, Berlin, Germany ☑
- McDowell, L., Sambell, K. and Davison, G. (2009) 'Assessment for learning: a brief history and review of terminology', *Improving Student Learning Through the Curriculum*, Oxford: Oxford Centre for Staff and Learning Development, pp. 56-64.
- McIntyre, D. (2000) 'Has classroom teaching served its day?' In B. Moon, M. Ben-Peretz and S. Brown (eds) *The Routledge International Companion to Education*, London: Routledge, pp. 83-108.
- McManus, I.C., Mooney-Somers, J., Dacre, J. and Vale, A. (2003) 'Reliability of the MRCP(UK) examination: 1984-2001', *Medical Education* 37(7), pp. 609-611.
- Mertens, D.M. (1998) *Research Methods in Education and Psychology: Integrating diversity with quantitative & qualitative approaches*, London: Sage Publications.
- Messick, S. (1996) *Research Report: Validity and Washback in Language Testing*, Princeton, Educational Testing Service.
- Meyer, R.E. (1995) Feedback. In: Anderson, L.W. (ed) *International Encyclopaedia of Teaching and Teacher Education*, 2nd edn. Oxford: Pergamon Press, pp249-251, in Van de Ridder, J.M.M., Stokking, K.M., McGaghie, W.C. and ten Cate, O.T.J. (2008) 'What is feedback in clinical education?' *Medical Education*, 42(2), pp. 189-197.
- Miles and Huberman (1994) Miles, M.B. and Huberman, A.M. (1994) *Qualitative Data Analysis: An Expanded Sourcebook* (2nd edition), London: Sage Publications.
- Miller, E.M., Walton, G.M., Dweck C.S., Job, V., Trzesniewski K.H., et al. (2012) 'Theories of Willpower Affect Sustained Learning', *PloS ONE*, 7(6), pp. 1-3.
- Moorhead, R., Maguire, P. and Thoo, S.L. (2004) 'Giving feedback to learners in the practice', *Australian Family Physician*, 33(9), pp. 691-695.
- Moss, F. and McManus, I.C. (1992) 'The anxieties of new clinical students', *Medical Education*, 26(1), pp. 17-20.
- Murdoch-Eaton, D. and Sargeant, J. (2012) 'Maturational differences in undergraduate medical students' perceptions about feedback', *Medical Education* 46(7), pp. 711-721.

- Murphy, P., Sharp, G. and Whitelegg, F. (2006) *Girls' experience of physics: A problem of identification and marginalization?* Buckingham, UK: Open University Press.
- Myerson, K.R. (1998) 'Can we assess professional behaviour in anaesthetists?' *Anaesthesia*, 53(11) pp. 1039-1040.
- National Association of Clinical Tutors (NACT) UK (2013) *Managing Trainees in Difficulty (version 3): Practical Advice for Educational and Clinical Supervisors*, UK: NACT. [http://www.gmc-uk.org/Final\\_Appendix\\_5\\_Trainees\\_in\\_Difficulty.pdf\\_53816759.pdf](http://www.gmc-uk.org/Final_Appendix_5_Trainees_in_Difficulty.pdf_53816759.pdf) (Last accessed August 2016).
- Nesbitt, A., Baird, F., Canning, B., Griffin, A. and Sturrock, A. (2014) 'Student perception of workplace-based assessment', *Clinical Teacher*, 10(6), pp. 399-404.
- Newton, P.E. (2007) 'Clarifying the purposes of educational assessment', *Assessment in Education*, 14(2), pp. 149-170.
- Newton, P.E. (2012) 'Validity, purpose and the recycling of results from educational Assessment', in J. Gardner (ed) *Assessment and Learning*, 2nd edition. London: Sage, pp. 264-276.
- Newton, P. and Shaw, S. (2014) *Validity in educational and psychological assessment*, London: Sage Publications Ltd.
- Nicol, D. (2009) Good designs for written feedback for students, in *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers* (13th Edition), New York: Houghton Mifflin.
- Nicol, D. (2010) 'From Monologue to Dialogue: Improving written feedback practices in mass higher education', *Assessment and Evaluation in Higher Education*, 35(5), pp. 501-517.
- Nicol, D. (2013) 'Changing Conceptions of Feedback', in *Feedback in higher and professional education: Understanding and doing it well*, ed. D. Boud & E. Molloy, Oxford: Routledge, pp. 11-33.
- Nicol, D. and Macfarlane-Dick, D. (2006) 'Formative assessment and self-regulated learning: A model and seven principles of good feedback practice', *Studies in Higher Education*, 31(2), pp. 199-218.

- Noel, G.L., Herbers, J.E. Jr, Caplow, M.P., Cooper, G.S., Pangaro, L.N. and Harvey, J. (1992) 'How well do internal medicine faculty members evaluate the clinical skills of residents?' *Annals of Internal Medicine*, 117(9), pp. 757-765.
- Norcini, J. & Burch, V. (2007) 'Workplace-based assessment as an educational tool: AMEE Guide No.31', *Medical Teacher*, 29(9) pp. 855-871.
- Oliver, M. (2013) 'Learning technology: Theorising the tools we study', *British Journal of Educational Technology*, 44(1), pp. 31-43.
- Orsmond, P., Merry, S. and Reiling, K. (2005) 'Biology students' utilisation of tutors' formative feedback: a qualitative interview study', *Assessment & Evaluation in Higher Education*, 30(4), pp. 369–386.
- Paas, F., Renkl, A. and Sweller, J. (2004) 'Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture', *Instructional Science*, 32(1-2), pp. 1-8.
- Paice, E., Aitken, M., Cowan, G., Heard, S. (2000) 'Trainee satisfaction before and after the Calman reforms of specialist training: questionnaire survey', *British Medical Journal*, 320(7237), pp. 832-836.
- Paul, A., Gilbert, K. and Remedios, L. (2013) 'Socio-cultural considerations in feedback', in *Feedback in higher and professional education: Understanding and doing it well*, ed. D. Boud & E. Molloy, Oxford: Routledge, pp. 72-89.
- Paul, S., Dawson, K.P., Lamphaer J.H. and Cheema, M.Y. (1998) 'Video recording feedback: a feasible and effective approach to teaching history-taking and physical examination skills in undergraduate paediatric medicine' *Medical Education* 32(3), pp. 332-336.
- Pedder, D. and James, M. (2012), 'Professional learning as a condition for assessment and learning' in *Assessment and Learning*, ed. J. Gardner, 2nd edn, London: Sage Publications Ltd., pp. 33-48.
- Pendleton, D., Schofield, T., Tate, P. and Havelock, P. (2000) *The consultation: an approach to teaching and learning*, Oxford: Oxford University Press.
- Perrenoud, P. (1998). 'From formative evaluation to a controlled regulation of learning: Towards a wider conceptual field', *Assessment in Education*, 5(1), pp. 85-102.
- Polanyi, M. (1967) *The Tacit Dimension*, New York: Doubleday.



- Prins, F.J., Sluijsmans, D.M.A. and Kirschner, P.A. (2006). 'Feedback for general practitioners in training: Styles, quality and preferences', *Advances in Health Science Education*, 11(3), pp. 289-303.
- Ragin, C.C. (1987) *The comparative method. Moving beyond qualitative and quantitative strategies*, Berkeley: University of California Press.
- Ragin, C.C. (2000) *Fuzzy-set social science*. Chicago: University of Chicago Press.
- Ragin, C.C. (2008) *Redesigning social inquiry: Fuzzy sets and beyond*, Chicago: University of Chicago Press.
- Ramaprasad, A. (1983) 'On the definition of feedback', *Behavioural Science*, 28(1), pp. 4-13.
- Rees, C.E., Cleland, J.A., Dennis, A., Kelly, N., Mattick, K. and Monrouxe, L.V. (2014) 'Supervised learning events in the Foundation Programme: A UK-wide narrative interview study', *BMJ Open*, 4, e005980. doi:10.1136/bmjopen-2014-005980 (Last accessed August 2016).
- Rees C.E., Knight L.V. and Cleland J.A. (2009) 'Medical educators' metaphoric talk about their assessment relationship with students: 'You don't want to sort of be the one who sticks the knife in them'', *Assessment and Evaluation in Higher Education*, 34(4), pp. 455–467.
- Robson, C. (2011) *Real World Research: A Resource for Social Scientists and Practitioner-Researchers* (3<sup>rd</sup> edition), Oxford: Blackwell Publishers Ltd.
- Royal College of Nursing (2012) RCN Factsheet: Nurse prescribing in the UK. [http://www.rcn.org.uk/data/assets/pdf\\_file/0008/443627/Nurse Prescribing in the UK - RCN Factsheet.pdf](http://www.rcn.org.uk/data/assets/pdf_file/0008/443627/Nurse_Prescribing_in_the_UK_-_RCN_Factsheet.pdf) (Accessed November 2015)
- Royal College of Physicians (2012) 'David Black: Q&A', *Commentary*, February 2012 pp. 10-13.
- Royal College of Physicians (2015) 'MRCPUK: Development of the exams', <https://www.mrcpuk.org/about-us/development-exams/past> (Last accessed 4 September 2015)
- Royal College of Physicians (2016) 'Our Specialties', <https://www.rcplondon.ac.uk/about-rcp/our-aims/our-specialties> (Last accessed August 2016)

- Royal Colleges of Physicians of Edinburgh, Glasgow and London (1996) 'Conferences on the MRCP(UK): abstracts', in Salter, R., Smith, S. (1998) 'How to pass the MRCP(UK) examination – ask a successful candidate!', *Postgraduate Medical Journal*, 74(876), pp.33-35.
- Royal College of Psychiatrists (2016) *Core Training in Psychiatry: A Competency Based Curriculum for Specialist Core training in Psychiatry*, London, The Royal College of Psychiatrists. [http://www.rcpsych.ac.uk/pdf/Core Psychiatry Curriculum August 2016.pdf](http://www.rcpsych.ac.uk/pdf/Core%20Psychiatry%20Curriculum%20August%202016.pdf) (Last accessed August 2016)
- Royal College of Radiologists (2004) *Structured Training in Clinical Radiology (Fourth Edition)*, London, The Royal College of Radiologists. <https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/STCR-2004.pdf> (Last accessed December 2015)
- Royal College of Radiologists (2007) *Structured Training Curriculum for Clinical Radiology*, London: The Royal College of Radiologists. <https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/CURRICULUM%20CR%20FINAL%20January%202009%20incorporating%20revised%20SI%20curricula.pdf> (Last accessed December 2015)
- Royal College of Radiologists (2010) *Specialty Training Curriculum for Clinical Radiology*, London: The Royal College of Radiologists. [https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/2010\\_Curriculum\\_CR.pdf](https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/2010_Curriculum_CR.pdf) (Accessed September 2015)
- Royal College of Radiologists (2014) *Specialty Training Curriculum for Clinical Radiology*, London: The Royal College of Radiologists. [https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/Curriculum CR 28 October 2014 FINAL.pdf](https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/Curriculum_CR_28_October_2014_FINAL.pdf) (Last accessed August 2016)
- Royal College of Radiologists (2015) *Specialty Training Curriculum for Clinical Radiology*, London, The Royal College of Radiologists. [https://www.rcr.ac.uk/sites/default/files/clinical\\_radiology\\_curriculum\\_2015.pdf](https://www.rcr.ac.uk/sites/default/files/clinical_radiology_curriculum_2015.pdf) (Last accessed August 2016)
- Royal College of Radiologists (2016) *History of the College* – web page. <https://www.rcr.ac.uk/college/about-college/history-college> (Last accessed August 2016)
- Rushton, A. (2005) 'Formative assessment: a key to deep learning?' *Medical Teacher*, 27(6), pp. 509-513.

- Sadler, D.R. (1989) 'Formative assessment and the design of instructional systems', *Instructional Science*, 18(2), pp. 119-144.
- Sadler, D.R. (1998) 'Formative assessment: revisiting the territory', *Assessment in Education*, 5(1), pp. 77-84.
- Saedon, H., Salleh, S., Balakrishnan, A., Imray, C. & Saedon, M. (2012) 'The role of feedback in improving the effectiveness of workplace-based assessments: a systematic review'. *BMC Medical Education*, 12(25), pp. 1-8. <http://dx.doi.org/10.1186/1472-6920-12-25> (Last assessed August 2016).
- Salter, R., Smith, S. (1998) 'How to pass the MRCP(UK) examination – ask a successful candidate!', *Postgraduate Medical Journal*, 74(876) pp. 33-35.
- Sandelowski, M. (1993) 'Rigor or rigor mortis: the problem of rigor in qualitative research revisited', *Advances in Nursing Science*, 16(2), pp. 1-8.
- Sargeant, J., Armson, H, Chesluk, B. *et al.* (2010) 'The processes and dimensions of informed self-assessment: A conceptual model.' *Academic Medicine*, 85(7), pp. 1212- 1220.
- Sargeant, J., Mann, K. and Ferrier, S. (2005) 'Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness', *Medical Education*, 39(5), pp. 497-504.
- Sargeant, J., Mann, K., Sinclair, D., van der Vleuten, C. and Metsemakers, J. (2007) 'Challenges in multisource feedback: intended and unintended outcomes', *Medical Education*, 41(6), pp. 583-591.
- Sawdon, M. & Finn, G. (2014) 'The 'unskilled and unaware' effect is linear in a real-world setting', *Journal of Anatomy*, 224(3), pp. 279-285.
- Sender-Liberman, A., Liberman, M., Steinert, Y., McLeod, P., and Meterissian, S (2005) 'Surgery residents and attending surgeons have different perceptions of feedback', *Medical Teacher*, 27(5), pp. 470-472.
- Shute, V.J. (2008) 'Focus on formative feedback', *Review of Educational Research*, 78(1), pp. 153-189.
- Sinclair, H. and Cleland, J.A. (2007) 'Medical undergraduate students—who seeks formative feedback?' *Medical Education*, 41(6), pp. 580–582.
- Stiggins, R.J. and Conklin, N.F. (1992) *In teachers' hands – Investigating the practices of classroom assessment*, Albany: State University of New York Press.

- Stobart, G. (2008) *Testing Times: The uses and abuses of assessment*, Oxford: Routledge.
- Stobart, G. (2012) 'Validity in Formative Assessment', in *Assessment and Learning*, ed. J. Gardner, 2nd edition, London: Sage Publications Ltd. pp. 233-242.
- Symonds, J.E. and Gorard, S. (2008) 'The death of mixed methods: research labels and their casualties', Paper presented at the The British Educational Research Association Annual Conference, Heriot Watt University, Edinburgh, September 3-6.  
<http://www.leeds.ac.uk/educol/documents/174130.pdf> (Last accessed 21 June 2016)
- Talbot, M. (2004) 'Monkey see, monkey do: a critique of the competency model in graduate medical education', *Medical Education*, 38(6), pp. 587-592.
- Tashakkori, A., and Teddlie, C. (1998) *Mixed methodology: Combining qualitative and quantitative approaches*, California: Sage.
- Ten Cate, O. (2013) 'Competency-Based Education, Entrustable Professional Activities and the Power of Language', *Journal of Graduate Medical Education*, 5(1), pp. 6-7.
- Teoh, N.C. and Bowden, F.J. (2008) 'A case for resurrecting the long case', *British Medical Journal*, 336(7655), p. 1250.
- Terry Page, G & Thomas, JB (1979) *International dictionary of education*, London: Kogan Page, in Van de Ridder, JMM, Stokking, KM, McGaghie, WC, ten Cate, OTJ (2008) 'What is feedback in clinical education?' *Medical Education*, 42(2), pp. 189-197.
- Thompson, T. and Richardson, A. (2001) 'Self-handicapping status, claimed self-handicaps and reduced practice effort following success and failure feedback', *British Journal of Educational Psychology*, 71(1), pp. 151-170.
- 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.
- Torrance, H. and Pryor, J. (1998) *Investigating Formative Assessment. Teaching, Learning and Assessment in the classroom*, Buckingham: Open University Press.
- Torrance, H. and Pryor, J. (2001) 'Developing Formative Assessment in the Classroom: Using Action Research to Explore and Modify Theory', *British Educational Research Journal*, 27(5), pp. 615-631.

- Van de Ridder, J.M.M., Stokking, K.M., McGaghie, W.C. and ten Cate, O.T.J. (2008) 'What is feedback in clinical education?' *Medical Education*, 42(2), pp. 189-197.
- Van der Vleuten, C.P.M (1996) 'The assessment of professional competence: Developments, research and practical implications', *Advances in Health Sciences Education*, 1(1) pp. 41-67.
- Vivekananda-Schmidt, P., MacKillop, L., Crossley, J., Wade, W. (2013) 'Do assessor comments on a multi-source feedback instrument provide learner-centred feedback?' *Medical Education*, 47(11), pp. 1080-1088.
- Wall, D. (1997) 'Impact and washback in language testing', in Clapham, C. and Corson, D. (Eds.) *Encyclopedia of Language and Education*, Netherlands: Kluwer Academic Publishers, pp. 291-302.
- Watling, C.J., Kenyon, C.F., Zibrowski, E.M., Schultz, V., Goldszmidt, M.A., Singh, I., Maddocks, H.L. and Lingard, L. (2008) 'Rules of engagement: residents' perceptions of the in-training evaluation process', *Academic Medicine*, 83(10 Suppl), pp. S97-100.
- Watling, C., Driessen, E., van der Vleuten, C.P.M., Vanstone, M. and Lingard, L. (2012a) 'Understanding responses to feedback: the potential and limitations of regulatory focus theory', *Medical Education*, 46(6), pp. 593-603.
- Watling, C., Driessen, E., van der Vleuten, C.P.M., Lingard, L. (2012b) 'Learning from clinical work: the roles of learning cues and credibility judgements', *Medical Education*, 46(2), pp. 192-200.
- Watson, M.J., Wong, D.M., Kluger, R., Chuan A. *et al.* (2014) 'Psychometric evaluation of a direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia', *Anaesthesia*, 69(6), pp. 604-612.
- Webb, G., Fawns, R., Harré, R. (2009) 'Professional identities and communities of practice', in Delany, C. and Molloy, E. (eds), *Clinical Education in the Health Professions*, Chatswood, N.S.W.: Elsevier Australia.
- Weber, R.P. (1990) *Basic Content Analysis (2nd edition)*, Thousand Oaks, CA: Sage.
- Weigle, S.C. and Jensen, L. (1997) 'Assessment issues for content-based instruction', in Snow, M.A. & D.M. Brinton, (eds.), *The content-based classroom: Perspectives on integrating language and content*, New York: Longman, pp. 201-212.
- Wiener, N. (1967) *The Human Use of Human Beings*, 2nd edition, New York: Avon.

- Weller, J.M., Jones, A., Merry, A.F., Jolly, B. and Saunders, D. (2009) 'Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: implications for implementation', *British Journal of Anaesthesia*, 103(4), pp524-530.
- Wiliam, D. (2004) 'Assessment and the Regulation of Learning', Paper presented at Invited Symposium 'What does it mean for classroom assessment to be valid? Reliable?' at the annual meeting of the National Council on Measurement in Education, April 2004, San Diego, CA.
- Wiliam, D. (2011) 'What is assessment for learning?', *Studies in Educational Evaluation*, 37(2011), pp. 3-14.
- Wiliam, D. and Thompson, M. (2007) 'Integrating assessment with instruction: what will it take to make it work?' in C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning* (pp. 53-82), New Jersey: Lawrence Erlbaum Associates.
- Wilkinson, J.R., Crossley, J.G.M., Wragg, A., Mills, P., et al. (2008) 'Implementing workplace-based assessment across the medical specialties in the United Kingdom', *Medical Education*, 42(4), pp. 364-373.
- Williams, J.G. (2007) 'Providing feedback on ESL students' Written Assignments', *TESL Journal*, 9(10) (Online: <http://iteslj.org/Techniques/Williams-Feedback.html> Last accessed August 2016).
- Williams, S.E. (1997) 'Teachers' written comments and students' responses: A socially constructed interaction', *Proceedings of the annual meeting of the Conference on College Composition and Communication*, Phoenix, AZ. <http://eric.ed.gov/?id=ED408589> Retrieved 26 September 2013.
- Willis, J. (2011) 'Affiliation, autonomy and Assessment for Learning', *Assessment in Education, Principles, Policy and Practice*, 18(4), pp. 399-415.
- Wilson, M. (2005) *Constructing Measures: An Item Response Modelling Approach*, New Jersey: Lawrence Erlbaum Associates, Inc.
- Wulf, G. and Shea, C.H. (2002) 'Principles derived from the study of simple skills do not generalize to complex skills learning', *Psychonomic Bulletin and Review* 9(2), pp. 185–211.

Yorke, M. (2005). *Issues in the assessment of practice-based professional learning: A report prepared for the Practice-based Professional Learning CETL at the Open University*. Milton Keynes, UK: Open University.

Zadeh, L. A. (1965) 'Fuzzy sets', *Information and Control*, 8(3), pp. 338–353.

## Appendix 1

# Rad-DOPS Guidance for Assessors



The Radiology Directly Observation of Procedural Skills (DOPS) focuses on the skills that trainees require when undertaking a clinical practical procedure. The DOPS is a focused observation or “snapshot” of a trainee undertaking a practical procedure. Not all elements need be assessed on each occasion. You may explore a trainee’s related knowledge where you feel appropriate.

### Instructions:

1. Please ensure that the patient is aware that the Rad-DOPS is being carried out.
2. You should directly observe the trainee performing the procedure to be assessed in a normal environment and explore knowledge where appropriate.
3. Please assess the trainee on the scale shown. Please note that your rating should reflect the performance of the trainee against that which you would reasonably expect at their stage of training and level of experience.
4. Please give an overall rating of the trainee’s performance using the options in question 13.
5. Please give feedback to the trainee after the assessment. This should include specific written comments on areas of good practise and constructive feedback on areas for further development.
6. Encourage the trainee to provide written comment on their performance and any actions required.

### Descriptors of competencies demonstrated during Rad-DOPS:

<b>Demonstrates understanding of indications, relevant anatomy and technique</b>	Does the trainee know the relevant indications, anatomical landmarks, and techniques relevant to the procedure?
<b>Explains procedure/risks to patient, obtains informed consent where appropriate</b>	Is there a clear explanation of the proposed procedure to the patient, with the patient given an opportunity to ask questions? Where informed consent is sought, is this documented appropriately?
<b>Uses appropriate analgesia or safe sedation</b>	Does the trainee use adequate amounts of appropriate drugs to minimise patient discomfort? Is this titrated where appropriate?
<b>Usage of Equipment</b>	Does the trainee show an understanding on the radiology equipment with appropriate tool/ probe selection and utilisation? Does he/she optimise equipment parameters for individual examinations?
<b>Infection prevention and control</b>	The trainee demonstrates good aseptic technique where appropriate with demonstration of principles of infection prevention and control.
<b>Technical ability</b>	Most pertinent to practical applications such as ultrasound and screening. Is there satisfactory hand/eye co-ordination?
<b>Seeks help if appropriate</b>	Does the trainee recognise his/her limitations and request assistance when appropriate?
<b>Minimises use of ionising radiation for procedures involving x-rays</b>	Where the procedure involves ionising radiation.
<b>Quality of Diagnostic images obtained</b>	The trainee tailors the number and quality of images to the procedure and patient.
<b>Communication skills with patient/staff</b>	Is the trainee polite, and exhibits a sense of self within a team structure? Is he/she able to convey understanding to others?
<b>Quality of report of procedure</b>	Does the report have a clear, concise, clinically appropriate and lucid appearance, within the context of other available clinico- radiological information?
<b>Judgement/insight</b>	For example, the trainee stops the procedure if unforeseen complications are encountered.



## Appendix 2 - Radiology Direct Observation of Procedural Skills (Rad-DOPS) form

Assessor's Registration Number

--	--	--	--	--	--	--	--

Trainee's GMC Number

--	--	--	--	--	--	--	--

Date of Assessment (DD/MM/YY)

		/			/		
--	--	---	--	--	---	--	--

Assessor's Name

--

Year of specialty training:     1    2    3    4    5    6

Clinical Setting:     Ultrasound    Computed Tomography    Paediatric Imaging    Fluoroscopy  
 MRI    Radionuclide Imaging    Interventional Radiology    Breast Imaging

Other setting:

Procedure Name:

Number of times this procedure previously performed by trainee:     0    1-4    5-10    >10

	<i>1. Well below expectation for stage of training</i>	<i>2. Below expectation for stage of training</i>	<i>3. Borderline for stage of training</i>	<i>4. Meets expectation for stage of training</i>	<i>5. Above expectation for stage of training</i>	<i>6. Well above expectation for stage of training</i>	<i>Unable to comment*</i>
<b>1. Demonstrates understanding of indications, relevant anatomy and technique</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>2. Explains procedure/risks to patient, obtains/confirms informed consent where appropriate</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>3. Uses appropriate analgesia or safe sedation/drugs</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>4. Usage of equipment</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>5. Infection prevention and control</b>	<input type="checkbox"/> Unsatisfactory		<input type="checkbox"/> Satisfactory		<input type="checkbox"/> Not applicable		
<b>6. Technical ability</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>7. Seeks help if appropriate</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>8. Minimises use of ionising radiation for procedures involving x-rays</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>9. Communication with patients/staff</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>10. Quality of diagnostic images</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>11. Judgement/Insight</b>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**12. Quality of report of procedure**

<b>13. OVERALL COMPETENCE</b>		
	<b>Rating</b>	<b>Description</b>
<input type="checkbox"/>	Trainee requires additional support and supervision	<i>Demonstrates basic radiological procedural skills resulting in incomplete examination findings. Shows limited clinical judgement following encounter</i>
<input type="checkbox"/>	Trainee requires direct supervision (performed at level expected during Core training)	<i>Demonstrates sound radiological procedural skills resulting in adequate examination findings. Shows basic clinical judgement following encounter</i>
<input type="checkbox"/>	Trainee requires minimal/indirect supervision (performed at the level expected on completion of Core Training)	<i>Demonstrates good radiological procedural skills resulting in sound examination findings. Shows good clinical judgement following encounter</i>
<input type="checkbox"/>	Trainee requires very little/no senior input and able to practise independently (performed at level expected during Higher Training)	<i>Demonstrates excellent and timely radiological procedural skills resulting in a comprehensive examination. Shows good clinical judgement following encounter</i>

**\*Unable to comment** – Please mark this if you have **not observed** the behaviour and feel unable to comment.  
**Further mandatory questions on the following page**

**Assessor's comments – state areas of good practice and areas for development (mandatory field)**

**Trainee's comments – comment on your performance and any actions required (mandatory field)**

### Appendix 3 – Clinical Radiology ARCP Decision Aid, 2010

**ARCP Decision Aid** The following decision aid grids offer guidance on the domains to be reviewed and level of attainments suggested to inform an ARCP panel.

Standards for satisfactory progression:

	ST1	ST2	ST3	ST4	ST5
Curriculum coverage: Generic	20-30% focus area content at core level descriptor	50% focus area content at core level descriptor	Competent in all focus area content at core level descriptor	50% focus areas content at Level1/2 descriptors	Competent in all focus area content at Level1/2 descriptors
Curriculum coverage: Radiology Specific	20-30% common presentations at core level descriptor	60% common presentations at core level descriptor	90% common presentations at core level descriptor	Complete common presentations L1 – Special interest area(s)	L2 – Special interest area/multiple L1 interest areas
Indicative Workplace based Assessments/yr	6 mini-IPX (minimum 2 per clinical attachment), 6 Rad-DOPS (minimum 2 per clinical attachment), 1 MSF, 1 Audit Assessment, 2 Teaching Observations WpBA should be undertaken in a timely and educationally appropriate manner <b>throughout</b> the training year. Progression predicated by satisfactory anchor statements				
Examinations	First FRCR Examination	Final FRCR Part A Examination: three modules	Final FRCR Part A Examination: all six modules	Final FRCR Part B Examination	--
Education Supervisor's Structured Report	All areas of personal and professional development addressed with overall progress at expectation or above.				

## **Appendix 4 – Evolution of the coding framework for assessors' comments**

### **Phase 1 - Initial inductive coding of assessors' feedback comments**

30/10/2011

#### Tentative coding ideas

Positive unconditional  
Positive conditional  
Negative conditional  
Negative unconditional

Within positive conditional or negative conditional;

Specific – refers to specific actions or behaviours of the trainee  
Linked – specific feedback linked to the assessment criteria  
Clinical – specific feedback linked to clinical practice

Within specific:

Knowledge  
Attitude  
Behaviour  
Skill

### **Phase 2 – Review of initial codes and further inductive coding**

28/11/2011

#### New codes

Linked to assessment criteria  
Clinically relevant  
Communication, positive comment  
Communication, negative comment  
Competence, positive comment  
Competence, negative comment  
Confidence, positive comment  
Confidence, negative comment  
Independence, positive comment  
Independence, negative comment  
Insight, positive comment  
Insight, negative comment

Unspecific positive comment  
Unspecific negative comment

30/11/2011

New codes created:

Recommendation for improvement – unspecific  
Recommendation for improvement – specific

**Phase 3 – Further review and inductive coding**

05/12/2011

New code ideas:

Assumed improvement  
Technical capability  
Reference to stage of training  
Reference to safety  
Reference to experience?  
Reference to global development ie overall progress to date?

**Phase 4 – Piloting a modified version of Canavan et al.'s (2010) framework**

31/01/2012

1. Global assessment (non-specific, directed at the self) e.g. *great guy, a good trainee etc.*

[Consider

- *global positive*
- *global negative] – rejected. Use overlapping 'valency' codes to indicate this.*

2. Behavioural

- *general behaviour e.g. good communication skills, not a team player*
- *specific behaviour e.g. managed to insert the needle at the first attempt*

[consider developing 'specific' into

- *knowledge*
- *skill*
- *attitude]*

3. 'Valency' of feedback

- positive
- negative

4. Linked (explicitly mentions one or more of the assessment criteria)

5. Suggestions for improvement

- Recommendations (general) e.g. *get more experience; see more patients* or where the trainee would need further clarification in order to know how to improve e.g. *must learn to insert the needle at the first attempt*
- Recommendations (specific) i.e. A clear recommendation as to specific actions that can be undertaken by the trainee in order to improve / action plan generated e.g. *“do 5 more of these under supervision and then complete another Rad-DOPS”*

6. Descriptive – the comment offered is limited to a descriptive account of the procedure being assessed with no impression of what has gone well or otherwise, no reinforcement of desirable behaviour, and no suggestions for improvement

7. Dismissive – either of the trainee or the training/assessment process; insubstantial; joking etc.

20/08/2012

New code created:

Overall procedure – assessor comments on the whole procedure rather than separate components

24/08/2012

New code created:

Assumed improvement – assessor expresses the notion that the trainee's skills will improve with more time/practice/experience etc.

'Overall procedure' code abandoned – not parsimonious as it overlapped with 'general comments on observed performance.

## Final coding framework

Code	Criteria for applying code to assessors' comments
Valency	
Positive	The comment was clearly intended to be positive in nature
Negative	The comment was negative in nature. This included any suggestion that improvement would be necessary, however constructively expressed.
Performance	
General comment on observed performance	The assessor commented on an aspect of the trainee's performance in a manner that may have required further explanation
Specific comment on observed performance	The assessor made a comment that was sufficiently clear as to make it unlikely that the trainee would have needed further explanation
Linked to assessment criteria	The comment clearly invoked one or more of the assessment criteria on the Rad-DOPS form
Descriptive	The comment is limited to a description of the procedure undertaken by the trainee, and lacks any judgement of their performance or suggestions for further development
Developmental.	
Specific recommendation	The assessor made a suggestion for improvement that is unlikely to need further clarification
Unspecific recommendation	The assessor made a suggestion for improvement that was unclear or ambiguous
Personal	The comment referred to some aspect of the trainee's personality or personal qualities
Global assessment	The comment referred to the trainee's overall progress within the training post
Assumed improvement	The assessor made a comment to the effect that time, or experience, or continued practice would necessarily bring about improvement
Absent	The assessor failed to provide a comment

## Appendix 5 – Confirmation of ethics approval



School of Education  
Research Office  
Queen's University Belfast  
Belfast  
BT7 1HL  
Tel +44 (0) 28 90975923  
Fax +44 (0) 28 90975066  
[www.qub.ac.uk](http://www.qub.ac.uk)

### Memorandum

⊕

To	Michael Page
From	Ulrike Niens, Chair, Ethics Committee
Date	4 October 2011
Distribution	John Gardner, Supervisor School of Education Office File
Subject	Ethics Approval

The School of Education Ethics Committee has approved your proposed research.

Note that this approval applies only to the procedures outlined in your submission.

Any departure from these must be discussed with your supervisor, and may require additional ethical approval.

**Note for the supervisor:** it is the responsibility of the supervisor to add any research projects involving human participants, material or data, to the University's Human Subjects Database for insurance purposes. (The Human Subjects Database is accessible through QOL under 'My Research').

The Committee wishes you every success with your research.