Relatively Idiosyncratic
Exploring variations in assessors' performance judgements within medical education

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy (PhD)
Faculty of Human and Medical Sciences
2012

Peter Yeates

School of Translational Medicine / Division of Respiratory Medicine
# Contents

Declaration: ............................................................................................................................ 5  
Copyright Statement: ............................................................................................................. 5  
Acknowledgements................................................................................................................ 6  
Brief candidate biography ...................................................................................................... 7  
Format of thesis ..................................................................................................................... 7  
Abstract ...................................................................................................................................... 8  

## Chapter 1: Background and Literature Review ................................................................. 9  
  Introduction ........................................................................................................................... 9  
  Performance assessments within healthcare education ...................................................... 9  
    Defining assessment .......................................................................................................... 9  
    The various purposes of assessment ............................................................................... 10  
    Approaches to authentic assessment .............................................................................. 11  
  Review of the literature ....................................................................................................... 14  
  Direct Observation Performance Assessments ................................................................. 14  
    Reliability of direct-observation performance assessments ............................................ 15  
    Components of variance within observed scores ............................................................ 16  
    Assessor score differences ............................................................................................... 17  
    Relationship of scores to external variables .................................................................... 18  
    Internal structure of score data ....................................................................................... 21  
    Perceived educational impact of direct observation performance assessments ............ 24  
    Attempts to improve assessment judgements ............................................................... 28  
    Summary of issues ........................................................................................................... 32  
  Performance Appraisal and Impression Formation .......................................................... 33  
    Section introduction ......................................................................................................... 33  
    Judgements as cognitive categorisations......................................................................... 34  
    Organisation of categorical thinking .............................................................................. 36  
    Implications of categorical judgement .......................................................................... 38  
    Biases arising from categorical thinking ........................................................................ 39  
    Conditions that promote category-overuse ................................................................... 41  
    Attribution processes ....................................................................................................... 44  
    Judgement as a social process .......................................................................................... 45  
    Summary of issues from performance appraisal and impression formation .................. 49  
Rater cognition in Medical Education .................................................................................. 50  
  Summary of issues in rater cognition within medical education ....................................... 56  
  Formulation of Research Aims: ............................................................................................ 57  

## Chapter 2: General Methodology ....................................................................................... 60  
  Introduction ......................................................................................................................... 60  
  Theoretical orientation of study .......................................................................................... 60  
  The nature of cognition ....................................................................................................... 60  
  Approaches to the study of cognition .................................................................................. 61  
    Choice of study population and context ......................................................................... 62  
    Consideration of Methodological approaches .................................................................. 66
Chapter 3: Seeing the same thing differently ................................................................. 89
  Developing study research question .............................................................................. 89
  Methodological Considerations in Study 1 ..................................................................... 91
Study 1: Seeing the same thing differently. Mechanisms that contribute to assessor
differences in directly-observed performance assessments ............................................. 94
  Additional results from study 1 ..................................................................................... 95
  Use of inference: .............................................................................................................. 95
  Non-performance factors: ............................................................................................... 96
  Conclusion: ......................................................................................................................... 101

Chapter 4: Mind who you stand next to ........................................................................ 102
  Formulating research aim ............................................................................................... 102
  Review of relevant literature ......................................................................................... 103
  Comparative influences on judgement ......................................................................... 103
  Contrast-relativity effects ............................................................................................. 104
  Anchoring assimilation effects ..................................................................................... 106
  Contrast relativity effects versus assimilative anchoring effects ................................... 109
  Additional methods for chapter 4 .................................................................................. 111
Study 2: Mind who you stand next to: Determining the influence of raters’ recent-
experiences on assessment judgments ............................................................................. 116

Chapter 5: You’re certainly relatively competent ........................................................... 117
  Introduction ..................................................................................................................... 117
  Rationale for further study ............................................................................................ 117
  Brief literature review on the role of confidence in judgements .................................... 121
  Confidence and the influence of context ....................................................................... 124
Study 3: You’re certainly relatively competent. The influence of recent experience, perceived
rank and confidence on assessors’ scores ......................................................................... 128

Chapter 6: General Discussion .......................................................................................... 128
  Summary of findings ....................................................................................................... 128
  Differential Salience ....................................................................................................... 128
  Criterion Uncertainty ..................................................................................................... 129
  Information Integration .................................................................................................. 129
  Idiosyncrasy within judgements ................................................................................... 130
  Influence of comparative judgements .......................................................................... 130
Influence of assessor characteristics on scores ............................................................. 130
Contrast effects on borderline performances ............................................................. 131
Influence of assessors’ “Hawk / Dove” differences ....................................................... 131
Perceived Rank of Performances ................................................................................... 132
Confidence in judgements ............................................................................................. 132
Theoretical implications of findings ................................................................................... 133
Cognitive schemas and idiosyncrasy .............................................................................. 133
Assessment criteria ........................................................................................................ 134
Judgemental “fuzziness” ................................................................................................. 136
Limitations.......................................................................................................................... 140
Remainder issues and future research ............................................................................. 142
Replication and extension of findings ............................................................................. 142
Social and interpersonal influences on judgement........................................................ 146
Organisational culture in which assessments occur ...................................................... 148
Implications for practice .................................................................................................... 150
Idiosyncrasy versus error, and different theoretical perspectives .............................. 150
Criterion or norm referencing of cut scores .................................................................. 151
Variations in competence-criteria between hospital settings ....................................... 152
Potential implications for OSCE exams ....................................................................... 153
Implications for improving uniformity of assessors’ judgements ............................... 154
Conclusions ........................................................................................................................ 156
References ................................................................................................................... 158
Appendices .................................................................................................................. 177
Appendix 1: Search terms used for literature review .................................................... 177
Appendix 2: MiniCEX mark sheet from the UK foundation Program............................. 178
Appendix 3: Characteristics of “Good” and “Poor” performance, and qualitative performance descriptors used in scripting cases .................................................... 179
Appendix 4: Example scripts for performances ............................................................. 181
Appendix 5: Consent form and participants information sheet for study 1.................. 192
Appendix 6: Sample transcript from study 1 ................................................................. 197
Appendix 7: Participants information sheet for study 2 ................................................ 222
Appendix 8: Data tables from study 2 ............................................................................ 224
Appendix 9: Participant information sheet for study 3 ................................................. 227
Appendix 10: Data tables from study 3.......................................................................... 230

List of Tables and Figures:

Table 1: Review panel ratings of scripts 82
Table 2: Review panel’s ratings of videos 86
Table 3: Reviewers ratings of videos for Case 1 (Respiratory) 113
Table 4: Reviewers ratings of shortened videos for Case 2 (Loss of consciousness) 114
Table 5: Reviewers ratings of shortened videos for Case 3 (Upper GI bleeding) 114

Figure 1: Still image from video for Case 2, Good Performance 85

Word count: 60,403
Declaration:

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement:

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

My profound thanks to my supervisors, Paul and Karen, for their guidance, tuition, patience and support over the last three years (as well as the time before that).

Thank you also to Kevin Eva for helping so much through his collaboration in this work; for his insight, tuition and patience.

Thank you to Jo Hart for being my advisor over the last two and a half years; thank you for insisting on such a strict timeline and making me stick to it. Seems to have worked!

It has been both a pleasure and a privilege to work with and be tutored by such eminent, knowledgeable, and nurturing people. Thank you.

.....

My sincere thanks to the UHSM Academy, and in particular to Professor Ged Byrne for providing salary and practical support for me during this thesis. Thank you for taking a chance on me.

Thank you to the Association for the Study of Medical Education (ASME) travelling fellowship committee for their award of a scholarship that supported this work.

.....

Thank you to the foundation doctors and simulated patients who participated in filming the videos used through the programme of study, and to the clinicians who gave their time to help in the videos development.

Thank you to the United Kingdom Foundation Programme and each of the regional foundation schools who helped in recruitment and in giving the work the endorsements it needed.

Thank you to all of the three studies’ participants who, despite being busy people, gave their time for no reward. I hope that this thesis means that your efforts will have a suitable impact on the assessment conversation in the UK and around the world.

Thank you to all of the colleagues and informal contacts whose conversations have stimulated and provoked me. Thank you to Julie Morris and Sigrid Whiteside for helping with my myriad statistical queries.

.....

Thank you lastly to Kathryn for her patience, love and support, as well as to Mum, Dad and Andrew. Andrew, the very last line is for you.
Brief candidate biography

I undertook medical undergraduate training at Newcastle University between 1995-2001, including an intercalated BSc in clinical pharmacology. I became involved in medical education in 2006, working as a teaching and education fellow in Northumbria Healthcare trust. I then moved to Manchester to take up an academic clinical fellowship in medical education, during which time I undertook a masters of clinical education (MClEd). I began this programme of study in October 2009.

Clinically I undertook general pre-registration training, then trained through general internal medical training, completing the Royal College of Physicians exams. I have subsequently specialised in respiratory medicine.

Format of thesis

The thesis presented here is organised in “alternative format”. This is a recognised and accepted thesis format under the University of Manchester’s regulations for doctoral degrees. Unlike the conventional thesis format, the alternative format permits that some sections included in the thesis are in a format suitable for publication. These sections are not required to be published, or even accepted for publication; they are simply required to be formatted in manner suitable for publication.

All the work presented in this thesis is primarily the work of the submitting candidate, Peter Yeates. Readers will note that the three papers included in the thesis are the work of multiple authors. Consequently it is necessary to clarify the role that Peter Yeates contributed to the research.

Peter Yeates took the lead role in design, conduct and analysis of all of the research presented in this thesis, under the supervision and with the assistance of his supervisors Paul O’Neill and Karen Mann. Kevin Eva, who collaborated on each of the three studies contributed to the design and analysis of the studies. Peter Yeates wrote the initial drafts of the entirety of the thesis. Sections other than those in a format suitable for publication were primarily written by Peter Yeates, with editorial comments and suggestions from both supervisors, and slight editing of contents. Each of the three papers in the thesis was initially written by Peter Yeates. In keeping with their multiple authorships, these papers then underwent iterative revisions involving all authors. The papers for studies 1 & 2 have also been revised in view of comments made by peer reviewers. None the less, Peter Yeates took a lead role throughout these revisions. As a result, all of the presented work is predominantly attributable to his work.

In line with regulations, sections that are multiply authored are identified, and are paginated separately from the main body of the thesis text.
Abstract

The University of Manchester, Peter Yeates, September 2012.
Thesis submitted for the degree of Doctor of Philosophy (PhD).

Title: Relatively idiosyncratic: Exploring variations in assessors’ performance judgments within medical education.

Background: Whilst direct-observation, workplace-based (or performance) assessments, sit at the conceptual epitome of assessment within medical education, their overall utility is limited by high-inter-assessor score variability. We conceptualised this issue as one of problematic judgements by assessors. Existing literature and evidence about judgements within performance appraisal and impression formation, as well as the small evolving literature on raters’ cognition within medical education, provided the theoretical context to study assessor’s judgement processes.

Methods and Results: In this thesis we present three studies. The first study adopted an exploratory approach to studying assessors’ judgements in direct observation performance assessments, by asking assessors to describe their thoughts whilst assessing standard videoed performances by junior doctors. Comments and follow up interviews were analysed qualitatively using grounded theory principles. Results showed that assessors attributed different levels of salience to different aspects of performances, understood criteria differently (often comparing performance against other trainees) and expressed their judgements in unique narrative language. Consequently assessors’ judgements were comparatively idiosyncratic, or unique.

The two subsequent follow up studies used experimental, internet based, experimental designs to further investigate the comparative judgements demonstrated in study 1. In study 2, participants were primed with either good or poor performances prior to watching intermediate (borderline) performances. In study 3 a similar design was employed but participants watched identical performances in either increasing or decreasing levels of proficiency. Collectively, the results of these two studies showed that recent experiences influenced assessors’ judgements, repeatedly showing a contrast effect (performances were scored unduly differently from earlier performances). These effects were greater than participants’ consistent tendency to be either lenient or stringent and occurred at multiple levels of performance. The effect appeared to be robust despite our attempting to reduce participants’ reliance on the immediate context. Moreover, assessors appeared to lack insight into the effect on their judgements.

Discussion: Collectively, these results indicate that assessors score variations can be substantially explained by idiosyncrasy in cognitive representations of the judgement task, and susceptibility to contrast effects through comparative judgements. Moreover, assessors appear to be incapable of judging in absolute terms, instead judging normatively. These findings have important implications for theory and practice and suggest numerous further lines of research.
Chapter 1

Background and Literature Review

Introduction

In order to situate this programme of study, this chapter will cover a number of pertinent issues. Firstly, it will describe the nature of assessment within healthcare education, explaining the position and purpose of workplace based (or performance) assessments within overall assessment strategies. Secondly it will review relevant literature. The literature review will initially consider performance assessments in medical education - illustrating strengths and weaknesses of these assessments. In so doing it will make a case for studying the judgement processes used by assessors. Next the review will illustrate existing knowledge from wider fields of psychology that are useful in understanding (in more general terms) the ways that humans make judgements about other humans and their performance. This will involve consideration of the “impression formation” and “performance appraisal” literatures. Finally the literature review will summarise the small but rapidly evolving field of “rater cognition” within medical education. To conclude, the reviewed principles will be formulated to develop research aims for the programme of study.

Performance assessments within healthcare education.

Defining assessment

Assessment in medical education has been defined as “a systematic procedure for measuring a trainee’s progress or level of achievement, against defined criteria to make a judgement about a trainee” (PMETB 2008). Whilst this definition has merit it may be noted that various accepted forms of assessment lie outside its parameters: formative assessment may not be systematic; criteria are sometimes not explicitly defined and assessments may
be used for re-certification of clinicians who are no longer trainees. These issues notwithstanding, the definition highlights that assessment seeks to ascertain and describe the ability or performance of an individual.

The various purposes of assessment

The purpose of assessment in medical education has been debated extensively over recent decades. The traditional view of assessment sought to ensure that individuals met an acceptable standard; that assessment comprised a formal set of processes which were used to determine whether an individual could be allowed to progress beyond their current level. Thus it was summative in nature (assessment of learning). Increasingly the formative value of assessment (or assessment for learning) has been recognised, with the intention that feedback from assessments should guide development (PMETB 2009). This has become particularly salient as research into the development of expertise has highlighted the need for accurate feedback (Ericsson 2004), and investigation of self-assessment has shown that individuals often have limited insight into their own performance ability (Davis et al. 2006). More contemporary concepts of assessment view it as “programmatic”: integrating various forms of information gathered throughout the curriculum to provide the basis for sound judgment on an individual, whilst offering much useful developmental feedback along the way (van der Vleuten & Schuwirth 2005).

Concurrent with changes in the purpose of assessment, our collective view of what should be assessed has also evolved. Traditional assessments focused on testing the acquisition of knowledge (Wass et al. 2001). Whilst knowledge remains foundational – especially in view of the context specificity of medical knowledge (Elstein 2009) – attention has moved to ensuring that this knowledge could be applied to clinical problems, and that adequate skills could be demonstrated (Harden et al. 1975). Methods of measuring applied knowledge and
skills followed. Whilst these methods have endured and been refined with time, they have been criticised on the basis that assessment conducted in this manner leads to both atomisation and decontextualisation of aspects of complex practice into discrete measurable units (Huddle & Heudebert 2007). This in turn led to an aspiration to measure individuals’ authentic performance in real practice. Miller’s hierarchy of assessment purposes (Miller 1990) views assessment of real performance as the epitome of assessment within medical education. Other authors have expressed similar aspirations to measure authentic or integrated real performance (Epstein & Hundert 2002). Authentic assessment of real performance is seen conceptually as giving the richest insight into individuals’ professional competence.

Approaches to authentic assessment

Whilst authentic assessment (or its synonyms: workplace-based (WPBA) or performance assessment) can be viewed conceptually as the epitome of assessment within healthcare education, a number of issues emerge in terms of the practicalities of how best to measure authentic performance. In a discussion article on the subject, Norcini (2003) debated a number of conceptual approaches – monitoring of outcomes, observations of performance and analysis of records of process. Problems exist with each of these: due to team-based care, patient outcomes are rarely attributable to individual clinicians; analysis of records is time consuming, especially where records are paper-based; and observation of performance may well alter trainees’ performance (sometimes referred to as “Hawthorne effect”). Indeed Norcini (2003) recommended that observations of performance should ideally be either covert or continuous to avoid this influence.

In comparison to these aspirations, our current methods of assessing workplace-based performance represent a pragmatic compromise. In broad terms two methods are used:
direct observation performance assessments and multi-source feedback. The former of these involves observation of routine clinical tasks by an assessor who, on each occasion, judges the observed performance, allocates scores and provides feedback. Scores are usually recorded on a Likert scale for a number of different “domains” of clinical practice – notionally discrete aspects of practice – and feedback is documented along with a plan for development. Usually a number of assessments are required over a period of training, usually with different assessors and with different types of clinical case material. This process fulfils the intent of authentic assessment in some ways: it involves real patients, with real clinical problems as they present to the healthcare setting in which the trainee works. Therefore, it should avoid much of the decontextualisation that is observed in assessment formats that use simulations of clinical scenarios – for example objective structure clinical exams (OSCE). Nonetheless trainees are aware they are being observed, and may alter their practice accordingly. Moreover cases are often selected by trainees, and thus may be chosen to represent their strengths rather than their weaknesses, and consequently may not be a representative sample of a trainee’s actual work.

By contrast, multi-source feedback involves obtaining a triangulated, retrospective view of the performance of an individual over time (Lockyer 2003). Opinions on the individual’s performance are sought from a range of people, who are deemed able to observe the individual in various aspects of their daily work. In medical education this often includes other trainees of the same or different grade; senior clinicians; nursing staff and allied healthcare workers or administrative or support staff. Some schemes also seek the views of patients. Individuals who provide their opinion do so by recollecting their impressions over time. Thus they may include the totality of their experience of the individual over that time, taking into account all the clinical and non-clinical experience the observer has of the target individual. These judgements are again documented as scores using a Likert scale, along with
space for free text feedback. Typically multi-source feedback involves quite a large number of score domains – often ten or more. Whilst multisource feedback may overcome the issue of Hawthorne effect, and include a larger sample of clinical material and a wider variety of judgements, it still has vulnerabilities. Some of the assessors may have limited experience of the individual being assessed, or there may be aspects of an individual’s practice that are unobserved. Moreover, assessors may incompletely recollect what they have observed. Consequently impressions may be quite general and feedback may lack specificity, with important implications for its developmental utility.

Consequently – despite the conceptual attraction of measuring what an individual “does” in their clinical practice – neither of the major currently-employed methods completely achieves this intent. In exploring authentic assessments further, this thesis will focus on one of these two types of assessment – namely on direct observation performance assessments, and will not give further consideration to multi-source feedback.

Numerous different types of direct-observation performance assessment exist (Pelgrim et al. 2011; Kogan et al. 2009). They vary in terms of the aspects of clinical performance they are used to assess, the level of trainees to whom they apply, the format of the scoring sheets that they employ, the domains of competence which they assess and the intended purpose of the scores and feedback which they generate. None the less, they have an important commonality which makes it meaningful to consider them together: they all involve judgement by an assessor on a discrete instance of clinical performance by a trainee for the purpose of assessment. Consequently, it is useful to review the existing literature concerning direct-observation performance assessments, both to summarise existing knowledge and to situate the subsequent research. The following literature review aims to be comprehensive but not exhaustive; papers were selected to develop the described argument whilst being
representative of the current state of knowledge in this field. Papers were identified by means of literature searching in MEDLINE (search terms are described in appendix 1) and through the reference lists of selected papers.

Review of the literature

As described earlier, this literature will firstly consider direct observation performance assessments, before considering two literatures relevant to making performance judgements: impression formation and performance appraisal. Literature on rater cognition in medical education will then be reviewed.

Direct Observation Performance Assessments

A number of authors have reviewed direct-observation performance assessments within medical education. Kogan et al (2009) found 55 different direct-observation tools described in the literature. Of these a large number had only been evaluated in terms of user attitudes, whereas more data was available to support the validity of one of the most prevalent formats known as the Mini Clinical Evaluation Exercise (Mini-CEX). Reviews by Hawkins et al (2010) and Pelgrim et al (2011) reached similar conclusions. The first of these focused specifically on Mini-CEX assessments, whereas the latter considered all direct-observation performance assessments in medical education. Both reviews found much evidence to support the overall utility of the assessments, but noted high inter-assessor variability in scoring, which they viewed as the chief area of concern. A number of the specific findings that arise from this literature are worthy of detailed consideration. The following review of the literature, whilst not attempting to be exhaustive, will describe and illustrate such themes.
Reliability of direct-observation performance assessments

The reliability of a measurement can be defined as the ratio of subject variance to total variance in observed data relating to the subject (Streiner & Norman 2008, p169). As total variance includes both variance due to the subject and variance due to error, this represents the precision of the measure relative to the amount the subject itself varies. Its value is always between 0 (completely unreliable) and 1 (completely reliable).

The reliability of several direct-observation performance assessments has been investigated across a number of contexts. Typically these analyses use generalisability theory (Crossley et al. 2002) to estimate the variance components of each facet of the assessment, before manipulating the data in a decision study to determine the number of assessment encounters (occasions on which a trainee is observed and assessed) required to achieve a specified level of reliability (Streiner & Norman 2008 p215-223). Thresholds for acceptable reliability are typically set at either 0.7 or 0.8, indicating that 70% or 80% respectively of the observed variance in scores is due to differences in trainees performance, whereas the remainder of observed score variance is attributable to error. Estimates vary across contexts. At the lower end, 3 observations achieved a reliability of 0.8 for procedural skills assessments in two separate studies (Beard et al. 2011; Wilkinson et al. 2008), and 5 observations were adequate to achieve the same level for clinical encounter assessment (Wilkinson et al. 2008). More typical estimates are in the region of 8 assessments (Nair et al. 2008), 10-12 assessments (Cruess et al. 2006), or 12-14 assessments (Norcini et al. 1995). Kogan et al (2003) showed that 8 raters were needed to achieve a reliability of 0.77. After reviewing the topic, Pelgrim et al (2011) concluded that 10 encounters was the most generally applicable estimate of the number of encounters needed to ensure adequate reliability.
Whilst encounter estimates in this range predominate, a few exceptions should be noted. McGill and colleagues (2011) estimated that 28 encounters would be required to reach a reliability of 0.8, although estimates for some individual competencies were lower. Alves de Lima et al (2007) estimated that 50 encounters were required to achieve the same level, although a follow-up study by this group in a laboratory setting (artificial assessments of standard performance videos rather than naturalistic data) needed only 9 encounters to achieve the same reliability (Alves de Lima et al. 2011). A study by Weller et al (2009) estimated that in a scenario where different assessors assess each trainee, 15 assessors, each assessing 5 encounters with each trainee would only achieve a reliability of 0.49. This improved to 0.73 when 15 assessors perform 5 encounters each on the same pool of trainees.

In summary, two observations from this literature are important: reliability estimates in the region of 10 encounters are the most prevalent, and estimates can vary substantially across contexts. From these observations we may infer that reliability is not an inherent parameter of this type of assessment, but that it varies substantially across settings. This of course is no surprise: classical test theory has always viewed reliability as property of a test as it was administered in a particular context with particular group of learners, rather than as an immutable or fixed characteristic of the assessment (Streiner & Norman 2008 p173). None the less, it is salient to notice how great the magnitude of these changes can be between contexts for direct observation performance assessments.

Components of variance within observed scores

A smaller number of studies have partitioned score variance estimates into their constituent components. Most naturalistic studies preclude this option as all facets of interest (case, assessor and trainee) are “nested” – or observed simultaneously in a manner that prevents comparison of their relative contribution to score variance. Consequently these studies are
less common. Weller et al (2009) found that “trainee ability” accounted for only 4% of score variance, whilst assessors’ stringency or leniency accounted for 40%. Case specificity (the tendency for trainees to perform differently on different clinical cases) accounted for a further 41% of score variance. Margolis et al (2006) used a fully crossed design in which the use of videoed performances enabled all assessors to view all trainees on every case. Variance components due to assessors were consistently higher than for trainees. Percentages are not quoted in their study but can be calculated from their presented data. Based on the “overall clinical competence” ratings, assessor-variance comprised 21% of the observed score variance and trainees comprised 20%. In contrast to Weller et al, there was very little case specificity (2%). Similarly, Wilkinson et al (2008) showed that trainees accounted for 15% of score variance, with assessors producing 21% of score variance. This partially nested design didn’t enable estimates of case specificity. Alves de Lima et al’s (2011) study is the only one of this nature in which trainees contributed a greater proportion of score variance (28%) than assessors (18%), although the influence of assessors is similar to that in both Margolis’ and Wilkinson’s studies. Alves de Lima et al’s study (similar to Margolis, but in contrast to Weller) showed a small influence of case specificity (9% of score variance). Thus again, the relative contributions to observed score variance of trainees, cases and assessors (and indeed their various interactions) vary across contexts. What is consistent across these studies is that assessors contribute a substantial component of the observed score variance.

**Assessor score differences**

In a different approach, other studies have looked at the differences in individual assessors’ mean scores: On a 9 point scale, Boulet et al (2002) showed raters’ mean scores varied from 4.5-6.6 ($F=2.3$, $p<0.05$); Wiles et al (2007) showed significant but lower magnitude variation in assessors’ mean scores of less than 1 point ($F=5.6$, $p<0.001$). Norcini et al (1997) showed examiners individual mean scores ranged from 5.5-8.0. Schuh et al (2009) compared
consistency of pass / fail decisions by different examiner groups, and showed a kappa correlation of only 0.32 (95% CIs 0.11-0.53). It is important to note that in these naturalistic studies, some of the observed differences in assessors’ mean scores may have resulted from real variations in trainees’ performances. None the less these results tend to indicate important differences in assessors’ stringency or leniency.

Holmboe and colleagues (2003) investigated the construct validity of Mini-CEX assessments by asking examiners to rate videoed performances that had been constructed to show differing levels of performance. Whilst their results supported the overall construct validity of the assessment – different levels of performance received different mean scores, which fell in the intended range – the study highlighted the degree to which assessors’ ratings of common performances may vary. This study used the American Board of Internal Medicine’s (ABIM) standard 9 point version of the Mini-CEX scale, which is subdivided into “poor” (1-3), “Marginal / Satisfactory” (4-6) and “High Satisfactory / Superior” (7-9). Using this format, assessors’ scores ranged between 1-6 for a performance scripted to be unsatisfactory, from 3-9 for a performance scripted to be satisfactory and 2-9 for a performance scripted to be superior. Moreover, of the 38 participants, 14 rated a performance scripted to be “unsatisfactory” as “marginal / satisfactory”. Consequently there is a consistent finding that assessors’ scoring of performance is variable. Moreover the magnitude of these differences are sufficiently large to substantially alter the educational message that is conveyed to a trainee for a given performance.

Relationship of scores to external variables

Investigation of concurrent or convergent evidence to support the validity of direct observation performance assessments has also been undertaken. The distinction between concurrence and convergence is important when considering these results. Concurrent validity seeks to measure the same construct via a different measure, whereas convergent
validity seeks to show that similar constructs operate similarly (Streiner & Norman 2008, p254 & p262). The initial premise upon which direct-observation performance assessments are based is that assessments of authentic performance in the workplace seek to understand something different from that which is measured in decontextualised tests of skills (such as OSCEs) or tests of reasoning or knowledge (vivas or written examinations). Thus if we were to envisage a perfectly valid measure of authentic clinical performance, we would not expect it to correlate perfectly with assessments of lower-order competence or knowledge. An inherent limitation of studies which investigate both concurrent and convergent evidence of validity is that no “gold standard” measures of either performance or competence currently exist. As a result, if a perfect correlation between two assessments were to occur, it would only illustrate that the both assessments possess the same degree of validity, without necessarily confirming the validity of either assessment in absolute terms.

Durning and colleagues (2002) investigated both the concurrent and convergent validity of Mini-CEX scores by comparison with residents’ monthly placement scores (a global score of performance given by their clinical supervisors) and comparisons with scores on an annual knowledge test examination. They showed moderate to high correlations (0.59-0.81, p<0.003) for the concurrent measures, and slightly lower correlations for the convergent measures (0.47-0.57, p<0.05). Notably Mini-CEX and monthly placement assessments appear to have been conducted by the same person, and thus these scores may not have been independent.

Hatala et al (2006) examined the convergent validity of Mini-CEX assessments with three separate components of the Royal College of Physicians and Surgeons of Canada Internal Medicine (RCPSC IM) exams. These included an oral exam, a bedside assessment and a written exam, and showed statistically significant correlations of 0.73, 0.67 and 0.72
respectively with the mini-CEX. Therefore, whilst convergence occurred it is noticeable that the skill that is arguably most similar to that tested in Mini-CEX assessments (bedside assessment) showed the lowest correlation of the 3 methods.

In similar studies, Boulet et al (2002) investigated the convergence of the Clinical Skills Assessment (CSA) – a decontextualised simulated patient based exam – with Mini-CEX scores. Most correlations were high, between 0.60 and 0.93. Baig et al (2009) examined convergence between OSCE-based assessment of communication skills and various forms of performance assessments of communications skills. All reported convergent correlations were small and non-significant, although some concurrent correlations between different performance assessments were significant and moderate (0.50-0.61). Kogan et al (2003) investigated concurrence between undergraduates’ Mini-CEX scores and their placement ratings and convergence between the same Mini-CEX scores and both national exam scores and patient-case write ups. Correlations were significant; the concurrent correlations were moderate (0.35-0.43), whereas convergent correlations were small (0.17-0.22).

A few studies have adopted a broader focus in their search for validity evidence. Willett and colleagues (2009), as well as examining concurrence with attending physicians’ evaluations, examined the convergence of Mini-CEX scores with a range of patient related endpoints that were determined from examination of patient records. These included frequency of colon and breast cancer screening, lipid screening, pneumonia vaccination, and smoking cessation advice. Correlations were small and statistically insignificant. The authors conclude performance assessment tools “are not adequate surrogates for documented patient outcomes”. The degree to which this subset of patient outcomes is a genuine reflection of overall clinical care is unclear; arguably this study compares health promotion activity with
more general clinical competence. None the less it highlights the manner in which current WPBA practice has deviated from the more comprehensive vision Norcini (2003) proposed.

Rather than examine concurrent or convergent validity, one study (Mitchell et al. 2011) examined the ability of workplace-based assessment scores to predict trainees that were subsequently classified as “in difficulty” (hence examining the assessments’ predictive validity). They showed that “trainees in difficulty” had small differences in their mean scores on Mini-CEX and case based discussion (CbD) assessments compared with those “not in difficulty” (for example mean score of 3.79 (difficulty) vs. 3.91 (no difficulty) on a 9 point scale). These differences were slightly increased for a subset whose “difficulties” were performance related rather than personal. None the less a sensitivity and specificity analysis for this subset gave an area under the receiver-operator-curve of only 0.76, with optimal sensitivity and specificity of between 60-75%. On this basis the authors conclude that WPBAs are not capable of adequately detecting clinical performance problems.

**Internal structure of score data**

As long as performance is judged against a fixed benchmark, scores from trainees’ assessments may reasonably be expected to grow over time, assuming that trainees are indeed increasing in clinical competence. On this basis a number of studies have investigated temporal score growth to support the validity of assessments. Alves de Lima et al showed mean scores of 7.16-7.64 in year 1 of a cardiology residency increasing to 8.00-8.20 by year 4 (Alves de Lima et al. 2007); Norcini showed scores increase from year 1 (6.45) to year 2 (6.98), p<0.05 (Norcini et al. 1995), and increases across the course of 1 year of 6.40 to 6.97 (Norcini et al. 2003). Similar increases were seen by Wiles et al (2007), and Schuh et al (2009) showed a decreased failure rate over successive years. Van Lohuizen et al used multi-level modelling to develop an average growth-curve of students’ performance during a 14
week clerkship. They demonstrated a small amount of average growth (7.6 to 7.8 on a 10 point scale), but more importantly showed that by including growth as a variance component in their reliability estimates, the assessments’ overall reliability was improved. This suggests that desirable score growth may have confounded previous reliability estimates (van Lohuizen et al. 2010).

Only one study has not shown score growth over time: Durning and colleagues showed static scores over 7 months (Durning et al. 2002), although the possibility of type II error exists over this shorter observation. Whilst most of these studies show score changes in the expected direction, the magnitude of the changes are fairly small. Little or no consideration is given in these studies to the degree of score growth that is expected over the relevant time frame and no a priori levels were set. Further, none of the studies used any degree of blinding to the level of the trainees. Consequently these results can’t exclude the potential for assessors’ score choices to have been biased (consciously or unconsciously) by knowledge of the trainees’ year, leading to confounding of these observations. These points notwithstanding this repeated finding tends to offer some support to the assessments’ validity.

In addition to temporal-score growth, a further aspect of internal structure has been investigated – namely the degree to which scores that pertain to different domains of the assessment act independently. Durning and colleagues (2002) showed inter-domain score correlation coefficients from 0.59 – 0.81 (p<0.05); Norcini et al (1995) similarly showed correlation coefficients of 0.62-0.78 (p<0.001), and 0.61-0.78 (p<0.001) (Norcini et al. 2003). Margolis et al (2006) showed even higher inter-relations, with inter-domain correlations ranging from 0.72-0.97 (no p-value quoted). In a different approach, Fernando et al (2008) showed that on a 6 point scale, 21% of assessments had the same score across all domains;
only 1.3% of assessments showed a score range of 3 out of 6 or more. Cook et al (2010) assessed the independence of scores from the 6 score domains of Mini-CEX assessments that had been undertaken on internal medicine residents. Despite retaining all factors that explained at least 5% of score variance, common factor analysis produced a 1-factor solution, with an internal consistency of \( \geq 0.86 \). Moreover they were able to replicate this finding on a second dataset.

The notable contrast to this fairly ubiquitous finding comes from the P-MEX assessment, which is used to rate and give feedback to learners on their “professionalism”. The original Canadian version, administered to undergraduates, produced a four-factor solution in exploratory factor analysis. Items within the factors grouped fairly easily into domains of “doctor-patient relationships”, “reflective skills”, “time management” and “interprofessional relationships” (Cruess et al. 2006). Moreover, the same assessment format returned the same four-factor structure when used in Japan in a postgraduate context (Tsugawa et al. 2009).

As a result – the discussed methodological and theoretical comments notwithstanding – scores from direct observation performance assessments tend to: fairly consistently indicate score growth with time, show an overall (although contextually variable) tendency to convergence and concurrence, and demonstrate a lack of domain independence. Collectively these findings tend to support the notion that direct observation performance assessments provide a fairly valid measure of some general clinical competence trait, but little information about the assessments’ sub-domains. This may be because the domains do not describe separate skills or because assessors cannot separate them. Reliability, whilst distinct from validity, may be considered a way to describe the maximum possible validity of scores (Streiner & Norman 2008, p248). Thus once scores from enough assessment
encounters have been amalgamated to ensure adequate reliability, we may view the scores they produce as having reasonable validity. Unfortunately the validity of a single assessment encounter will be limited by the low reliability of individual encounters, which in turn appears to arise mostly from inter-assessor variability.

**Perceived educational impact of direct observation performance assessments**

In most contexts in which they are used the stated purpose of direct observation performance assessments is formative – the assessments are intended to provide feedback which will assist trainees’ development rather than simply measure competence (PMETB 2009; Norcini & Burch 2007). Some consideration has been given to summative use of these assessments in both undergraduate (De Haes et al. 2005; Kogan et al. 2003), and postgraduate contexts (Nair et al. 2008). In view of their generally formative purpose, a more important issue in constructing a validity argument for the assessments concerns the degree to which they assist in trainees’ development.

No controlled studies have compared whether direct observation performance assessments have a measurable influence on trainees’ educational development; Alves de Lima et al. (2012) have presented work of this nature in a conference short communication but publication is awaited. Their results did not show any improvement in performance for trainees using Mini-CEX assessments compared with those who did not, although their intervention was short. Instead, researchers have generally sought users’ perceptions of the assessments’ educational impact. As the assessment mark sheets often include ratings of both the trainees’ and assessors’ “satisfaction” with the assessment, user satisfaction has frequently been reported and is generally high (Alves de Lima et al. 2007; Kogan et al. 2003; Kogan et al. 2002; Norcini et al. 1995; Norcini et al. 2003; Hauer 2000; Weller et al. 2009). The degree to which these ratings reflect perceived educational benefit, as opposed to other constructs (agreement with the scores / enjoyment of the process) is unclear.
Reporting free text responses from survey evaluation of Mini-CEX assessments, Cohen et al (2009) reported that trainees felt they benefitted from being observed, receiving feedback and gaining learning points. Wilkinson et al (2008) reported that 10% of assessors and 22% of trainees suggested in free text responses that the Mini-CEX usefully initiated feedback and discussion. Dattner et al (2010) used a survey method to compare faculty and paediatrics residents’ perceptions of feedback prior to and following introduction of an assessment they termed the “Structured Clinical Observation” (SCO). Results indicated that the assessments had increased the frequency of feedback, and caused assessors to focus more on communication skills than decision making and knowledge. Hamburger and colleagues (2011) coded free text written feedback on paediatric residents SCO assessments. They determined that specific feedback was given in 85% of encounters.

Alves de Lima et al (2005) adopted a phenomenological approach to understanding the impact of Mini-CEX assessments with cardiology trainees. They described the assessments as constructively contributing to learning and helping to foster intrinsic interest. Malhotra et al (2008) described the assessments being anxiety provoking for trainees, but also a valued source of feedback which they ultimately felt helped them to prepare for subsequent exams.

Holmboe and colleagues (2004b) performed a theoretically orientated analysis of tape-recorded feedback following Mini-CEX assessments. They showed that on 80% of occasions the assessors made one or more recommendations. This, however, only resulted in an action plan in 10% of cases. Importantly, in terms of assessing the potential educational impact of these assessments, this study highlights that feedback may often fail to be delivered in a manner that is likely to translate into educational change. These results are striking as assessors were aware they were being recorded which may have induced
Hawthorne effect (the tendency for performance to improve under observation) (Adair, 1984).

Whilst many of these studies are very positive about feedback or educational benefit from assessments, Daelmans et al (2006) describe differing results. They used semi structured interviews to understand users’ perceptions of a newly implemented workplace based assessment programme. Students described receiving very little feedback in the context of assessments, and that the assessments rarely linked to development. The authors described a complex series of organisational influences beginning with poor information and misunderstanding, limited time within the clinical environment, and unwillingness by students to seek feedback actively in case it was negative or they appeared “stupid”. Concordantly, assessors described infrequently assessing students and an unwillingness to fail students. This study importantly illustrates the manner in which local culture may influence the assessment process, and how (despite receiving scores) the assessments can lack educational benefit.

Sabey and colleagues (2011) initially surveyed and then conducted focus groups with UK primary care trainees on their perceptions of WPBAs conducted whilst they were in secondary care training placements. A majority (72%) reported that trainers delivered useful verbal feedback. Focus groups indicated that the usefulness of assessments was reduced (at times) by timing within placements (too near to the end), unwillingness by assessors to fail, feedback that was bland, or incomplete observation of the clinical task by assessors. As well as highlighting the difference in results that can be found by differing methods, this study shows that even when assessments appear to be generally beneficial, the assessors’ behaviour can have a negative impact.
Weller et al (2009) interviewed specialist and trainee anaesthetists to understand their perceptions of Mini-CEX assessments. They described perceived benefit arising from the assessments as they gave routine clinical sessions an educational focus, and increased the tendency for assessors to give feedback. Overall, they viewed the assessments as an improvement on prior arrangements “based on someone’s recollection at the end of four months”. None the less, quite pronounced concerns were expressed about the role of assessors and their ability to choose scores. Trainees felt that assessors needed “calibration”, describing variations in assessors’ interpretations of descriptors, or a tendency to ignore descriptors. Trainees described perceived uncertainty by assessors as to the meaning of scores, and a lack of certainty about whether to be hard or lenient. Trainees perceived that on occasion the assessor “doesn’t want to offend you”, and that personal interactions could lead to scoring bias. Moreover they perceived that some specialists had personal “hobby horses” that influenced their judgements. They described perceived unwillingness to fail trainees due to the additional work load this might impose, or discomfort at the thought of being unkind in a face to face interaction. These findings show considerable resonance with the “failure to fail” literature which has been described as a problem in assessment more widely (Dudek et al. 2005; Cleland et al. 2008; Rees et al. 2009).

In summary, we see that within the literature which uses psychometric methods there is general (but not universal) support for the assessments’ validity. None the less this literature highlights concerns about the degree of inter-assessor score variability, and the impact of this on the assessments’ reliability. This in turn limits the validity of the assessments’ scores. Across the user-perceptions-orientated literature we see a general trend of perceived educational benefit, and some support for the usefulness of feedback that emerges from the assessments, but concerns arise from users’ (particularly trainees’) perceptions of the accuracy and fairness of assessors’ judgements which tend to reduce trainees perceptions of
the overall value of the assessments. Triangulating between these two literatures we may conclude that whilst there is a fair amount of evidence to support the overall validity of the assessments (at least where they have been implemented well), assessment judgements by assessors are problematic, in that they are very variable between assessors, and may be subject to a number of biases which reduce their educational value.

Attempts to improve assessment judgements

As this literature has evolved, focus has shifted to the role of assessors and attempts have been made to improve their judgements and scoring. In broad terms these interventions have taken two forms: attempts to guide score choices by alterations in the scoring format, and attempts to train assessors so that their judgements become more accurate or consistent.

Changes in scale format

Donato et al (2008) developed and tested a novel rating format for direct-observation performance assessments. In contrast to the traditional Mini-CEX format which has 9 rating options for each domain, the new format (which they termed the “MiniCARD”) has only 4: excellent, good, marginal and poor. The new format had 12 different skill domains (compared to the Mini-CEX’s 7 domains), which were grouped into 4 categories: “history”, “physical exam”, “assessment of findings”, and a composite category of “presentation of plan to patient/counselling/behavioural change”. Each of the 12 domains has content specific prompts and behavioural anchors for each level of performance. Participants were randomised to use either the miniCARD or the traditional Mini-CEX format to rate common videoed performances by 2nd year internal medicine residents. The miniCARD group were more accurate at identifying unsatisfactory performances than the control group (96% vs. 52%, Odds ratio (OR) 25.1 p<0.001), but less accurate at identifying satisfactory
performances (73% vs. 95%, OR 0.15 p<0.001). Overall the miniCARD group classified more cases correctly (85% vs. 73%, OR 2.13, p=0.002) and inter-rater agreement rose from low (Fleiss kappa 0.290 z=4.69, p<0.001) to moderate (0.520, z=15.65, p<0.001) agreement. Importantly these classifications rely on considering the “marginal” category as a fail rather than minimal pass. As this assumption is not stated on the MiniCARD form, some assessors may not have interpreted the category this way. None the less these results indicate that this innovation produced a modest increase in overall accuracy and agreement, but at the cost of inducing some rater stringency.

Cook and Beckman (2009) compared a novel 5 point scale with the traditional 9 point rating scale for Mini-CEX assessments, by asking participants to use both concurrently to rate each of the assessments’ domains. Notably this design means that ratings on both scales were based on the same judgements, and are consequently a reflection of how well judgements are translated into scores. Scores from the 9 point scale were slightly more accurate when used to categorise pre-determined levels of performance (54% vs 44%, p<0.0001), and overall inter-rater reliability was slightly higher for the 9 point scale than the 5 point scale (0.43 vs 0.40). Consequently the novel 5 point scale did not lead to any improvement in rating characteristics. It also serves as a further illustration of the degree of variation between assessors.

Crossley et al (2011) developed novel rating scales for a range of direct-observation performance assessments based on two novel constructs: developing independence and developing clinical sophistication. Conceptually these scales hope to align with the constructs that assessors naturally use whilst judging trainees’ performance, which in turn is derived from ten Cate’s suggested theory of “entrustability” (ten Cate 2005). Notably this theory lacks empirical basis. They compared the new scales with the existing scales in real-
world use, and showed that the component of observed score variance that is attributable
to trainees rose for all assessments. For example: Mini-CEX conventional format 29% to new
format 53%, Case based discussion (CbD) conventional format 23% to new format 47%. They
showed corresponding falls in assessor score variance, for example: Mini-CEX conventional
format 43% to new format 23% and CbD conventional format 37% to new format 26%.
These findings are encouraging; however, an important limitation (which the authors
acknowledge) should be noted. Descriptions of developing sophistication and independence
were presented on the new scoring format alongside pre-determined training level anchors.
As a consequence assessors may have tended to select ratings in part based on knowledge of
the year of training of the person they were assessing. As the study included all years of
training, this may have increased agreement and tended to spread trainees from different
years. Consequently this may have confounded the observed results.

Assessor training

A number of studies have investigated the influence of training on assessors’ judgements,
with the intention of improving either their accuracy or reliability. Crossingham et al (2012)
implemented a novel assessment of non-technical skills in anaesthesia. Initial in-theatre use
showed poor inter-assessor agreement, and poor reliability (mean kappa 0.20, G = 0.50). In
response, the authors used a 45 minute training session including instruction on the use of
the scoring sheet and a benchmarking exercise in which assessors independently scored
videos at different levels of performance. Inter-rater agreement on the training videos was
high (mean kappa 0.70), but then fell when the format was used again in real practice in the
operating theatre (mean kappa 0.14, G=0.42). Correspondingly the proportions of score
variance attributable to trainees actually declined following training (19% to 16%) and the
proportion of variance due to assessors increased (31% to 38%). Potentially these results
could be explained by a large training effect that was very short-lived. Alternatively the
effectiveness of training may not translate from the artificial training environment to real
practice. Notably the assessment format had a much higher inter-rater agreement when used in a selection centre (mean kappa 0.61) which increased slightly following training (mean kappa 0.68). As a result the latter explanation seems more probable.

Whilst the duration of training in Crossingham et al’s study may have been too short to achieve an effect, Graham et al (2010) used an 8 hour training programme to train anaesthetic assessors of non-technical skills. Training included all four of the types of training which have previously been advocated within the assessment centre literature (Woehr & Huffcutt 1994): rater error training, performance dimension training, frame of reference training and behavioural observation training. Reliability of assessors’ scores after training failed to meet the authors’ a priori threshold of 0.7, ranging from 0.11 – 0.62. Unfortunately the study had no control group and no pre-training measure of reliability. Consequently no inferences can be made from this study about the effect of training on reliability.

Two studies have used randomised controlled trial designs to investigate the influence of assessor training. Holmboe et al (2004a) included 40 participants who were randomised to control and intervention groups. They used scripted videoed performance to compare baseline Mini-CEX scores with follow-up 8 months later. Consequently there was a considerable washout period prior to the follow-up measurement. Analysis used an intention-to-treat hierarchical approach that adjusted for baseline rating behaviour and geographical clustering. Training included performance dimension training, frame of reference training and behavioural observation training, and was administered to the intervention group shortly after the baseline measurement. Comparisons at follow-up showed that the intervention group rated more stringently than the control group, across all levels of performance (for example: intervention 2.7 vs. control 3.8 for unsatisfactory videos,
intervention 6.0 vs. control 7.5 for superior videos, p<0.001). The range of scores given by the intervention group was smaller for 6 out of the 9 videos compared to the control group (for example intervention 1-4 vs. control 2-7 for unsatisfactory, or intervention 5-8 vs. control 3-9 for superior). Thus this study provided some evidence of a training effect although this may have been at the overall cost of increased assessor stringency.

Cook et al (2009) performed a similar study, except that their baseline measurement and follow-up measurements were only separated by a month, and their training intervention was shorter. Cook et al used reliability as their primary endpoint, and showed that differences in initial and follow-up intra-class correlations (ICC) were not significant between the two groups (intervention group: baseline 0.4, follow up 0.43, control group baseline 0.40, follow-up 0.53, p=0.19). Similarly the accuracy rating did not show a significant difference between groups in the proportion of cases classified correctly (intervention 54.1%, vs. control group 55.2%, p=0.92). The authors concluded that rater training does not improve reliability or accuracy of assessors’ scores.

**Summary of issues**

In summary we see that the reliability of direct-observation performance assessments varies across context, but is fairly universally subject to a large influence of inter-rater variability which limits the assessments’ validity. With a few exceptions, other score parameters offer reasonable evidence in support of the assessments’ validity. Trainees mostly perceive educational benefit from the assessments, but raise concerns about assessors’ score choices and the accuracy and fairness of their score decisions. Attempts to improve score characteristics through either changes in format or training of assessors have generally had either limited or no effect. Crossley et al’s (2011) study is an exception to this, but methodological criticisms of this study have been discussed. Thus the situation remains that assessors’ score choices within direct-observation performance assessments are
problematic, and attempts to improve them have not achieved the desired degree of benefit. This has lead to a desire within the medical education research community to understand the basis of assessors’ judgements and score choices, in the hope that understanding can suggest novel approaches to improving score judgements.

Performance Appraisal and Impression Formation

Section introduction

The judgement processes used by assessors within medical education have so far been subject to very little investigation – what studies there have been will be considered later. None the less, two other areas of literature are useful as they give extensive insight into the way that humans make judgements on other people. One of these is the “performance appraisal” literature; the other is the “impression formation literature”. Neither process is directly analogous to direct-observation performance assessment. The first concerns the study of judgements by managers on the performance of their staff but, unlike direct-observation performance assessments, performance appraisal judgements are based on impressions gathered over a long period of time: weeks to months or even years. By contrast the latter (impression formation) concerns more immediate, but non-technical judgements. As such it concerns issues such as likeability, trust and other interpersonal judgements, rather than professional performance. None the less, by triangulating between these literatures many relevant principles emerge. The purpose of this section of the literature review will be to consider these principles. Rather than systematically describing these large literatures, this review will illustrate important principles.

In both of these literatures the cognitive processes used by individuals making judgements have been a central investigative theme. The study of performance appraisal judgements has a long history – one which already dated back fifty years when Landy and Farr reviewed the
topic thirty years ago (Landy & Farr 1980). Early study was predominantly psychometric, and described a frustratingly consistent range of biases (Denisi 1984; Landy & Farr 1980). These included: the Halo effect, which is the tendency to unduly allow impressions in one area to influence impressions of other areas; leniency or stringency effects, which is the consistent tendency by different raters to score either unduly harshly or benevolently; and central tendency, which is the tendency by assessors to avoid the extreme ends of any rating scale.

Attempts to improve on these issues focused on both alterations to the scale format, and attempts to train raters to improve their judgements (Denisi 1984). Landy and Farr (1980) concluded that changes to the rating format had produced few gains, other than understanding that rating scales should have between 5 and 9 response options. This finding is probably theoretically explained by the finding that human working memory can only actively deal with seven pieces of information concurrently (Baddeley 1994). Conversely some effect of rater training was demonstrated, albeit inconsistently between studies, but Halo effect in particular seemed resistant to this intervention (Landy & Farr 1980). In response to these findings, a number of authors reconceptualised the problem from a psychometric issue to a cognitive issue. Thus the parallels with the literature on direct-observation performance assessments are striking.

Judgements as cognitive categorisations

Within both the performance appraisal literature (Landy & Farr 1980; Denisi 1984; Feldman 1981) and the impression formation literature (Fiske 1993; Macrae & Bodenhausen 2000) “cognitive categorisation” is the central principle around which cognitive theory is organised. Categories may be defined in that “a category exists whenever two or more distinguishable objects or events are treated equivalently” (Mervis & Rosch 1981). In this respect categorisation seeks to group objects together so they can be thought about similarly. As a
result categorisation emphasises commonality between objects or events and therefore reduces or overlooks differences. Humans live in a world that is rich in stimuli (Macrae & Bodenhausen 2000). If we consider for a moment a direct-observation performance assessment, every tiny feature of the trainee’s appearance, speech, behaviour, the interaction between the patient and the trainee, the environment and the clinical context all contribute to myriad range of available stimuli. Conversely humans have a limited processing capacity (Baddeley 1994). Categorisation is a tool to organise and make this array of stimuli manageable (Macrae & Bodenhausen 2001).

Once categories are used to organise information, two implications become inevitable. For any given object or event (let us call this an instance), some of its features will be very typical of the features in the category, and some will be less typical. On this basis Mervis and Rosch (1981) asserted that categories have “fuzzy boundaries”, and that there are gradients of category representativeness, with some instances typifying the category closely whilst other instances are less typical, and perhaps therefore harder to categorise. Thus any given instance will have category-consistent features and category inconsistent features. The former are served well by the category: the category is useful shorthand for these features of the instance. The latter (category inconsistent features) represent more of a difficulty, as the category serves them poorly. Many of the implications of categorical thinking stem from the combination of these two issues.

Categories are not simply passive labels. Their purpose is to help the individual to order the perceived world, and from there to make judgments, form impressions and ultimately to guide actions (Fiske 1993). Consequently, imprecision in their use (as we will see) translates into imprecision in the judgements which they produce. This presents a paradox: the more that instances can be categorised, the more the world is ordered, and the cognitive work of
interpreting it is reduced. Equally, the broader the categories used, the less precisely they describe instances and therefore the more category inconsistent features each instance will possess. Humans possess two systems for interpreting stimulus information. The first (located in the neocortex) stores stable representations of the world which can only change slowly. The second, within the hippocampus, forms temporary representations of surprising or new stimuli; these are lost unless repeatedly encountered (Mcclelland et al. 1995). Human judgement therefore occurs on a continuum which moves between easy (but imprecise) categorical thinking, through subcategories to difficult (and potentially overwhelming) individuated thinking (in which instances are individually represented). The position on this continuum which is adopted depends on the degree to which the perceiver is concerned with accuracy in their impression (Fiske & Neuberg 1990). Thus categories are used to perceive the world in a “good enough” manner (Fiske 1993).

Organisation of categorical thinking

In forming an impression of another person, a perceiver might organise their impressions in a number of ways. Firstly, they might assign a number of category labels to that person. A person could be perceived in terms of their gender, race, age, and numerous other categories such as their facial appearance, clothing, height or weight; for example a tall, well dressed, caucasian, male, who has a cruel face. Stangor et al (1992) showed that perceivers primarily categorise other people based on their sex, or using a composite sex-race category (i.e. black female). Following this primary categorisation they pay attention to subordinate categories such as clothing type or colour only if it seems useful to the purpose they are trying to attain. Similarly age serves as a primary category (Brewer et al. 1981), with associated sub-ordinate categories (grandmother, elder statesman).

Perhaps more striking is that despite the numerous types of category that might be applied, all categories seem to group into one of two traits – namely: warmth and competence. A
number of studies have shown that when ratings from an array of categories are factor analysed, results consistently show two underlying, almost orthogonal dimensions which correspond to these two traits (Fiske et al. 2006; Abele et al. 2008). Moreover, if the traits are dichotomised then the four combinations that result (high or low warmth and high or low competence) produce predictable responses: for example when an individual displays low competence and low warmth, disgust is produced. Conversely, when a person is perceived as low in warmth, but high in competence, envy typically results (Cuddy et al. 2008; Fiske et al. 2006). Consequently perceivers categorise individuals on primary and then subordinate categories, driven by the purpose they seek. These categories are then synthesised through their positions on the two underlying fundamental traits to reach an overall impression.

At the opposite end of the continuum from categorisation (as already described) lies individuation (Fiske & Neuberg 1990). Under these circumstances, perceivers form a rich description of the individual rather than viewing them in categorical terms. Park et al (1994) investigated this phenomenon. They showed that perceivers who have read descriptions of the same behaviour interpret them in idiosyncratically different ways. This particularly occurs if perceivers are exposed to multiple examples of different behaviours by the same individual in different contexts. Under these circumstances, perceivers form abstract inferences about the person’s character that go beyond the observed behaviour to more general representations. Surprisingly, despite the potentially vast array of idiosyncratic impressions that different perceivers might be expected to produce, Park et al were able to group participants’ impressions of any given target individuals into only two or three different broad types of impression.
Mohr et al (2006) repeated this finding using a more ecologically valid approach based on video stimulus material, and found that three different stories existed for each individual, although (again) the stories differed by individual. Additionally the stories varied in that of the three, one always had a markedly opposite valence to the other two (for example very positive about the person, whereas others were very negative). This “person model” of impression formation exists somewhat in contrast to the more dominant theories of categorisation, and its implications have been incompletely examined. Particularly, it is unclear why only three broad explanations would be generated by different people, or if people group based on the type of description they are likely to generate. None the less it offers a different approach to understanding the organisation of impressions.

**Implications of categorical judgement**

Having established that judgements on individuals rely on cognitive categorisation, we can now consider the various ways in which categorisation processes might explain the frustrating discrepancies in judgements which occur. To do this we will consider the way that categories are activated, and then consider a number of biases that can result from category use.

**Category activation**

If humans understand the world around them through the use of stable categories, then logically (in order to interpret the complexity of our environments) we must either possess or develop vast numbers of categories. The key to using these categories lies in the choice of which ones are activated. The classical conception of category activation supposed that category activation was an inevitable consequence of encountering a cueing stimulus (Fiske 1993). Problems, however, exist with this idea. Firstly, perceivers are able to control category activation to some degree (Blair & Banaji 1996). Secondly, when encountering real people (as has already been described), a vast array of potential categorisations exist.
Through these a smaller number must be selected to enable useful cognition (Macrae & Bodenhausen 2000).

In an elegant demonstration, Macrae et al. (1995) showed that viewing images of a Chinese woman eating noodles with chopsticks activated ideas of “Chineseness” in participants, but inhibited ideas of womanhood compared to a control group. Conversely, images of a Chinese woman applying makeup in front of a mirror inhibited ideas of “Chineseness” and activated ideas of womanhood relative to the control group. Based on these findings they proposed that contextual cues enable the perceiver to determine the salience of potential categories, and activate the most salient, whilst inhibiting less salient categories. Other features also seem to contribute to category activation: the perceivers’ pre-existing beliefs or prejudices (Lepore & Brown 1997), and the extent to which the perceiver is processing other information (Spencer et al. 1998). Consequently a number of contextual factors influence category activation, and therefore the way that an object or person is perceived.

Biases arising from categorical thinking

Category-consistency and inconsistency effects

When making categorical judgements, biases arise in a number of ways. As discussed, individual instances contain features that are both category consistent and inconsistent. Once a category has been activated, subsequent recollection of the instance tends to favour category consistent features, rather than category inconsistent features. Fyock & Stangor (1994) conducted a meta-analysis of 26 studies investigating this phenomenon. Results indicated an overall tendency to recollect category-consistent information more strongly than category-inconsistent information. This seems to arise because strong associations form between the activated category and the features of the instance that are category-consistent. These then dominate the category-inconsistent features, biasing recollection of
the event. As well as biasing recall, this can lead to distortions of memory. Tsujimoto and colleagues (1978) showed that once a category schema (or an organisational representation, based on the category) has been developed, participants confidently reported having previously seen category-consistent instances which they had in fact not encountered. Conversely they confidently reported that they had not encountered category-inconsistent instances which had in fact been present. Thus categorical thinking can both induce spurious recollection of events which did not occur or failure to recall real events.

In contrast to these findings, a number of studies have demonstrated an opposite pattern indicating that category-inconsistent information is preferentially recalled in free recall tasks (Hastie & Kumar 1979). This is supposed to occur because category-inconsistent instances require more mental work, and therefore achieve greater salience, making them more readily recalled. Stangor and McMillan (1992) meta-analysed 54 studies that examined either recall (ability to suggest items freely) or recognition (determining whether an instance has previously been encountered) of category-consistent or inconsistent instances. They showed a weak overall predominance of category-consistent effects, but more importantly showed that recall strongly favoured inconsistent instances, whereas recollection strongly favoured consistent instances. They inferred that guessing due to uncertainty relies on category-consistent information, but description of specific examples favours category-inconsistent information. As a result, rather than cancelling each other out, both types of instance can bias the impression formation process in different ways.

**Illusory-correlations**

A further type of bias that results from categorical thinking is termed “illusory correlations”. These occur when there is a perceived relationship between two variables, when they are in fact unrelated (Chapman, 1967). For example, a perceiver may observe that a person has both a deep voice and is good at their job. The perceiver subsequently unduly (and
unconsciously) develops a positive view of the performance of subsequent people they encounter with deep voices (Denisi 1984). Thus the impression which develops is distorted by irrelevant information. In a similar fashion, the illusory correlation of different characteristics helps to explain the development of both stereotypes, and the halo effect – where attributes in one domain unduly influence perceptions in another (Fiedler 2000).

**Contextual influences on category salience**

As described earlier, the salience of different categories within a given context influences their activation. Context can, however, unduly manipulate category salience in a manner that leads to further distortions of impression formation. For example the salience of a woman’s “femaleness” is increased when she is the only woman in a group, compared to being surrounded by other women (Tversky 1977; Feldman 1981) Consequently, activation of the category “female” will unduly influence perceptions in the ways already described, compared to the same woman behaving the same way in a context that does not make her femaleness as salient. This is especially true if the perceivers possess pre-existing gender stereotypes (Lepore & Brown 1997). Similarly, when a perceiver meets an individual from an unfamiliar group (for example a previously unencountered nationality) their unfamiliarity makes this category unduly salient, and tends to produce illusory correlations which the perceiver will generalise to other group members on subsequent meetings (Risen et al. 2007). Consequently, perceivers tend to stereotype members of unfamiliar groups more than those from familiar groups (Risen et al. 2007).

**Conditions that promote category-overuse**

As described, categorical thinking serves to make order in a complex world, and thus serves as an energy saving strategy. Thus categorisation may be used optimally to categorise that which may be easily categorised, freeing resources to individuate and interpret remaining stimuli (Macrae & Bodenhausen 2000). Correspondingly, overuse of categorical thinking
results in bias. A number of conditions have been shown to promote over-reliance on categorisation, resulting in bias in impression formation.

**Cognitive load**

Processing information under conditions of increased cognitive load or reduced cognitive capacity promotes categorical overuse. For example, people who prefer mornings are more likely to exhibit stereotypical thinking (an example of over-categorisation) when making decisions at night compared with making decisions in the morning (Bodenhausen 1990). Similarly, when reading information under time pressure, perceivers recalled more stereotypical information and made more stereotypical judgements than perceivers who read without time pressure (Dijksterhuis & Van Knippenberg 1995). Making decisions under time pressure or whilst exposed to environmental noise also shifted decision mode (Kruglanski & Webster 1991).

**Emotional state**

The emotional state of the perceiver influences reliance on categorical judgment. People who are angry rely more on stereotypical information than people who are sad whilst making judgements (Bodenhausen et al. 1994). Notably sad people performed similarly to people with neutral emotions indicating that anger produces an unusual over-reliance on stereotypes. Similarly, in comparison to controls, envy utilises cognitive resources, allocating attention and memory to targets of the person’s envy and reducing focus on other cognitive tasks (Hill et al. 2011). Consequently envy is likely to increase over-reliance on categorical thinking. Additionally, both anxiety (Wilder 1989), and low self esteem (on the part of the perceiver) (Pittman & Agostino 1989) favour simplistic impression formation. In formulating a model of emotion in impression formation, Forgas (1995) described current emotion (or mood) resulting in the unconscious selective recall of instances from memory that were associated with similar emotions. As such the process is similar to illusory correlation, except
that it relies on an irrelevant feature in the perceiver rather than in the target person. Thus the current emotional state of the perceiver can bias impression formation in both these ways.

**Motivation**

The motivation that the perceiver has in forming an impression of another individual importantly influences the extent that they categorise or individuate. Fiske (1993) divides perceivers’ motivations into a desire for accuracy, and a desire to avoid indecision. For example, Neuberg and Fiske (1987) showed that participants who believed that they needed an accurate impression of a person with schizophrenia, in order to cooperate with them to achieve a financial reward were more likely to ignore stereotypical features, and instead to individuate, than controls for whom an accurate impression was less important. Conversely, if people are motivated to avoid indecision, they tend to rely more on categorical thought. One example of this (as described) is time pressure. A similar scenario occurs if the perceiver needs to publicly report their impression immediately after it is made (Bond & Omar 1990); in this instance, perceivers tend to stop attending to new information and instead attend to formulating their response, thus promoting over-categorisation.

**Pre-existing beliefs**

Finally various long-standing perceiver attributes influence the degree to which the perceiver utilises categorical thinking. One of these is their dispositional tendency to think deeply – or their “need for cognition”. Crawford and Skowronski (1998) showed that people who were high on the “need for cognition” scale – surprisingly – remembered more stereotypical information about target individuals than people low on the scale. Notably though, judgements by the high need-for-cognition individuals were less influenced by these stereotypes than participants with low need-for-cognition. This implies that people with high need for cognition do indeed activate categorical thinking, but ameliorate its influence
before judging. Similarly people’s long-term beliefs (or prejudices) can influence both their category activation, but also the degree to which those categories subsequently influence judgements (Kunda & Spencer 2003). These authors further showed that individuals’ motivation to avoid stereotyping mediates the extent to which activated stereotypes are applied in judgements.

**Attribution processes**

Another important set of processes runs alongside the tendency to either categorise or individuate, namely attribution processes. Let us suppose that under a certain set of circumstances (for example time pressure or prejudicial beliefs), a perceiver notices a feature of the person about whom they are forming an impression that contradicts their categorical representation (or their stereotypical belief). At this point they must either muster both the effort and the willingness to individuate the individual, or they must somehow dismiss the observation. Attribution represents a cognitively easy mechanism by which they perform the latter (Bodenhausen 2005).

In simple terms attribution involves generating causal explanations for encountered events – especially those that contradict expectations. The classical conception of this is that when a negative event befalls another person, humans typically attribute responsibility for the event to that person, but (conversely) when a negative event befalls themselves, they blame external circumstances (Bradley 1978). Inter-personal attribution is more complex, in that it seeks explanations of motivation and control. For example if a person falls down because they are drunk (internal control) or because they are blind (external control) this provokes differing responses in onlookers in terms of emotion and desire to help (Rudolph et al. 2004). Similar processes have been shown to operate in making inter-personal judgements. For example, success by women in business is attributed to strength in the economy (external control), rather then personal effectiveness (internal control) as a way of
preserving gender stereotypes (Deaux 1984). Thus, whilst forming impressions, perceivers can find ways to dismiss, or otherwise attribute observations which don’t fit their evolving model. This may enable the perceiver to perpetuate inaccurate stereotypical impression, or to reject disconfirmatory evidence that would otherwise require alteration of their judgement. As such faulty attribution can lead to the introduction of further bias in the impression formation process.

Judgement as a social process

The literature described so far has considered the cognitive processes used by individuals as they make judgements on other individuals. Within the performance appraisal domain another vein of research exists which goes beyond individuals’ cognitive processing during judgements, to consider the interpersonal context in which those judgements occur (Levy & Williams 2004). As interpersonal processes also have the potential to unduly influence assessors’ judgements, it is useful to consider principles which emerge from this literature.

Appraisal purpose

Whilst the cognitive implications of a perceiver’s goals have already been discussed, other research has looked at the social implications of differing purposes of the appraisal process. A long history of research has considered whether the appraisal results are intended for promotion or salary purposes (high stakes), or whether they are to be used for developmental processes (low stakes). Jawahar and Williams (1997) reviewed and then conducted a meta-analysis on this topic. They showed that high stakes appraisals were scored more leniently than low stakes appraisals by one third of a standard deviation. Moreover this effect was stronger in data collected in the field compared with data collected in laboratory studies. Extending this line of enquiry, Wang et al (2010) investigated whether a range of qualitatively different goals influence appraisal scores. They showed that appraisers who were experimentally manipulated to seek harmony, fairness, or to motivate
their staff all gave comparatively lenient ratings for poor performances, compared to when they were manipulated to seek accuracy in their ratings. Conversely, scores for high performers’ were comparatively reduced in the harmony, fairness and motivation conditions compared with accuracy. Notably, though, the influence on high performers’ scores was less than on low performers’ scores, indicating that these biases had a greater influence on lower levels of performance.

**Appraisal accountability**

Similar to the appraisal purpose, other research has examined the accountability and reporting procedures of appraisal. Klimoski and Ink (1990) examined the influence of subordinates’ self assessment (good, bad, or none) along with mode of feedback (face to face, written, none). Both conditions showed a significant main effect, although the effect of self-assessment was more pronounced. The two conditions showed no interaction. Consequently both interpersonal communication and unwillingness to contradict a subordinate’s self-impression seem to independently influence appraisers score choices. Mero et al (2007) showed that when appraisers are accountable to the people they are rating they give more lenient ratings than when they are accountable to senior administrators. Being unaccountable for ratings similarly induced leniency. This study also replicated similar face-to-face leniency effects. Thus a combination of accountability and interpersonal factors, presumably due to a desire to avoid social awkwardness, further influences ratings.

**Appraiser characteristics**

Similar to effects on cognition, some fixed characteristics of appraisers can be seen to influence scoring patterns. Based on the five-factor model of personality (Costa & McCrae 1992), individuals who are high on “conscientiousness” give more negative scores than individuals low on this trait, and individuals high on “agreeableness” give more lenient
scores (Bernardin et al. 2000). Moreover these authors showed that when individuals exhibit a combination of low conscientiousness and high agreeableness they give particularly inflated ratings. Villanova et al. (1993) developed a self-rated scale of “appraisal discomfort” – that is, a scale which measured appraisers’ level of discomfort at performing appraisals. Psychometric evaluation of the scale supported its validity, and their results show that individuals higher in appraisal discomfort rated more leniently than individuals with lower discomfort ratings. The authors contend that this helps to ameliorate discomfort that these appraisers experience. Notably neither of these studies’ designs involved experimental manipulations, and thus causality of the observed relationships cannot be assumed. This notwithstanding, these studies suggest that appraisers’ dispositions influence their scoring.

**Relationship between appraiser and subordinate**

The nature of the relationship between appraisers and the subordinates that they appraise has been studied. The degree of positive affective regard (or “liking”) of the subordinate by the appraising supervisor has been examined. In simple terms, liking of a subordinate is associated with higher ratings of their performance by supervisors (Lefkowitz 2000; Wayne & Ferris 1990; Varma et al. 1996), although further investigation reveals that the relationship is complex. Notably, the direction of the relationship can be questioned, with some data suggesting supervisors’ liking of subordinates may increase due to high quality of their work rather than the inverse (Varma et al. 1996).

Further aspects of inter-personal relationship between the supervisor and the subordinate appear to influence both liking and performance appraisal ratings. Impression management involves subordinates making efforts to manipulate their own portrayal. These efforts can be supervisor focused (agreeing / flattery / doing favours) or self focused (smiling / politeness). In either case, their intention is to favourably influence the attribution processes that the supervisor makes. Wayne and Liden (1995) studied these phenomena along with
demographic similarity between supervisor and subordinate. They showed that supervisor-focused impression management techniques early in a relationship, and demographic similarity both contributed to liking, but also to the supervisor’s sense of similarity with the subordinate. Both of these then influenced performance appraisal ratings, although notably perceived demographic similarity had a slightly greater influence than liking on ratings.

A further influence, termed organisational citizenship behaviour (OCB), is similar to impression management. This describes activities that go “above and beyond” what a role formally requires – for example willingness to take on additional work, or being cooperative. Notably, OCB does not mean that the employee performs their role any better, and so should still not directly influence performance ratings. Allen and Rush (1998) showed that a combination of OCB and task performance better predict appraisal ratings than task performance alone. Moreover, the influence of OCB seems to occur through increased liking.

Closely related to these concepts is the notion of “leader-member exchange” (LMX). This states that supervisors allocate their subordinates to “in-groups” and “out-groups”. These allocations are made partly on the grounds of perceived competence, but also include phenomena such as dependability and interpersonal compatibility. In-group relationships are characterised by trust whereas out-group members remain distant (Duarte et al. 1994). Duarte and colleagues (1994) showed that high LMX relationships were protective of poor performance whereas low LMX relationships were more accurately rated. Kacmar et al (2003) showed that the influence of high or low LMX on performance appraisal ratings are much greater if the supervisor and subordinate communicate frequently, compared to pairs that communicate infrequently – possibly because increased communication highlights the nature of these relationships.
Employee reactions

Whilst these interpersonal influences are an interesting and important source of bias in appraisal ratings, another topic is relevant to consider: that of trainees’ “reactions” to performance appraisal and the manner in which this influences their subsequent development. As well as enabling decisions on pay and promotion, one of the major aims of performance appraisal is to stimulate development in employees (Fletcher 2001). In that respect it is very similar to direct-observation performance assessments. Feedback is intended to facilitate this goal. Unfortunately previous review of the literature has shown an inconclusive relationship between feedback and development, with up to a third of studies reporting a negative relationship (Kluger & Denisi 1996).

To investigate why feedback appeared to sometimes have a negative impact, Jawahar (2010) defined “reactions” in terms of the perceived accuracy of the feedback, its perceived utility and employees’ satisfaction with it. Jawahar used structural equation modelling to control for the influence of employees’ prior performance, and then demonstrated that his reactions model significantly explained future performance by employees. Moreover the study was able to attribute most of the variation in employees’ reactions to specified characteristics of the feedback. This area of research represents an interesting novel arena, which parallels concerns in medical education which have already been discussed.

Summary of issues from performance appraisal and impression formation

In this section we have illustrated how, within wider fields of enquiry, judgement processes of other people have been construed as a socially-situated cognitive categorisation task (Govaerts et al. 2007). Cognitive-categorisation processes give rise to a number of possible biases: over-emphasis of both category consistent and inconsistent features (depending on the task), distortions of memory, illusory correlations, susceptibility to contextual features
which make particular categories unduly salient, and a range of conditions which promote
category over-use thus tending to make judgements stereotypical. These phenomena are
then compounded when faulty attribution processes enable the perceiver to dismiss
inconsistent features.

The cognitive processes occur within the context of a range of social features which further
influence the judgements that appraisers reach – or at least the judgements that they
report. These social processes tend to revolve around preservation of relationships or
avoidance of social discomfort. They include influences on judgements due to the purpose of
the appraisal and the appraiser’s level of accountability, various dispositional features of the
appraiser and the nature of the relationship between the appraiser and the person they
appraise. Notably, the way in which the employee reacts to feedback then has an important
role in determining the developmental value of the feedback.

Thus it is clear that human impression formation is vulnerable to many forms of subjective
influence. Whilst much of this theory is likely to apply to the formation of judgments within
medical education, the precise operation of these processes, the peculiar vulnerabilities of
assessors and the ways that these interact with perceptions of clinical competence have not
yet been discussed. Although this area was almost entirely devoid of empirical study only a
few years ago, a number of studies now exist, and their contribution to our evolving
understanding of this field will now be considered.

Rater cognition in Medical Education

Within medical education, comparatively little investigation has sought to understand the
judgement processes that assessors use as they reach judgements on observed
performance. At the commencement of this course of study publications within this area of medical education comprised one empirical study of examiners thinking during an oral examination (Yaphe & Street 2003), and a few papers that give the topic theoretical consideration (Williams et al. 2003; Govaerts et al. 2007; Downing 2005). Since then a small number of further empirical studies have been published.

Yaphe et al (2003) used an investigative technique known as interpersonal process recall to study the decision making by examiners for the oral component of the Royal College of General Practitioners membership exam. In this study, they investigated an assessment format that is quite different to performance assessment in that the assessment was for high stakes purposes and didn’t involve observation of clinical performance. They used video recordings of the exam. These were shown to examiners who were asked to describe their recollected thoughts, and explain their reasoning.

In their model, examiners formed a rapid initial impression, which was followed by exploratory then confirmatory questions which progressively refined the impression. Examiners perceived that strong candidates were fluent, self aware, reflective, focused, comprehensive and capable of answering with little prompting. Weaker candidates, by contrast, were described as faltering, inflexible and incoherent, not capable of developing an argument, showing only superficial understanding, deviating from the question or answering only from personal experience rather than literature. The study doesn’t describe the degree to which examiners uniformly agreed with these descriptions, or the extent to which these characteristics were seen as universally applicable to either strong or weak candidates. None the less, the ability by examiners’ in this study to describe typical features of either good or poor candidates is consistent with the use of categorical thinking.
More recently a few studies have specifically investigated assessors’ judgement processes in performance assessments. Kogan et al (2010) investigated relationships between a number of assessor-characteristics and assessors’ scores on Mini-CEX assessments. Assessors provided details of their age, gender, academic rank, years of precepting experience, the frequency with which they assessed trainees and clinical assessment training. Assessors’ clinical skills were then assessed using an 8 station objective structure clinical examination (OSCE) before they observed and scored four standard videoed performances of residents’ performance using the Mini-CEX assessment format.

Demographic data, and assessors’ OSCE scores (predictor variables) were then correlated with the scores they gave to residents in the Mini-CEX videos (dependent variables). Forty-two (apparently independent) correlations were reported between assessors’ OSCE scores and the scores they gave residents. Nine of these were statistically significant – thus exceeding the number expected by chance. They uniformly showed a negative relationship between the assessors’ clinical skills and the scores they gave. Conversely demographic data showed no relationship. Notably (as this study’s design was observational) these associations cannot be assumed to be causal. None the less, these findings suggest that assessors may use their own clinical skills as a frame of reference – in essence that their category systems are based on their own ability or performance. This correspondingly produces differences in scores.

Following on from this study, Kogan et al (2011) used a grounded theory approach to explore factors that impact how assessors judge and rate residents when observing their clinical skills. They asked 38 assessors to observe and judge four scripted videoed performances and two live (but also scripted) cases by residents – again using the Mini-CEX assessment format. Live residents were also scripted to give particular responses to
feedback. Assessors were interviewed following each videoed performance and provided feedback, then underwent further interview after the live cases. Transcripts were analysed using principles of grounded theory.

The researchers developed a model in which assessors viewed observed performance by residents through dual lenses of “inferences about trainees”, and either “internal” or “external” frames of reference. The “external” frames of reference lens varied between existing frameworks, other residents at similar stages of training, more experienced physicians or patient related outcomes. The lens of “Inferences” involved assessors making observations of trainees and then adding their own interpretations to these. These interpretations then appeared to influence assessors’ subsequent judgements. Assessors described considerable uncertainty in how to translate judgements into numerical scores, with varied approaches between different assessors. Assessors were further influenced by external factors – the context of the assessment and the residents’ response to feedback – which influenced their score choices. As with their previous study, these findings suggest that assessors use different categorical systems whilst judging trainees. Moreover the study shows the influence of interpersonal factors with the decision process. Consequently its findings are consistent with the literature already described.

Govaerts et al (2011) investigated the role of assessment-expertise on performance assessments, by comparing the cognitive processes of expert and non-expert assessors. They developed a number of specific hypotheses from the expertise literature concerning differences between experts and non-experts in the time to problem representation and differences in response to contextual cues, numbers of inferences and numbers of self monitoring statements. They then asked both expert and non-expert assessors to watch and scores videos of a simple and a complex case. They recorded the time required for assessors
to develop a first impression, and obtained verbal protocols at a number of points in the process. Their analysis segmented the verbal protocols, and coded statements based on the type of cognition they displayed. They then compared frequencies of different types of cognition between experts and non-experts for both simple and complex cases.

Govaerts et al’s results partly confirmed their hypotheses: experts took longer to produce problem representations than non-experts for the complex case but not for the simple case. Experts also made more inferences and produced fewer literal descriptions than non-experts, and were more sensitive to contextual cues early in the performance. Other differences in cognition were not statistically significant. Notably expert and non-expert assessors produced similar scores. The authors interpret these data to mean that expert assessors possess more elaborate assessment schemata (or, in essence, more developed categorical systems) than non-experts, and that these enable them to integrate information into more meaningful patterns, rather than viewing data as multiple discrete chunks. The authors suggested, on this basis, that experts analyse information more in view of its context. Notably, as scores between experts and non-experts did not vary, these useful insights are not helpful in explaining score variations.

In a recent follow-up study, Govaerts et al (2012) extended their exploration of this phenomenon. Using a similar methodology to their 2011 study, they showed that assessors use their own “performance theories” whilst judging – that is their own personal theories about how to effectively undertake the job. Secondly they showed that different behaviours are expected in relation to particular tasks, but, importantly different assessors have different specific task-related expectations. Lastly they showed that assessors begin to develop “person schemas” – inferences about the trainees’ dispositional nature – very soon after beginning to observe them. Importantly their work (similar to the study presented in
chapter 3 of this thesis, although published later) showed that assessors’ task related expectations were highly idiosyncratic: the combination of factors in which assessors were interested varied between both assessors and cases making each assessors task-related expectations comparatively unique.

Tweed & Ingham (2010) examined the relationship between assessors’ self rated confidence in their assessment scores and their accuracy, whilst marking videoed performances of consultations between undergraduates and simulated patients. Researchers calculated a “correct” checklist score for the performances by means of a borderline regression technique. They used this to determine the “accuracy” of assessors’ scores and then related these to assessors’ confidence.

Overall there was no significant correlation between accuracy and confidence. Around the threshold of acceptable performance accurate and inaccurate assessors had similar confidence whereas at levels of performance further from the threshold accurate assessors tended to be more confident then inaccurate assessors. Further data showed that assessors’ confidence exceeded their accuracy around the threshold of acceptability (overconfidence), whereas their accuracy exceeded their confidence at extremes of performance (underconfidence). These findings indicate that judgements near to the threshold of acceptable performance are more difficult for assessors than performances which are obviously either good or bad. Moreover, assessors’ overconfidence indicates that they may be unaware of the extent of this difficulty.

Two other studies, although not directly involved in investigating assessment judgments, are useful to consider as they illustrate a further relevant principle. Ginsburg et al (2010) interviewed experienced attending physicians about their perceptions of specific recollected
outstanding, average and problematic residents. These impressions were based on prolonged interactions with residents over a period of time and consequently differed from the briefer judgements formed during performance assessments. Participants’ descriptions of residents utilised a range of common themes - knowledge, professionalism, patient interactions, team interactions, systems, disposition, trust, and impact on staff – but combined these in idiosyncratically different ways depending on the resident they were describing. Some features were emphasised and others discounted based on judgements as to which aspects were most pertinent for that resident. Consequently real-life impressions of residents were holistic and subjective, and not a linear sum of components. Notably then, assessors’ use of categories may not be stable but instead they may combine categories in different ways under different circumstances. Considering the prior literature, we might conjecture that this occurs due to contextual stimuli (in the resident or the environment) that cause different categories to be activated.

Similarly Thammasitboon et al (2008) examined differences in assessors’ conceptions of resident competence. Participants were presented with randomly ordered statements describing resident behaviours and asked to rank them in terms of the order that described an “ideal” resident. Factor analysis showed that participants’ responses clustered into four separable clusters. Consequently, different assessors do not appear to possess uniform perceptions of competence, but rather construct ideals in different ways. As such different assessors seem to have constructed their categorical representations of competence differently from each other.

**Summary of issues in rater cognition within medical education**

Collectively these studies give us a number of insights into assessors’ judgement processes. Yaphe et al’s study suggests that examiners may possess stereotypical notions of good or poor performers – thus suggesting a categorical approach to judgment. Ginsburgh et al and
Thammistboon et al show that conceptions of competence seem to vary between assessors, and may be employed in idiosyncratically different ways to different trainees. Govaerts et al’s most recent study developed this picture of individual uniqueness (or idiosyncrasy) employed in the judgement approach further. Thus categorical systems may be complex, individual and not uniformly activated. Govaerts et al’s (2011) study also suggests that categorical thinking is employed, and that it becomes richer and more complex with time. Particularly the increased use of interference by experts suggests that categorical thinking serves a purpose beyond simple classification in the judgement process.

The two studies discussed by Kogan et al extend this understanding to show that assessors’ conceptions of competence seem to rely in part on their own self-reference, but also varying sources of external references. Moreover it illustrates that social influences are perceived to influence these processes. The major contribution of Tweed et al’s study is to illustrate the judgements around the threshold of competence seem to be more difficult for assessors. Moreover, the undue overconfidence which they display may indicate that some of the judgement processes at this stage are unconscious. Consequently a picture starts to emerge of an uncertain, variously constructed cognitive categorisation process which is subject to social influences. None the less, this picture is preliminary and theory offers numerous opportunities for further development.

Formulation of Research Aims:

At the outset of this programme of study very little research had investigated the judgement processes used by assessors during performance assessment in medical education. Several problems with performance assessments’ score characteristics were well described: high inter-rater score variability, low score growth over time and poor discrimination of domains
of competence. Moreover, attempts to improve score characteristics, namely changes in the format of mark sheet and assessor training, had not achieved the desired degree of improvement. Consequently a problem existed without a clear solution. In response, this programme of study has had the following overall aim:

*To develop an understanding of the judgement processes used by assessors as they reach judgments and allocate scores whilst undertaking clinical performance assessments.*

The purpose of developing this understanding was to contribute to theoretical understanding of assessors’ judgement processes in a way that might eventually explain why attempts to improve the scores have been unsuccessful, and to suggest alternative approaches. Consequently this research has sought to respond to the defined problem by generating new theoretical understanding that will in turn lead to developments in practice.

As described, the intervening period of time has seen the publication of some important initial contributions to what can be seen to be an emerging field of enquiry within medical education. These publications have not invalidated the intention of this programme of study; rather they have confirmed both its importance and current relevance. None the less, we have responded to these developments by focusing the intent of individual studies within the programme of research without needing to alter our overall goal. The aim of the first study focused on exploring the development of differences within assessors’ judgement: a theoretically based research question that is developed in that chapter. The following study builds on a concept that emerged from the first study: namely that of comparative
judgements by assessors, and the third study replicates and extends the second study’s findings, additionally investigating conditions that may mediate the observed effect.
Chapter 2
General Methodology

Introduction

In order to address the research aim set out at the end of the previous chapter, it was necessary to theoretically situate our intended research: initially by determining an appropriate theoretical stance; then a methodology; and from there to determine appropriate methods by which to conduct the research. This chapter will describe these processes, focusing on the reasoning for methodological choices and their theoretical underpinnings and assumptions. It will also describe the development of materials that were used throughout the course of study. The precise methods of individual studies will, however, be considered individually in the relevant chapters.

Theoretical orientation of study

The nature of cognition

This programme of study aimed to explore assessors’ judgement processes during performance assessments. Before making methodological choices, it was necessary to situate this phenomenon ontologically. Judgement processes are a subset of human cognition. Cognition in turn may be defined as:

*The mental process of knowing, including aspects such as awareness, perception, reasoning, and judgment*

(definition: thefreedictionary.com)
As a result it is clear that cognition exists in the mind of the individual who uses it to reason, perceive or judge. At this point it is tempting to adopt a relativist view of the ontological nature of cognition (Denzin & Lincoln 2000 p. 21). A few observations are, however, important to consider. Firstly, cognition can occur without the awareness of the individual. This so called “automatic” or “unconscious” processing is well documented and occurs in all types of human judgement (Bargh & Chartrand 1999). Consequently, cognition does not need to be consciously constructed or perceived in order to occur. Moreover, individuals’ descriptions of their reasons for a particular choice can be shown (at times) to involve spurious, incorrect explanations (Nisbett & Wilson 1977). As a result, individuals’ perceptions of the cognitive processes they use to make a decision or choice can be different from the actual cognitive processes that we can logically infer must have occurred. Consequently, cognition possesses many attributes of an ontologically real phenomenon: it exists regardless of whether it is perceived, it does not need to be consciously constructed by the individual, and its actual nature may be different from the way it is perceived (Crotty 1998 p. 10). It is therefore not surprising to find that cognition is almost universally viewed as ontologically “real” within cognitive science.

**Approaches to the study of cognition**

Having decided to view cognition as ontologically real, we then sought to “uncover” assessors’ cognitions in their unaltered form (rather than to participate in constructing an understanding of assessors’ cognitions) (Crotty 1998 p. 27). As a result we wished to be as objective as possible in our research, aiming to minimise interpretation or subjectivity. Unfortunately no truly objective direct measure of cognition currently exists. Instead researchers must choose between two broad methodological strategies: approaches that are subjective, but capable of giving rich insight; or approaches that gather data more
objectively, but which only relate to indirect indicators of cognition and rely on inferences to draw conclusions.

The former of these approaches relies on “introspection” by research participants, or the process of a participant describing their thoughts. This has the potential to be informative, but is hampered by subjectivity. A range of methods have been described to reduce this subjectivity (which will be explored later), but the issue none the less persists. The latter approach uses either experimental designs to test the influence of a particular variable, or objective measures of “process traces” from which inferences can be drawn. These approaches are less subjective but are much narrower in scope. Therefore introspective methods offer a valuable tool to explore and suggest research hypotheses, whereas experimental methods or objective process-trace methods are useful for empirically testing either exploratory findings or those that can be suggested theoretically.

The initial research aim described in chapter 1 was exploratory. Consistent with this initial aim, this programme of research began with an introspective method to conduct exploratory research, and then used the results of this first exploratory study to develop specific hypotheses that were subsequently tested experimentally.

Choice of study population and context

As a first step, it was necessary to define a study population. Our research aim concerned “assessors”: specifically the subset of assessors that engage in clinical performance assessments. We chose to narrow this population to a more focused sub-population in order to provide appropriate methodological control for the comparisons we wished to make.

Consistent with our theoretical stance, we wished to study assessors whilst they engaged in judging performance (rather than relying on retrospective accounts or opinions). In order to
do this, we required the capability to engage assessors in judging a specific clinical performance. Moreover, as we sought to understand differences between assessors, it was helpful to make these the same performances for all assessors, so that we could attribute observed differences in judgements to differences between assessors rather than variations in the judgement task.

Performance assessment judgements must logically rely, at least in part, on judgements about the extent to which the clinical content of the performance is correct. In order to determine this, a clinician must have an opinion about what the clinical content should be. Depending on the scenario this might involve an opinion about what information should be gathered, what diagnosis reached, what information should be conveyed or the sequence of procedural steps that are required. If an assessor is unable to form valid opinions on these issues then it is likely that their ability to judge will be impeded. Correspondingly, Kogan et al’s (2010) study (as previously described) showed that assessors’ clinical skills were inversely related to the scores they gave. Therefore we judged it to be important for assessors participating in our study to be competent at the clinical topics involved in the performances they were asked to judge, or at least for us to be able to make a reasonable presumption of their competence.

As a result we firstly needed to use the same stimulus material for all participants, and secondly all participants had to be competent at the clinical topics that the stimulus material depicted. To achieve this it was necessary to restrict our study population to assessors who were competent in the same clinical discipline.

Consistent with this decision, we chose consultant physicians as our study population. We defined “physicians” as “doctors working in specialties associated with general internal
medicine, who had trained via the Royal College of Physicians exams system. As such we expected that they would all be familiar with common case presentations in internal medicine, either through current practice or through extensive exposure whilst training. The selection of physicians, as opposed to another speciality, was arbitrary yet convenient: methodologically there was little reason to presume any speciality was preferable to another; equally both my clinical supervisor and I are physicians, and hence are very familiar with case material in this speciality. Moreover physicians are a substantial subset of practicing doctors in the UK, and so results would be capable of generalising to a sizeable population.

Next it was important to consider the assessment context in which assessors’ judgments were to be studied. As described in the literature review, a large variety of performance assessments exists (Kogan et al. 2009), which employ many different formats of rating scale and which assess a range of different clinical tasks. Little in the literature suggested whether these different types of assessment would result in differences in the judgment approach used by assessors. One finding, however, seemed important, namely that assessments of procedural skills generally required fewer assessments in order to achieve acceptable reliability than was required for assessments of clinical consultations (Beard et al. 2011; Pelgrim et al. 2011). Moreover, these assessments concern tasks which correspond with different parts of Bloom’s taxonomy (Crossley et al. 2002). It therefore seemed plausible that different judgement processes might be involved in assessing procedural skills, and consultation skills.

Again with the intention of aiding comparisons between participants, we chose to study assessments in a single context. We selected consultation skills, and selected the Mini-CEX format as an exemplar of a direct-observation performance assessment. The Mini-CEX
assessment format was introduced by the American board of Internal Medicine (ABIM) in 1995 to assess the clinical skills of internal medicine residents (Norcini et al. 1995). The format involves a trainee performing a routine clinical task, and being observed and judged by a more senior doctor who provides scores and feedback. Scores are provided for a range of domains on a Likert-type scale and recommendations for development are provided. As described in chapter 1, a number of formats of the Mini-CEX exist; this programme of study used the UK Foundation Programme format of the Mini-CEX (Norcini 2005): see appendix 2. The Foundation Programme is a mandatory stage of training for all newly qualified doctors in the UK comprising postgraduate years 1 (denoted F1) and 2 (denoted F2). The scoring format for Foundation Programme Mini-CEX asks assessors to provide judgements for domains of history taking; physical examination; communication skills; critical judgement; professionalism; organisation/efficiency; and overall clinical care. Judgements are scored on a 6-point Likert scale that has descriptors of ‘well below expectation for F1 completion’, ‘below expectations for F1 completion’, ‘borderline for F1 completion’, ‘meets expectations for F1 completion’, ‘above expectations for F1 completion’ and ‘well above expectations for F1 completion’.

In addition to the considerations already described, it is plausible that assessors’ approaches to making judgements may vary for different levels of learners. Because of this, we chose to standardise the level of learners in the assessments we studied. Correspondingly, we studied consultant physicians’ judgements whilst performing Mini-CEX assessments on performances by foundation year 1 (F1) doctors. Performances by these doctors showed them in consultations with new patients presenting with acute general internal medicine problems.
Having outlined these choices, in the next section of this chapter we will consider two methodological approaches that we used during this programme of study, and our reasons for their selection. Firstly we will consider introspective methods. Next we will consider internet-based research. This is pertinent as both experimental studies contained in this programme of research were administered via the internet. Following consideration of these methodological approaches, the chapter will conclude with a description of the development of the stimulus material used throughout the programme of research.

Consideration of Methodological approaches

Verbal protocols as a tool for studying cognition

The process of an individual attending to and then describing their cognition is termed “introspection”. Within psychology its use as a research technique dates back to the late 19th century. It fell out of favour in the early 20th due to concerns about its verifiability, before regaining popularity in the 1960s. In a prominent critique, Nisbett and Wilson (1977) reviewed studies that investigated peoples’ abilities to describe their thought processes whilst completing tasks. They summarised these findings by saying that

“Subjects are sometimes (a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response.”

The authors suggested that when attempting to describe their thoughts, people generate explanations based on pre-existing beliefs and explanations, and then concluded that reports may be accurate at times, but only when influential stimuli are prominent in the individual’s thinking.
In contrast to this view, Ericsson and Simon (1980) described a model that suggested a series of conditions under which verbal protocols, or “Think Aloud (TA)”, would be a valid representation of cognition. They suggested that in some tasks (referred to as level 1), cognition during the task occurs in a verbal form and requires no recoding. For example this might involve a task such as re-ordering words within sentences. Verbalising these thoughts was not believed to alter them. Other tasks (level 2) would require recoding from other forms to allow them to be spoken. For example this might involve grouping pictures of animals or performing calculations out loud. As these examples would have required recoding of visual images or numerical concepts into language, they were expected to slow cognition but not to alter it. A third level of verbalisation was viewed as requiring additional cognition (for example asking participants to explain their reasons for certain choices in addition to performing the task itself). This third level of verbalisation was expected to interfere with normal task completion. Importantly, describing steps in a decision or judgement process can be seen as a level 2 process as long as participants describe their thoughts, but do not explain them.

Ericsson and Simon recommended a series of practical measures to ensure that only level 1 or 2 verbalisations are collected. Principally these involved: ensuring that participants are actively engaged in the task in question (rather than giving general opinions from their remembered experiences); reducing the time between participants’ thoughts and their verbalisations, either by verbalising during the task (concurrent verbalisation) or immediately following task completion (retrospective verbalisation); and by asking participants to describe their thoughts whilst specifically instructing them not to offer any additional explanation. They contended that by following these procedures, the resulting verbal protocol would be an accurate representation of the actual sequence of cognitions used to complete the task.
The degree to which this model achieves its aims has been controversial for the subsequent 30 years, still resulting in ongoing debate (Ericsson & Fox 2011; Schooler 2011). As well as the potential to slow down task completion – which Ericsson and Simon (1980) acknowledged in their model – two other potential flaws may occur. Protocol reactivity occurs if verbalisation alters the outcome of the task – by creating either a categorically or quantitatively different outcome. By contrast, a second flaw known as protocol “nonveridicality” occurs if the protocol is not an accurate representation of the thought processes used in the task. Nonveridicality can result from omission (where some thoughts remain unreported), or commission (where thoughts that did not occur are reported) (Russo et al. 1989).

A number of studies have examined the degree to which these flaws occur. Before proceeding to examine these, one further point is important to note: verbal protocols can only ever collect conscious thought, and therefore will only gather deliberative, conscious reasoning. Thus even if procedures can be established that are completely veridical (i.e. the verbal protocol accurately represents the conscious thoughts used in the task) and non-reactive, they will still not report any of the unconscious cognitions that may have been used in task completion (Wilson 1994).

Russo et al (1989) examined the reactivity of protocols collected under a range of conditions that either met or violated Ericsson and Simon’s conditions in comparison to silent controls. They showed that for one task (mental addition of 3 digit numbers – designed to involve a high mental work load), accuracy was substantially reduced by both concurrent think-aloud (TA) (14% worse than silent control, p<0.05) and 3 variations of retrospective TA (8 to 23% worse than silent control, p<0.05). Two other tasks (one a verbal task based on anagrams,
the other a pictorial task) showed no change with think aloud. A fourth task, a numerical task involving choice between 2 gambles, was expected to perform similarly to the anagram task. In fact, both concurrent and 2 out of 3 variations of retrospective TA improved performance on this task (16% to 20% improvement, p<0.05). Response times also varied, but not in a related pattern. Importantly, the changes did not follow Ericsson and Simon’s (1980) 3 stage model. Russo et al concluded that whilst Think Aloud may be accurate under some circumstances, its reactivity is unpredictable.

More recently (Fox et al. 2011) performed a meta-analysis of 94 studies that compared participants’ silent performance at a task with participants performance at tasks whilst performing either: 1/. concurrent think-aloud performed according to Ericsson and Simon’s (1980) specifications (E&S TA), or 2/. Think-aloud under a range of different conditions: “explanatory”, “directed”, or “unspecified”. They showed that difference in task performance between E&S TA and silent-control approximated zero (r = -0.03; 95% CI [-0.10, 0.03], Z = -0.94, p=0.35). Conversely “explanatory” think-aloud (in which participants were asked to either describe or explain their thoughts during task completion) did show reactivity (r = 0.23; 95% CI [0.14, 0.31] Z=5.05, p=0.001). The study also compared visually-based tasks with verbal tasks and showed results to be the same for both types of task. Consequently this study provides strong evidence to show that task performance is not affected when think-aloud is performed under Ericsson and Simon’s (1980) conditions, but that it is affected when explanations are concurrently sought.

Studies that have attempted to examine the veridicality of TA (the degree to which spoken thoughts are an accurate reflection of the thoughts actually used in completing the task) have typically compared TA transcripts with process measures. Depending on the task being studied, process measures can be participants’ response times to subsets of the task, eye
tracking measures of gaze-fixation, or measures of the order and timing of data acquisition, such as mouse clicks or the order in which different data-sources were accessed. In each case inferences are made about participants’ cognition from the sequence or pattern of these measures. Importantly therefore, studies that attempt to validate the veridicality of Think-Aloud pre-suppose that the process measures themselves are veridical markers of cognition.

Brinkman (1993) compared the content of both concurrent TA and retrospective TA protocols with process measures that consisted of observations of the order in which data were revealed and tests made. From this, researchers inferred the participants’ cognitive strategy. Results showed moderate agreement between TA and process traces (Cohen’s Kappa ranged from 0.53-0.61 between experimental conditions). Guan et al (2006) demonstrated that subjects’ eyes fixated on the same items they subsequently described in retrospective TA protocols between 83-92% of occasions whilst performing a range of visually orientated tasks. Finer-grain analysis of eye movements reduced the overlap with verbal reports to between 47-59%. In both of these studies, results indicate the same finding: some similarity was found between TA protocols and process measures, but this similarity was incomplete. In a further study, Klasen et al (2008) compared think-aloud transcripts with functional magnetic resonance imaging (fMRI) data of cerebral activation. They described increased activation in areas of participants’ brains known to process functions similar to those reported in participants’ verbal protocols. From this they inferred support for TA as a method.

The difference between concurrent and retrospective verbalisations is worthy of some consideration. Ericsson and Simon’s (1980) original model presumed that concurrent verbalisations involve articulation of thought that it is currently active in working memory.
As such no retrieval from either short or long term memory is required, but the process may exert some additional mental workload. Conversely retrospective verbalisation involves retrieval of information from short term memory. Russo et al’s (1989) study (as mentioned) showed that performance on more complex tasks is impaired, even when verbalisation is retrospective. This presumably occurs because anticipation of verbalisation exerts an additional mental workload.

A few studies have examined whether either the content or accuracy of protocols vary between concurrent and retrospective verbalisations. (Kuusela & Paul 2009) found that concurrent verbalisation generated many more statements than did retrospective (mean of 66 vs 11, t(56)=8.27, p < 0.001). Particularly this involved greater description of features of the task, and also less reporting of difficulties during the task. Retrospective verbalisation did, however, show an increase in the number of statements regarding the final choice in the task (2.65 vs. 1.38, t(56) = -3.24, p < 0.002). Taylor and Dionne (2000) similarly showed that compared to retrospective verbalisation, concurrent verbalisation described many more reports of actions, but many fewer reports of strategies or beliefs. They conclude that a mixed strategy using both concurrent and retrospective verbalisations offers the fullest approach. Somewhat in contrast to these findings, van den Haak et al (2003) showed that concurrent verbalisation identified substantially fewer problems compared to the retrospective condition in a computer usability task (0.5 vs. 4.5, t(38) =5.168, p<0.001) which the authors accounted for by participants lacking time to articulate their thoughts completely during task completion. Neither condition reported all the problems that were seen to have been identified by observational process traces.

In an associated commentary on Fox and Ericsson’s (2011) meta-analysis, Schooler (2011) pointed out that whilst Fox and Ericsson’s study demonstrated that appropriately
administered concurrent think-aloud may not be reactive, it remains unproven that it captures cognition completely. Schooler highlights concerns with the ability of think-aloud to capture brief transitory thoughts which occur as interim steps between more substantive reportable thoughts. He also expresses concerns over the verbal ability of participants to articulate all cognition, as (he suggests) many cognitions do not correspond to appropriate language.

The position described by Schooler (2011) aligns with that adopted by Wilson (1994) who concluded that think-aloud methodology offered an excellent source of hypothesis generation, as it is able to “tap the content of consciousness” but was not adequate for hypothesis testing. In view of the literature reviewed, and the described limitations of the technique, we adopted the same position on the role of Think-Aloud: we viewed it as useful for exploration and hypothesis generation but inadequate for hypothesis testing.

Internet-based research approaches

The internet has been used by researchers to gather data since the mid 1990s when it became technically possible for website users to enter data into websites (Reips 2002b). Since then a number of authors have considered the implications of internet research, and a number of thorough reviews are available (Skitka & Sargis 2006; Barchard & Williams 2008; Birnbaum 2004). The purpose of this section is to illustrate the major principles that emerge from this literature.

Perhaps the most obvious benefit of internet based research is its ability to reach a broad sample of participants (Reips 2002b); traditional psychology lab experiments require that either participants attend the researchers’ lab, or that a researcher visits the participants. This produces constraints due to travel time and cost, and requires appointment scheduling, thus reducing flexibility. By contrast, internet-based research may be conducted at any time
that suits participants, regardless of researcher availability, and in very geographically dispersed locations. For example, McGraw et al. (2000) found that participation can occur throughout the 24-hour period of the day, and Dandurand et al.’s (2008) Canada-based study recruited participants from Europe and Asia as well as Canada and the United States.

This ease of reaching participants is of more than simply pragmatic benefit. Studies conducted via the internet have (at least theoretically) the potential to recruit a more representative sample of the study’s intended population than is achieved in lab-based studies. Prior to the internet, much of experimental psychology was conducted on psychology undergraduates due to simple recruitment necessity. As psychology undergraduates may be different from the general population in numerous ways, the generalisability of these studies is impeded (Skitka & Sargis 2006). Even where researchers try to include more representative samples they are often constrained to participants from a limited geographical area or they use an opportunistic sample, such as conference delegates (Azar 2011). Internet based research therefore has the potential to reach more representative samples of study populations; whether in practice it actually does is likely to depend on the way that recruitment is conducted and whether all subsets of the study population use the internet evenly. If such considerations are ignored internet research may simply reach a different but equally non-representative subset of the study population. Consequently, thought should be given to these issues during study design.

Unfortunately, whilst internet use is highly prevalent through most parts of society, there is evidence that internet-users are not entirely representative of wider society. Research has shown that 24% of Americans do not use the internet at all. Compared with internet users, non-users comprise a higher prevalence of older people, those who are less wealthy, those who live in rural locations and those from African-American or Hispanic backgrounds.
(Lenhart et al. 2003). Thus researchers must consider the implications of these population demographics to their research population. As already stated, this research used consultant physicians as its research population. The use of computers is very prevalent within healthcare, and although not empirically demonstrated, it seems unlikely that there would be many, if any, consultant physicians who do not have access to the internet, or who are not at least moderately comfortable with its use.

Another striking aspect of internet-based research is its anonymity (Skitka & Sargis 2006; Barchard & Williams 2008) – researchers don’t meet participants during the course of the study, and (depending on procedures) may never be aware of their identity. Principally this removes the opportunity for researchers to (consciously or unconsciously) influence participants’ responses through inter-personal interaction. Notably, anonymity may alter the way that participants respond. Chang and Krosnick (2010) used a laboratory experiment to compare participants’ responses to telephone interviews with those given to a self administered on-line survey. They showed that compared with telephone interviews, the internet-based approach reduced social desirability effects (answering questions in a manner that is expected to be favourably-perceived) and satisficing (giving responses that are satisfactory but not optimal).

Perhaps the most fundamental question regarding the use of the internet to conduct experiments is whether results of studies remain consistent with those expected from a laboratory situation: in essence whether internet-based research is a valid proxy for laboratory experimentation. McGraw et al (2000) replicated a series of classic psychological studies via the internet rather than in a laboratory setting. Consequently they did not perform a controlled comparison, but instead relied on comparison with accepted patterns of data. The studies included within- and between- participant comparisons, and involved
timed responses requiring accuracy in the order of 10 milliseconds. In each of the reported studies on-line results demonstrated the expected patterns, leading the authors to conclude that the internet is a feasible method for psychological research. Importantly their software ensured that all data required by the study was downloaded prior to the study commencing. Consequently the study did not need to “stream” data over the internet, and so was not sensitive to the speed of the internet connection (or its “bandwidth”). Without this feature, it seems unlikely that such accuracy would have been possible.

In a different approach Dandurand et al (2008) directly compared participants’ performance on the same problem-solving experiment conducted in both on-line and laboratory conditions. Results showed that participants were significantly less accurate when completing the study on-line compared with laboratory conditions (F(1,114) = 6.10, p=0.02, effect size 0.66), but that this difference in accuracy was distributed evenly across groups within the on-line experiment. Consequently the overall results were not altered. The authors suggested that reduced accuracy may have occurred due to uncontrolled distraction or reduced effort by on-line participants. Consistent with Dandurand et al’s results, reviews of on-line research generally conclude that both modalities obtain similar results (Skitka & Sargis 2006; Birnbaum 2004), although the described reduction in accuracy may produce more “noise” in the data, requiring a larger sample to achieve statistical significance. Skitka & Sargis (2006) conclude that the increased sampling potential more than compensates for the observed reduction in accuracy.

Apart from reduced accuracy, Dandurand et al’s (2008) study illustrates a persistent issue with internet-based investigations: high dropout rates from the studies. 600 potential participants visited the study’s information page; 376 (63%) chose not to participate and a further 98 (16%) were unable to proceed due to technical reasons with participants’
computers. These individuals were considered as non-starters rather than drop outs, having effectively declined or failed to enrol. Subsequently 103 participants enrolled, and 63 completed the study. The remainder ceased participation part way through, by choosing either not to navigate to the end of the study, or not to enter all required responses. This gave a dropout rate of 39%. By contrast, just 4% of the lab condition dropped out. Importantly, however, dropout rates were similar across different study groups and so do not appear to have confounded results.

In other studies, dropout rates of up to 89% have been reported (Reips 2002a), and are widely recognised as a threat to internet methodologies (Reips 2002b; Skitka & Sargis 2006; Birnbaum 2004). This phenomenon is suggested to arise from the comparative ease of leaving internet studies compared with laboratory studies. Although participants in lab-based studies are “free” to leave at any time, the combination of the time a participant has already invested in attending, and the social desire to avoid disappointing the researcher probably both serve as barriers to leaving (Birnbaum 2004). As long as dropout has no systematic effect across groups, it is not anticipated to be problematic, but it if one study arm requires more effort by participants then less motivated participants may preferentially dropout of the more effort-requiring study arm – confounding results. Reips (2002b) recommends using dropout rates as a measure of “motivational confounding” to ensure this does not occur.

A few further potential limitations of internet research should also be noted. The absence of a researcher means that no-one is readily available to give further instructions or clarifications if participants have questions whilst undertaking the research. Barchard et al (2008) recommends that at very least researchers should provide contact information so that participants may discuss information further before deciding whether to participate.
Moreover, this lack of researcher contact means that participants have less opportunity to highlight any part of the study’s stimulus material or responses that are counter-intuitive or unclear. Consequently it is recommended that materials are pre-tested (Reips 2002b) to ensure that they are easily understood and operated by potential participants thus ensuring that both participants’ on-line consent and responses are meaningful.

Technical issues may occur that prevent the study from working on a participants’ computer; as described earlier 98 potential participants were lost in Dandurand et al’s study (nearly equal to the total number enrolled) because of technical issues. As long as there is no reason why this would differentially influence different experimental groups there is little reason to suspect that this will produce a systematic bias in results. Consequently it is important to ensure that different study groups don’t require higher processing power or better internet connections. Perhaps more relevant is the influence on sampling; if technical issues differentially affect a subset of the population (i.e. rural inhabitants, poorer demographics with older computers) then the generalisability of results may be reduced.

It is possible that a single participant might attempt to complete a study on multiple occasions (multiple submissions). This in turn may bias the data, by over-representing a particular view or feature. Survey methods are more vulnerable to this phenomenon than experimental designs, where multiple submissions are likely to be randomised to different groups avoiding any systemic effect. Preventing a motivated individual from providing multiple submissions is very difficult, as they may invent identities, use new email addresses, or different computers to supply new submissions (Birnbaum 2004). More simply, it may be adequate simply to request that participants only take part once.
Masquerading by participants can be problematic, if participants who are not part of the study’s population present false identities. This has been noted to occur in research focusing on gender identity (Mathy et al. 2002); similar to the issues with multiple submissions it is perhaps most important for researchers to consider whether individuals will have any motivation to masquerade within a given study. In conducting our current research programme, we judged that it was very unlikely that consultant physicians would possess any motivation to complete multiple submissions.

Finally the ethics of on-line research have received some consideration (Barchard & Williams 2008). Gaining on-line consent is recommended, as is providing debrief information to participants where experiments may have been upsetting or involved deception. The importance of adequate data security to ensure confidentiality is stressed.

In view of these considerations, we concluded that internet-based experimental designs were a useful approach to addressing our research aims. In particular, consultant physicians have fairly low prevalence in the population, and are distributed across wide geographical areas. Moreover, owing to their working patterns they often have limited time in which to participate in research. Consequently we judged that internet based research offered a method capable of recruiting an adequate sample from this population whilst retaining an acceptable ability to control and manipulate variables in order to answer research questions.

Development of stimulus materials for study

In order to present common stimuli to participants, videos were produced of performances by foundation doctors as they performed consultations. We opted to script these performances rather than film naturalistic performances for two reasons. Firstly, we wished to sample assessors’ judgements across different levels of performance. Consequently it was
important to ensure that stimulus material showed clearly distinct levels of performance. Scripting offered a more certain way to achieve this aim. Secondly, as some videos would necessarily depict poor performance, we had concerns that showing real, unscripted videos of poor performance to a large audience of senior doctors could detrimentally affect the careers of the depicted doctors. Scripting avoided this concern. None the less we recognised that scripting performances reduced their ecological validity (the degree to which research conditions approximate real-life) compared with videoing real performances. To compensate, we used a thorough script development process and tried to ensure a highly realistic feel to the videos whilst filming.

**Development of performance scripts**

Although our research aim was to explain differences between assessors in their judgements, the purpose of assessments more generally is to distinguish between different levels of performance. Consequently we wished to compare assessors as they in turn judged performances at different levels of proficiency.

To achieve this, we decided to depict performances at three different levels: “good” (performance that was competent with some features of excellence), “borderline” (performance around the threshold of competence) and “poor” (performance that was less than competent). Neither literature search nor examination of the foundation programme curriculum (FP 2007) revealed adequately detailed descriptions of performance at these levels. Instead (as we intended videos to show foundation doctors’ performances in patient consultations) we used published literature to develop a list of features of good consultations (Kurtz et al. 2003; Martin 2003), and combined this with specific principles from the expertise and diagnostic reasoning literatures. The first principle was that experts possess well organised, subject specific knowledge structures (McLaughlin et al. 2007). The second principle was that knowledge acquisition by novices is brief and incomplete; by
intermediates is extended but poorly focused; and by experts is concise but focused (Boshuizen & Schmidt 1992). We used these to write a list of criteria characteristic of “good” and “poor” performances, and then, by referring to these lists, wrote qualitative descriptions of performance at each of the three levels (available in appendix 3).

Next we used these descriptors to write scripts of performance in specific clinical cases. In order to ensure plentiful stimulus material for later study, scripts were written depicting each level of performance (good, borderline and poor) for each of three different clinical cases. All cases were common presentations to acute general (internal) medicine, ensuring both that they were likely topics of assessment for foundation doctors and that assessors would be highly familiar with such presentations. The three clinical cases were:

1. Pleuritic chest pain in a 54 year old woman
2. Unexplained loss of consciousness in a 34 year old man
3. Upper gastrointestinal bleeding in a 44 year old man

To ensure authenticity of the depicted clinical cases, case descriptions were written based on the experience of clinicians in the research team, with reference to relevant clinical guidelines. Resources used included:

UptoDate Evidence Based resource (UptoDate 2010) for

- Pulmonary embolus
- Wells scoring system
- Transient loss of Consciousness
- Epilepsy
- Peptic Ulcer Disease
• Presentations of Upper GI bleeding

As well as:

• Draft NICE guidance on transient loss of consciousness (NICE 2010)
• Scottish Intercollegiate Guideline Network (SIGN) Guideline on Acute Upper and Lower GI Bleeding (SIGN 2008)

Cases were written to be typical in content, but to contain sufficient variation in symptoms to pose some diagnostic uncertainty.

Based on our performance descriptions, nine scripts were written – one depicting each level of performance on each clinical case. Scripts included areas of omission judged to be consistent with the level of described performance and showed variations in ability to filter clinical information and weigh diagnostic possibilities. Scripts also varied in the communication skills, patient explanations and interpersonal behaviour they depicted. The level of depicted performance was scripted to be roughly consistent across different facets of each performance (diagnostic reasoning / communication skills etc) and to remain fairly constant throughout each performance. Examples of the scripted performances are available to view in appendix 4.

Supporting the scripts’ validity

Scripts were independently reviewed by a panel of three reviewers. Reviewers were all consultant physicians who worked in acute medical care and who were substantially involved in supervising and appraising foundation doctors. Reviewers were blinded to the intended performance level of scripts and read them in varied orders determined by a Latin
square. Reviewers were instructed to rate scripts sequentially, indicating their perceptions of the scripts’ level and plausibility. Performance was rated as “good”, “borderline” “poor” or “other – specify”. Plausibility was rated descriptively. Scripts were sent to reviewers via email without interactions with the researcher.

Ratings from the review panel’s ratings are shown below in table 1:

Table 1: Review panel ratings of scripts.

<table>
<thead>
<tr>
<th>Script</th>
<th>Construct</th>
<th>1\textsuperscript{st} reviewer</th>
<th>2\textsuperscript{nd} reviewer</th>
<th>3\textsuperscript{rd} reviewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>Good</td>
<td>GOOD (I can’t believe such detail from FY1)</td>
<td>Excellent</td>
<td>Good, nice and concise for an on-call history. Very realistic *</td>
</tr>
<tr>
<td></td>
<td>Borderline</td>
<td>Good (believable) *</td>
<td>Borderline plausible *</td>
<td>Good but missed haemoptysis</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor (Don’t believe such low quality</td>
<td>Borderline plausible *</td>
<td>Poor - I would be surprised to see this sort of history</td>
</tr>
<tr>
<td>Case 2</td>
<td>Good</td>
<td>Good (believable) *</td>
<td>Excellent quite unrealistic</td>
<td>Good but missed incontinence biting tongue. Believable - anything less would be unacceptable *</td>
</tr>
<tr>
<td></td>
<td>Borderline</td>
<td>Borderline (believable *</td>
<td>Very good plausible *</td>
<td>Borderline - sadly this is also believable *</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor (not very plausible</td>
<td>Borderline plausible *</td>
<td>Poor - I would be surprised to see this sort of rubbish history</td>
</tr>
<tr>
<td>Case 3</td>
<td>Good</td>
<td>Good, not very believable</td>
<td>Reviewer did not respond</td>
<td>Good and this is what we see in practice *. However, bit slow and</td>
</tr>
</tbody>
</table>

82
As indicated, the results showed broad agreement with the intended constructs but not complete agreement. Case 2 matched the constructed performance levels on 2 out of the 3 ratings. Reviewer 2’s ratings of case 2 were different to the constructed performance level, but the scripts still rank in the intended order. For case 1, no reviewer’s ratings matched the construct exactly but there was broad similarity, and ratings ranked in the intended order. In the 3rd case, reviewer 1’s rating followed the intended constructs, and reviewer 3’s ratings follow the intended order, but the borderline case was rated better than borderline.

As table 1 indicates, in 15 out of the 24 instances when reviewers gave opinions, scripts were rated as plausible, or a qualitatively similar description (marked with *). 1 response did not rate plausibility. The remaining 8 responses predominantly rated the scripts as implausible.
due to infrequent contact with performance at that level rather than due to authenticity of the content.

Results did not indicate complete consensus. This variation was consistent with the variation described in the literature (for example Holmboe et al. (2003) in which experienced, trained examiners showed considerable variations rating the same performance). The results did, however, follow the intended order, albeit with discrepancies regarding the absolute level of the scripted performances. Moreover in a majority of occasions (16/24) reviewers’ absolute ratings matched the intended construct, and there was broad support for the scripts plausibility. We deemed this sufficient to proceed with the scripts.

Filming the scripts

Selection of participants for filming:

We used real foundation doctors to play the scripted roles in our videos. Owing to the possibility that both gender and ethnicity might influence assessment scores (Woolf et al. 2008; McManus et al. 2008) we limited video-participants to junior doctors of a single gender (female) and ethnicity (white). White females were chosen simply as the largest demographic group in our hospital and were recruited via group email to all foundation year 1 doctors in our hospital.

Patients in the videos were performed by professional “simulated patients” (SP) who were hired via a commercial SP agency (www.medicalroleplayers.co.uk). Each case was played by a different SP, but the same SP featured in all videos based on the same case. By contrast no foundation doctor appeared twice within the same case, although they featured in different cases. Consequently, within each case, a different foundation doctor played each different level of performance, but the SP remained constant. Scripts and project information were
Filming procedures were approved as part of the first study’s ethics application.

**Filming scripts:**

Videos were filmed by the research team. To enhance the videos’ ecological validity, the films’ environment was designed to look clinically authentic (see figure 1) and was kept constant between videos. Foundation doctors’ clothing (blue hospital scrubs) and patients’ attire (hospital gowns) were standardised. Recording was done in high definition, and videos were stored securely in accordance with conditions agreed with the ethics committee.

Figure 1: Still image from video for Case 2, Good Performance
Supporting videos’ validity

To ensure that filming had not altered the performance level depicted in the scripts, the videos were also reviewed. In view of the level of discrepancy in the script reviews, a fourth reviewer was added to the panel.

In a similar procedure to previously, reviewers independently sequentially rated the level of the performances in a varied order (Latin square), whilst blinded to the intended performance level, indicating whether videos were “Good”, “Borderline”, “Poor” or “other – specify”. Although present, the researcher took care not to interact with reviewers during either video observation or rating. Reviewers’ ratings are shown in table 2.

Table 2: Review panel’s ratings of videos

Case 1

<table>
<thead>
<tr>
<th>Constructed level</th>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Reviewer 3</th>
<th>Reviewer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Borderline</td>
<td>Borderline</td>
<td>Poor</td>
<td>Awful = other</td>
<td>Poor</td>
</tr>
<tr>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
</tbody>
</table>

Case 2

<table>
<thead>
<tr>
<th>Constructed level</th>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Reviewer 3</th>
<th>Reviewer 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Good</td>
<td>N/A</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Borderline</td>
<td>Borderline</td>
<td>N/A</td>
<td>Poor</td>
<td>Borderline</td>
</tr>
<tr>
<td>Poor</td>
<td>Poor</td>
<td>Other – fail. Stop her practising medicine</td>
<td>Poor</td>
<td>Poor</td>
</tr>
</tbody>
</table>
Similar to the script review process, the rank order of the ratings corresponds with the intended order in almost every case; reviewer 3’s rating of the borderline video in case 1 as “awful = other” is difficult to interpret in relation to the grade of “poor” given to the poor video. For case 2, two out of the three responding reviewers agreed with the absolute intended level for all 3 videos rather than simply agreeing with the performances’ order. Absolute-level agreement for all 3 videos was indicated by 1 out of 4 reviewers for case 1’s video, and 0 out of 4 for case 3’s video, although for case 3, two reviewers’ ratings (reviewer 1 and reviewer 4) were only marginally different from the absolute intended level. Thus, on the basis of these data, the validity of all videos can be supported.

Conclusion

Whilst these findings support the overall validity of these videos for their purposes in the subsequent studies, they should not be seen as supporting any singular absolute level of performance, or “true score” for the videos. Instead this validation demonstrates that the videos showed clearly distinct levels of performance which were approximately in line with the intended constructs of “good”, “borderline” and “poor”. The videos were used in various combinations in each of the subsequent studies, as described further in later chapters.
Summary

In this chapter we have ontologically situated cognition, recognising its “real” nature and hence the desire to study it objectively. We have considered various approaches to studying cognition, noting that no direct objective method exists. Following on from this theoretical orientation, we have considered two research methods: introspection (or Think-Aloud) and internet-based research. We concluded that introspective methods offer a useful exploratory approach, but not one that is valid for hypothesis testing. We subsequently concluded that internet based research is a valid and feasible research approach which would be useful within this programme of study.

Following on from this we have described the development and validation of video stimulus material to be used in the programme of study. These videos depict foundation doctors engaged in patient consultations in three different clinical cases, each showing performance that is nominally “good”, “borderline” and “poor”. On the basis of the presented validity evidence we judged that these video performances were adequate for their intended purpose within the programme of study.
Chapter 3

Seeing the same thing differently.

Developing study research question:

So far this thesis has provided a review of relevant literature on performance assessments and summarised salient principles relating to human judgements. This resulted in the formulation of the programme of study’s global research aim:

*To develop an understanding of the judgement processes used by assessors as they reach judgments and allocate scores whilst undertaking clinical performance assessments.*

Chapter 2 considered various aspects of methodology, and explained our choices in terms of methodological approach: to begin with an exploratory study, based on a Think-Aloud methodology, and then to test hypotheses that are suggested by the exploratory study in follow-up experiments. This chapter will describe the first of these studies.

In response to these considerations we posed the following research aim for our first study:

*To explore assessors’ cognitive processes whilst undertaking performance assessments, and thereby to illuminate the sources of variability in assessor judgements.*

Before proceeding, a further consideration is pertinent. In chapter 1 we mentioned that prior to commencing this research a few authors had given theoretical consideration to performance assessments within medical education. An important difference in orientation
can be seen between these papers. Two of the papers (Williams et al. 2003; Downing 2005) predominantly focus on potential sources of bias in the rating process: factors that (when they apply) may be expected to have a fairly uniform influence on the rating process. Numerous issues are described, but include the influence of structured data organisation techniques, varying frequencies of observation, the use of structured forms to guide observation and the influence of face to face feedback on judgement. Downing’s (2005) paper focuses on statistical methods to overcome rater variations rather than an explanation of their occurrence.

In contrast to these papers, Govaerts’ et al’s (2007) review proposes that performance assessments should be viewed through a different theoretical lens. The perspective they advance views assessors as active processors of information, that are influenced by a range of pre-existing differences in the cognitive schemas (or categories) which they possess, and who are variably influenced by a number of additional social influences within the judgement situation. Consequently they suggest that assessors’ judgements should be viewed as idiosyncratic; unique or different but equally legitimate. As such they propose that viewing assessors’ judgements as “faulty measurements” of an underlying “true score” is unhelpful and that instead assessors’ variations should be viewed as resulting from multiple equally valid true scores.

These different perspectives owe much to different philosophical approaches to the topic (different epistemologies). What they illustrate is that whilst some influences may be expected to have a fairly uniform influence on most assessors (biases), other differences may arise between assessors that are due to individual quirks or areas of comparative uniqueness in assessors’ understanding or approach (idiosyncrasy). Both may be important. None the less, in the study presented in this chapter we focused our research on
understanding ways in which assessors’ judgements become comparatively unique or individual. We chose this focus as little empirical research has addressed this topic whereas (especially within the wider literature (Williams et al. 2003) biases have been more thoroughly investigated.

The study described in this chapter produced quite extensive findings. To make this paper most helpful to the research community, we focused the published paper reporting this study on findings that were novel. The study also produced a number of findings which were similar to those already reported by other authors. These are not reported in the paper, but, for completeness, are described at the end of this chapter.

Methodological Considerations in Study 1

The methods of Study 1 used three of the videoed performances described in chapter 2 to provide a common stimulus to assessors. This enabled comparison of assessors’ approaches as they judged the same performances. In order to sample across levels of performance, the “good”, “borderline” and “poor” performances were selected from the same case (case 2). As described in chapter 2, we considered “Think-Aloud” to be a valuable approach to explore differences between assessors’ judgement approaches. Participants performed “Think-Aloud” concurrently whilst observing and judging performances, and retrospectively whilst determining scores and an appropriate feedback message. Follow-up interviews were conducted with participants to provide supplementary insight into their perceptions of the judgment process.

Analysis of assessors’ “Think-Aloud” comments and interview responses was undertaken using principles of grounded theory (Strauss & Corbin 1998). Grounded theory is a complete methodological approach to the study of social phenomena. The assumptions which
underpin grounded theory are post-positivistic – they assume that the investigative phenomenon is real, singular and apprehendable, but acknowledge that the difficulties of studying social phenomena ensure that resulting theory remains provisional and incomplete (Crotty 1998 p.29-41). Consequently this methodology aligned well with our chosen stance.

Grounded theory begins without any overt pre-existing theory relating to the study phenomenon, and seeks to develop theory directly from data. Data is collected iteratively and theory is developed inductively – undergoing refinement as each progressive iteration provides new data. Precise details of the analysis are described within the methods section of the paper within this chapter.

Subsequent to developing their “Think-Aloud” method, Ericsson and Simon (1993 ch. 6-7) recommended a formal analysis technique called “Protocol Analysis” (also sometimes referred to as “Propositional Analysis”). This technique is used to segment data into propositions (discrete thoughts which are described in the data) which are labelled according to a coding structure and counted. As a result it produces quantitative results. This approach is consistent with Ericsson and Simon’s view that “Think-aloud”, when collected under their precise conditions gives an accurate record of cognition during a task. In chapter 2 we considered this position, and instead aligned this research with a more conservative view that sees Think-Aloud as a useful “window” into cognition, but not a complete record (Wilson 1994; Schooler 2011).

Consistent with this view of the limitations of Think-Aloud, and our exploratory approach, we judged that it would be more informative to analyse participants’ responses with a technique more suited to exploration and theory development. Yang (2003), following theoretical consideration of different approaches, recommended coding “Think-Aloud” data
in a “context sensitive” manner, and thus supported the idea of deviating from Ericsson and Simon’s strict coding framework. Moreover, other studies investigating assessors’ judgements in medical education have successfully employed grounded theory analysis with either “Think-Aloud” (Kogan et al. 2011) or stimulated-interviewing (Ginsburg et al. 2010). On this basis we chose to collect data via “Think-Aloud”, but analyse it by means of grounded theory analysis.

In addition to the data presented in the paper for study 1, participant information sheets and consent forms used in the study can be viewed in appendix 5, and an example transcript can be viewed in appendix 6.
Study number 1 is described in the paper titled:

**Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments**

Yeates P, O’Neill P, Mann K, Eva K.

Advances in Health Science Education (2012)

DOI 10.1007/s10459-012-9372-1

Published on-line ahead of print, Available at


This paper should be obtained and read as part of the thesis, but is omitted from this on-line publication to prevent copyright infringement.
Additional results from study 1:

As described earlier, some findings from study 1 were very similar to those already reported by other authors. These were not described in the published paper, but are described here, and related to the existing work.

Use of inference:

Similar to the results presented by Kogan et al (2011) and Govaerts et al (2011), assessors in this study made inferences whilst observing performances, that went beyond observable behaviour to unobservable mental processes:

she still hasn’t taken that information and she’s processing it poorly

Assessor 6, borderline, concurrent

that’s good, she understands the difference between sitting and standing.

Assessor 7, good, concurrent

she was thinking about all the questions she needs to ask

Assessor 12, poor, concurrent

Importantly, this tendency varied between assessors. Some made no inferences whilst others made them comparatively frequently. Assessor 7’s comment (above) seems to indicate that this inference contributed to the evolving judgement that assessor 7 made. Similar to inferences, assessors sometimes demonstrated that, given the performance they
had seen so far, they had expectations about how the candidate would subsequently perform. For example:

*See I would have expected somebody like her to sort of say oh it’s a bit too much, you shouldn’t be drinking that much*

Assessor 3, good, concurrent

In this comment, the phrase “somebody like her” appears to indicate that assessor 3 had placed the observed performance (and indeed the person) into a category of similar performances or people. This category then seems to have enabled the assessor to develop an expectation of how the trainee would subsequently perform. When the performance deviated from the expectation, this was noted by the assessor.

Overall, these inferences and expectations were a comparatively small subset of assessors’ comments. None the less, they demonstrate that these processes occurred, but that assessors seem to have differed in their reliance on them. Moreover, they support the theoretical idea that assessors’ judgements utilise categorical schemas.

**Non-performance factors:**

Kogan *et al* (2011) reported a theme called “factors external to resident performance that drive ratings”. This described a range of issues which, although not part of the residents’ performance, none the less appeared to influence raters’ scores. These included case complexity, perceptions of residents’ familiarity with the case, the duration of the rater-to-faculty relationship, perceived response to feedback, aversion to giving negative scores and more general institutional culture.
Correspondingly, assessors in our study described the potential for the judgements that they made to be influenced by a range of factors that were unrelated to the content of the performance, or “non-performance factors”. Assessors commented on a range of emotions whilst observing performances: frustration, boredom, irritation, and even anger, acknowledging that these could affect their concentration on the process, and bias their subsequent judgement:

*if I was watching that as Mini-CEX it would really frustrate me. I would lose interest. And even the little good things which they might have will get lost because I’ve formed my judgment*

Assessor 3

Similarly, another influence due to non-performance factors related to the trainees’ appearance. Some assessors indicated that the trainees’ appearances would influence their judgement during performance assessments:

*if they’re tattooed and pierced then I’ve already taken a slight, um, against them if you like and they would have to perform a little bit better wouldn’t they?*

Assessor 7, follow-up interview

In this comment, this assessor indicated that if they developed a negative opinion of a trainee’s appearance, the trainee’s clinical performance would need to exceed that of other trainees to achieve a comparable grade. Other assessors demonstrated that an awareness of trainees’ appearance was prominent in their thinking, even if they didn’t link this awareness to the scores they gave:
Within a few minutes she was competent, she was making the right judgements, the right questions, big earrings but I was happy with her

Assessor 6, retrospective, good

Consequently, it seems possible that this awareness may have unconsciously influenced the assessor’s judgement. During the follow up interview, assessor 3 explained that they felt it was their duty as an assessor to highlight issues relating to appearance during the assessment process. It is reasonable that assessors might consider a trainee’s appearance or personal presentation as a facet of professionalism. None the less, these comments demonstrate the way these dimensions seem to overlap within the judgement process, possibly causing some halo effect between assessment domains (the halo effect concerns the undue interplay of ratings from different domains (Downing 2005).

A number of other non-performance factors were observed, which related to leniency in assessors’ scores. Some assessors commented that when a performance was poor, it was difficult to provide low scores or negative feedback in a face to face context:

I: would it be difficult to deliver that message face to face to a doctor who had just performed that way?

A: Yes it will be difficult, I don’t want to demoralise whoever it is and saying that you are doing completely nonsense … work.

Assessor 12, follow-up interview

Notably this assessor justified this approach by seeking to preserve the trainee’s self-esteem. Assessors admitted that this emotion-based reasoning could inflate the scores they gave.
Another assessor suggested that foundation doctors were too early in their career to be given strongly negative feedback.

A number of assessors described instances when they believed that prior knowledge or opinions of a trainee influenced their judgement. Typically this occurred when trainees had performed below the level that the assessor believed was typical of their usual performance:

\begin{quote}
A: I think it’s difficult. If you know your trainee well ... you’re willing them to be good – if they’re good
\end{quote}

\begin{quote}
I: And what would that tend to do to the scores?
A: I think I’d be more generous. Assessor 3, follow-up interview
\end{quote}

Assessors justified this approach by explaining that their accumulated experience of a trainee was more representative than the limited sample in a Mini-CEX, or that the pressure of direct observation might influence a trainee’s performance:

\begin{quote}
I know it’s an artificial situation and you’ve got me here and maybe that was putting you off a little bit in how you were doing it
\end{quote}

Assessor 11, Retrospective, poor

\begin{quote}
my regular ward round I ask him questions, that’s where I do my real assessment ... if that chap comes and does a mini-CEX ... I have already based my opinion on them
\end{quote}

Assessor 12, follow-up interview
Two assessors admitted to a tendency to leniency on first encounters with trainees, scoring more stringently later if relevant improvements didn’t occur:

*I was tempted to mark as borderline ... on the assumption that this was a first encounter... I was willing to give them the benefit of the doubt...*

Assessor 10, follow-up interview

By contrast, some assessors described judging comparatively unfamiliar trainees more stringently. A number of assessors were confident that support and training could improve the performance by weaker trainees, but some felt that this justified leniency in scores.

Consequently, these non-performance factors show a number of similarities with the data already reported by Kogan et al (2011). In particular, our findings of perceived effect of prior knowledge of the trainee, the difficulty of delivering difficult feedback in a face to face context and aversion by some assessors to giving negative scores under various circumstances all resonate closely with Kogan et al’s study. The latter two of these issues also resonates with the “failure to fail” literature (for example Dudek et al. 2005; Rees et al. 2009), which has described reasons why assessors choose not to score failing trainees as poorly as they might. The findings of both trainee-appearance and assessors’ emotions interacting with judgements appear to be novel within medical education.

Whilst neither of these themes (inference and non-performance factors) represent particularly novel theoretical insights, they are none the less valuable as they show that these findings can be replicated in the context used in this study.
Conclusion:

The study presented in this chapter sought to explore assessors’ judgement processes as they reach judgments and allocate scores. We have presented three themes which describe novel findings: differential noticing, criterion uncertainty and information integration. Collectively these themes offer considerable insight into the ways that assessors’ judgements become comparatively individual or unique, and therefore how (depending on the stance from which the phenomenon is viewed) they either contributed random error to measurements or they represent multiple true scores from individual perspectives. In addition to these three novel themes, we have also presented data relating to two further themes: use of inferences and non performance factors. These replicate work from other contexts, although both may also contribute to variations between assessors.

Having developed these initial insights based on our exploratory approach, we next sought to develop testable hypotheses in order to refine and develop these initial theoretical ideas. This process is described in the subsequent chapters.
Chapter 4
Mind who you stand next.

Formulating research aim

The findings described in chapter 3 suggested a number of possible avenues for further study. The theme “differential salience” suggested the potential to further investigate assessors’ attentional focus during observations. Assessors’ differing constructions of competence (described in the “criterion uncertainty” theme) suggested the need to formally compare the degree of uniformity in these conceptions. The third theme, “information integration” suggested the need to test the degree of judgemental difference (rather than score difference) which occurs between assessors.

We chose to focus on one issue described in the “criterion uncertainty” theme, namely that assessors frequently appeared to compare a trainee’s performance against other trainees (rather than against a fixed standard) whilst they made judgements. We selected this option for several reasons. Firstly, we were unaware of any investigation within medical education into relative or comparative judgments by assessors. Secondly, it seemed to be theoretically important: it had the potential to inform our understanding of assessors’ conceptions of assessment criteria. Thirdly, it seemed practically important: owing to hospital working patterns, assessors’ recent experiences of trainees are likely to be different from each other. Consequently, assessors will most probably make comparisons against different exemplar trainees. As a result this phenomenon might potentially help to explain assessors’ judgemental variations. Fourthly, it represented an unexplored influence on assessment scores, one which had the potential to be unfair.
Consequently we chose to investigate the influence of comparative judgements on assessors’ scores, in particular investigating whether different recent experiences could produce different scores.

Review of relevant literature

We explored the wider academic literature for helpful theoretical insights. Within cognitive psychology, two distinct phenomena seemed relevant: contrast-relativity effects and assimilative-anchoring effects. Both influences arise due to comparisons between items, and both effects therefore bias the scores that an individual provides. Importantly understanding which effect is dominant in performance assessment judgements within medical education is important to developing a theoretical understanding of the issue. In the section that follows major principles from these literatures will be reviewed.

Before proceeding one terminological clarification is useful. Both phenomena make reference to “anchor” stimuli. These are previously encountered items which influence subsequent judgements. In the case of contrast relativity effects the subsequent judgement is unduly different from the anchor whereas in assimilative anchoring effects the subsequent judgement is unduly similar to the anchor. The item that is influenced by the anchor (the object of concern) is referred to as the “target”.

Comparative influences on judgement

The ability of individuals’ perceptions of objects to be influenced by context is long established. For example, perceptions of the weight of a target object can be altered by changing the weights of objects which precede the target in a series (Brown 1953; Anderson 1975). More generally, perceptions of many target stimuli are influenced by their ‘contextual framing’; the information presented alongside target objects markedly influences the way
the targets themselves are perceived (Tversky & Kahneman 1981; Norbert Schwarz 1999). Taken further, some authors argue that all human judgements of absolute quantities rely on relative comparisons rather than retrieval of stored representations of absolute amounts (Stewart et al. 2005). Although this latter theory remains controversial (Brown et al. 2009), debate only concerns whether “absolute” judgment processes have either some or no influence on judgements (Brown et al. 2009; Azar 2007); there is uniform agreement that human judgements are influenced by relative comparisons. Similar to cognitive categorisation, comparative thinking offers an energy saving approach (Mussweiler & Epstude 2009) which is promoted by high cognitive load or emotion (Saini & Thota 2010).

**Contrast-relativity effects**

A range of explanations exist for the occurrence of contrast effects (Niedrich et al. 2001); the best supported is Parducci’s (1965) “Range-Frequency Theory”. This states that when an observer makes a judgement on a target item, they compare it with a distribution of items. This distribution may be retrieved from memory, but stimuli that have been encountered recently or frequently exert the most influence (Wedell et al. 2005). Notably, this means that individuals only use part of their knowledge whilst judging (combined with contextual cues) rather than the totality of their available knowledge (Brewer & Chapman 2003). The person’s judgement of the target item is based on the item’s distance from the top of the distribution (the range principle), and by its ordinal rank, or the number of items above it in the distribution (which Parducci et al (1965) called the frequency principle). Consequently changes in the item’s context influence its perceived value: adding a high-value anchor will increase the range of the target item; introducing an item higher in the distribution than the target will reduce the target’s rank. Both alterations will serve to reduce the perceived value of the target item.
This principle has been demonstrated across a number of contexts. For example, when asked to judge the width of human silhouettes, participants’ judgements were markedly influenced by the distributions in which they were shown (Wedell et al. 2005). Their judgements showed a contrast relativity effect; viewing body shapes either in the context of narrower images, or further from the top of a distribution, made them appear wider. Similar findings have been demonstrated for a range of concepts. In psychology, individuals’ ratings of happiness (Smith et al. 1989); levels of gratitude in response to help (Wood et al. 2011) and life satisfaction in relation to earning (Boyce et al. 2010) have all shown contrast relativity effects. Within economics, perceptions of value show similar influences (Azar 2011; Niedrich et al. 2009).

A few studies have examined the influence of contrast-relativity effects on performance ratings. Murphy et al (1985) compared undergraduates’ ratings of a lecturer having viewed preceding good or poor examples of lecturer performance. Findings showed a contrast effect with students rating the lecture more favourably following initial poor performance compared with initial good performance. Importantly this study showed that when ratings were obtained after a 24 hour delay, the contrast effect (although reduced) still persisted. Becker and Villanova (1995) showed similar findings examining ratings of managers’ performance. Various conditions mediated the effect: “simultaneous” observation, in which all performances were observed before any of them were rated, showed a smaller effect than sequential observation and rating. Again delayed rating reduced the contrast effect compared with immediate rating.

In a follow up study, Becker and Miller (2002) again replicated the contrast-relativity effect but this time none of the previously-demonstrated mediating factors (simultaneous vs. sequential, immediate vs. delayed, changes in the person being rated (same vs. different))
had any interaction with the contrast effect. Consequently whilst the contrast-relativity effect occurred consistently across these studies, the mediating effect of other variables is less clear.

Importantly, contrast relativity effects can be experimentally induced in experts (for example experienced economists) just as easily as in novices (economics undergraduates) (Azar 2007). This tends to indicate that despite extensive experience immediate context still has an important effect. The strength of these contrast effects differs across different types of judgement: individuals experienced a stronger contrast effect in response to material purchases than in response to experiential ones (Carter & Thomas Gilovich 2010).

**Anchoring assimilation effects**

The anchoring-assimilation effect involves a person being exposed to an initial anchor and subsequently making judgments of a target that are unduly similar to the anchor. Consequently, a high anchor tends to inflate subsequent judgements whereas a low anchor tends to decrease subsequent judgements. The effect is therefore the direct opposite of the contrast-relativity effect. As an example, whilst trying to judge the same quantity (the number of countries in Africa) people who were given the number 10 as an anchor provided smaller estimates than people given the number 60 as anchor (group means were 25 countries and 45 countries respectively (Tversky & Kahneman 1974)).

Two competing theories exist to explain how assimilative-anchoring effects may arise. Tversky and Kahneman (1974) suggested a mechanism they called “insufficient adjustment”. In this an individual entertains an initial idea that the anchor value may be an adequate estimate of the target. They subsequently realise that it isn’t and, in response, adjust their estimate in the direction they believe the target to be. Adjustment, however, tends to be insufficient and so results in estimates which are nearer the original anchor than the true
target-value. Adjustment appears to be insufficient because it ceases when the first plausible estimate is reached (Epley & Gilovich 2001), and as a result the effect occurs less (ie responses are more accurate) when participants are motivated to seek accuracy (Epley & Gilovich 2006).

Strack and Mussweiler (1997) proposed an alternative mechanism based on “selective accessibility”. In their model (Mussweiler & Fritz Strack 1999; Mussweiler 2003), when a person focuses on an anchor they activate mental representations associated with the anchor. These remain active whilst the person considers the target. This causes them to pay attention to features of the target that are similar to the anchor. As a result they tend to notice similarities between the anchor and the target, whilst ignoring less consistent features, producing an anchoring effect. More recently Simmons et al (2010) have proposed that both “insufficient-adjustment” and “selective-accessibility” may operate concurrently.

Anchoring-assimilation can occur despite the provision of an anchor that is overtly irrelevant or non-informative source, such as a spinning “wheel of fortune” (Tversky & Kahneman 1974). The effect also occurs for data with real-world relevance, in which individuals might be expected to have more skill in making estimates: estate agents property estimates (Northcroft & Neale 1987), college students’ estimates of their own grades (Cervone & Peake 1986) and court judges’ sentencing (Englich & Mussweiler 2001) have all shown assimilation. Notably in Cervone and Peakes’s study, the provision of an anchor not only influenced students’ estimates of their ability, but also their behaviour and overall performance. This demonstrates the degree of influence that anchoring-assimilation effects may have. Whilst all of these effects relied on the provision of anchors to participants by researchers, people can also generate their own anchor values from cues in the task (Epley &
Importantly this shows that anchoring-assimilation influences may arise spontaneously in real decision processes.

Of further importance is whether anchors of one type (or modality) can influence targets of a different type (or modality). Carroll et al (2009) showed that anchoring effects could be produced using both semantic anchors “very low, low, medium, high, very high” and letters “A, C, E, G, I, K” in a manner similar to that achieved with numeric anchors. Notably though, when the anchor was provided in semantic form, and the target was presented in numeric form, no anchoring-assimilation effect was observed. In direct contrast, Oppenheimer et al (2008) showed that a general sense of “largeness” or “smallness” could be induced by anchors in one modality, which subsequently influenced judgments on items from a different modality (for example initially copying long or short lines, and later estimating the length of a river). The reason for these different findings is unclear; consequently the conditions that mediate cross-modality anchoring-assimilation effects remain to be elucidated.

Other influences have been shown to mediate the extent of anchoring assimilation effects. Inbar and Gilovich (2011) showed that the anchors have more influence on people who are either sad or fearful, than on people who are either angry or feeling disgust. This effect was more pronounced for self-generated anchors than for researcher-given anchors. Jacowitz and Kahneman (1995) showed an inverse relationship between confidence and anchoring influences: more confident people were influenced less. Asking participants to describe reasons why a given anchor is inappropriate (called “considering the opposite”) also reduced the influence of anchoring manipulations (Mussweiler et al. 2000; Chapman & Johnson 1999).
Contrast relativity effects versus assimilative anchoring effects

Having considered both of these effects we might wonder why in some instances contrast-relativity effects occur, whereas in other instances assimilative-anchoring effects occur. A number of studies have explored this question.

Mussweiler and colleagues showed that when people are manipulated to consider similarities between an anchor and a target, assimilation occurs, whereas if they are manipulated to consider differences then contrast effects occur (Mussweiler 2001b; Mussweiler 2001a). Developing this idea, Mussweiler (2003) proposed that in real-life decisions, an initial appraisal creates an impression of either similarity of difference between an anchor and a target. This may be mediated by salient aspects of category membership, extremity, ease of processing, and the individual’s motivation in forming a judgement. This initial impression subsequently determines whether assimilation or contrast effects prevail.

A range of other factors have influenced whether assimilation or contrast effects occur. Presenting anchor stimuli together with target stimuli tended to produce assimilation, whereas as presenting them separately (sequentially) tended to produce contrast effects (Tanner 2008). Reminding participants of prior anchors caused them to contrast anchors and targets; without a reminder their judgements showed assimilation (Strack & Schwarz 1993). Presenting anchor phrases in a “fluent”, easy to read format produced assimilation whereas disrupting the word order (making them “non-fluent”) produce contrast effects (Greifeneder & Bless 2010). Collectively these findings support Mussweiler’s (2003) assertion that assimilation may be the default method of comparison, because humans are attuned to search for similarity (rather than difference) in social judgements.
In a more complex comparison Rashotte and Wedell (2012) showed that quantities that increase on a linear scale (such as the tempo of music) are prone to contrast effects when contextually manipulated. Conversely, quantities that rise to a peak and fall in a parabolic distribution (such as the pleasantness of music, at increasing tempos) show assimilation effects when contextually manipulated. Consequently an awareness of the shape of the underlying distribution of the quantity being judged may mediate which effect occurs.

In the performance appraisal literature Sumer et al (1996) showed that when anchor performances were observed and rated, a contrast effect occurred, whereas when anchor performances were observed without being rated, an assimilation effect occurred. Similarly, Smither et al (1988) showed that observation and rating of videoed performances produced contrast effects whereas rating written descriptions of performance produced assimilation effects.

In further investigation of performance ratings Damisch et al (2006) examined judgements on sporting performance. Within Olympic gymnastics judgements they showed that ratings given to performances were positively correlated with the judgements given to prior performances ($r(1307) = 0.30, p<0.01$). As a result, they demonstrated an assimilation effect. Moreover the effect persisted, but reduced, across the 2nd and 3rd most recent competitors. To establish causality of these observations, Damisch and colleagues then experimentally manipulated the proficiency displayed by initial anchor performances. In each case, participants’ judgements defaulted to assimilation, only showing contrast effects when manipulated to do so, although fairly small manipulation (the nationality of the competitor) could induce this switch. Notably, Damisch et al’s findings contradicted those of Tanner et al (2008) (discussed above) in that on this occasion sequential observation and rating of performances produced assimilation rather than contrast.
In summary, a range of issues influence whether (for a given context) contrast-relativity effects or assimilative-anchoring effects prevail: simultaneous vs. sequential presentation, fluency, recall, anchor/target format, the phenomenon’s distribution and a range of other contextual influences that produce focus on either similarity or difference. As a result it is not currently clear whether general predictions can be made about whether contrast-relativity or assimilative-anchoring effects will occur in a given context.

Looking more specifically at studies that involved ratings of performance, several studies (Murphy et al. 1985; Becker & Villanova 1995; Becker & C. Miller 2002) showed contrast effects. Conversely the studies by Damisch et al (2006) showed a predominance of assimilation. Consequently in performance ratings, as in other arenas in which either assimilation or contrast effects occur, it is difficult to predict which effect will prevail.

On this basis we chose to investigate the influence of recent-experience on assessors’ judgements within mini-CEX assessments, to determine whether recent-experience influenced score choices, and if so, whether contrast-relativity or assimilative-anchoring effects would predominate.

Additional methods for chapter 4

The methods, findings and discussion of the second study are described in the paper contained within this chapter. The description of the development of stimulus materials within the paper is brief, owing to word-limit constraints imposed by the journal format. Consequently these methods are now described in more detail here.
For this study we used all nine videos developed in the first study (see Chapter 2 for more on their development), but recruited a new sample of participants and developed a new methodological approach. The study design required that participants were asked to observe and judge performance in six sequentially presented videos. We wished to ensure that observation of the videos was not unduly taxing on participants as this had the potential to reduce concentration and to impose undue cognitive load. To this end the existing videos were shortened to approximately four minutes each. Consequently watching six videos took 24 minutes, and so we anticipated that study participation would take roughly 30 minutes. We judged that this was a feasible study duration given the constraints of internet experimentation that have already been described.

In study 1 (described in chapter 3) we noted that most participants’ comments were in response to either the “history of the presenting complaint” section or sections at the end of the consultation in which foundation doctors gave explanations to the patients. Fewer comments arose in response to sections in which routine data was gathered: past medical history, drug history, family history etc. Consequently videos were shortened by removing these portions. A screen message was added to each video to make it clear that a portion of the interview had been removed. This ensured that study participants were aware that the intervening section had been removed from the consultation by researchers.

Following shortening, the new shortened videos were revalidated. This involved asking a panel of expert reviewers to view the shortened videos. Three of these reviewers had participated in the validation exercise for the prior study; the remainder were new. Reviewers were all consultant physicians who were experienced in clinical supervision of foundation doctors. Because of the broad range of scores that videos received in the first study we chose to validate the shortened videos using ordinal rankings rather than scores.
Reviewers were asked to watch each of the three performances within each case. After watching the three performances contained within a case, they were asked to place them in order of “best”, “worst”, or “middle”. They also had the option to indicate that two or more videos showed equivalent performances and so could not be ordered. Following this process, they moved to the next case. We asked reviewers to order performances within cases as we judged that ordering all 9 videos in one task would place an undue burden on reviewers’ short-term memory, introducing unreliability. Consequently we obtained within-case ordinal rankings for each of the three performances within each case; we did not cross validate performances across cases. Within each case, reviewers viewed the performances in different orders as determined by a Latin square. Consequently it is unlikely that reviewers ordering resulted as an artefact of presentation order. Results of this process are presented in tables 1-3:

Table 3: Reviewers ratings of shortened videos for Case 1 (Respiratory)

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>Good</th>
<th>Borderline</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>2</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>3</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>4</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>5</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>6</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>Proportional agreement</td>
<td>6/6</td>
<td>6/6</td>
<td>6/6</td>
</tr>
</tbody>
</table>
Table 4: Reviewers ratings of shortened videos for Case 2 (Loss of consciousness)

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>Good</th>
<th>Borderline</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>2</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>3</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>4</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>5</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>6</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>Proportional agreement</td>
<td>6/6</td>
<td>6/6</td>
<td>6/6</td>
</tr>
</tbody>
</table>

Table 5: Reviewers ratings of shortened videos for Case 3 (Upper GI bleeding)

<table>
<thead>
<tr>
<th>Reviewer</th>
<th>Good</th>
<th>Borderline</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>2</td>
<td>Best</td>
<td>Worst</td>
<td>Middle</td>
</tr>
<tr>
<td>3</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>4</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>5</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>6</td>
<td>Best</td>
<td>Middle</td>
<td>Worst</td>
</tr>
<tr>
<td>Proportional agreement</td>
<td>6/6</td>
<td>5/6</td>
<td>5/6</td>
</tr>
</tbody>
</table>

*2 other reviewers commented after scoring (whilst still blind to intended order) that Poor and Borderline for this case were fairly similar, but still agreed with stated order
As the results demonstrate, for cases 1 and 2 agreement with the intended order was complete; for case 3 five out of six reviewers ordered the cases as intended whereas one reviewer placed the “borderline” and “poor” in the opposite order from intended. On this basis we judged that the shortened videos had adequate support to be used in the intended manner in the subsequent study.

Study 2 is described in the paper presented next. Participant information sheets and SPSS output tables relating to the study can respectively be viewed in appendices 7 & 8.
Study 2 is described in the paper titled:

**Effect of Exposure to Good vs Poor Medical Trainee Performance on Attending Physician Ratings of Subsequent Performances**
Yeates P, O'Neill P, Mann K, Eva K.
Available at:


This paper should be obtained and read as part of the thesis, but is omitted from the on-line publication to prevent copyright infringement.
Chapter 5
You’re certainly relatively competent.

Introduction

In the previous chapter we investigated whether assessors’ judgements are influenced by either contrast-relativity or assimilative anchoring effects. We showed in an experimental setting that assessors’ recent experiences produced a moderate-sized contrast effect on their scores. Moreover, we compared this with the influence of assessors’ more fixed “Hawk / Dove” tendencies and showed that the influence of recent experience accounted for a greater proportion of the observed score variance than Hawk / Dove differences between assessors.

Rationale for further study

Whilst these findings are useful and informative, they leave a number of issues unresolved. Notably, the previous study only examined the influence of assessors’ recent experience on “borderline” performances; we cannot assume that the same phenomenon will apply to performances that are either good or poor. It may be that assessors possess clear ideas (or categories) of “good” and “poor” but have less clear representations of intermediate performance. One assessor in the study described in chapter 3 made a comment that seems consistent with this position:

\[\text{if you get real big outliers ... that’s clear. But ... there may be a little bit of a soft focus in the middle}\]

Assessor 1, follow-up interview
This idea is also consistent with the findings of Tweed and Ingham (2010) (described in Chapter 1) which showed that assessors were over-confident when considering performances near to the threshold of acceptability, but under-confident at extremes of performance. “Over-confidence” as defined in this context arises when individuals are correct less often than they believe. As a result, Tweed and Ingham’s findings suggest that assessors were comparatively poor at making judgements around the threshold of acceptable performance. Thus their study supports the idea that borderline decisions may be particularly difficult. Consequently it is difficult to predict whether the contrast-relativity effect that we observed for borderline performances will also occur for good or poor performances. As a result this warrants further investigation.

In Study 1 (described in chapter 3), assessors, whilst forming judgments, made frequent comparisons to other trainees. In addition to these comparisons, some assessors also commented on how they thought the current performance would rank amidst a cohort of the foundation doctor’s peers. One assessor stated that the poor performance they observed was typical of the way that 70% of trainees at that level would perform; by contrast a different assessor stated (with respect to the same performance) that they believed they had rarely if ever seen a trainee at that level perform so poorly. Moreover, the scores given by these assessors reflected these disparities.

Consequently it appears that in addition to comparisons with singular other exemplar trainees, assessors may also have compared observed performances with a mentally-held “cohort” of typical trainees – or at least with what they perceived to be typical. In so doing, it appears that assessors’ competence scores may be influenced by their perceptions of the rank or ordinal position of the observed performance, i.e., how it would rank if the entire
cohort were to perform the same task. This has a number of implications. Firstly, it may help to explain assessors’ score variations: if assessors’ perceptions of the cohort’s performance are different, then the rank position of the current performance will vary, and so may influence the scores that assessors give. Secondly, this may represent further evidence that assessors’ judgements are not truly criterion-referenced (judged against a fixed uniform criterion (Wass et al. 2001)), as assessors may interpret “fixed” criteria by comparing against their mentally held cohorts of typical performance. As a result these criterion-based judgements may actually be comparative.

Logically we may suppose that in order to determine what is “typical” of trainees at a particular level, an assessor must refer to an abstracted idea of the distribution of trainees’ performances and the frequencies at which they occur. Consequently it is likely that in order to determine how a given performance would rank amidst a typical cohort, an assessor must draw on representations of the performance of numerous trainees held in long-term memory. Comparisons based on “perceived-rank” most likely differ from comparisons with recently-experienced trainees: the former relies on long-term memory, whereas the latter relies on what is available in the immediate context. In both approaches, assessors make comparisons, but in the former they compare against memories which may be comparatively fixed, whereas in the latter they compare against the variability of immediate context.

As a result it is possible that by appealing to the perceived rank of a performance, assessors may draw more on long term memories of trainees’ performance and less on immediate context. Consequently the influence of recent experience may be ameliorated. Alternatively, it is possible that even long-term representations of trainee performance are fairly weakly
represented, and that despite appeals to the perceived rank of a performance, judgements remain susceptible to recent experience.

Three issues have so far emerged through these considerations: whether recent experience will influence all levels of performance or just borderline performances; whether variations in assessors’ scores can be explained to some degree by variations in the perceived rank of performances; and whether appealing to perceived rank will mitigate the influence of recent experience or (conversely) whether the perceived rank of performances will also be influenced by recent experience. In order to further understand the influence of assessors’ recent experiences in producing contrast effects on their scores, the study presented in this chapter will address each of these issues.

Before proceeding a further consideration is pertinent. As described earlier, Tweed and Ingham (2010) showed that assessors’ confidence in their scores showed little overall relationship with their accuracy, with assessors tending to be overconfident around the threshold of acceptable performance and under confident at the extremes of performance. Confidence in a decision can be defined as the subjective probability that a decision is accurate (Paese & Sniezek 1991). It therefore follows that the more that a person is aware that their decision may lack accuracy, the less they should feel confident in it.

It is difficult to determine whether the contrast effect described in Study 2 reduced assessors’ accuracy per se as the performances had no objectively defined true score. None the less, we might expect that if an assessor is aware of possessing only a weak representation of the assessment’s criteria they would experience reduced confidence in their decision. Correspondingly, we might expect that assessors who possess the weakest representations of criteria would have the greatest tendency to draw on cues in their
context, and that consequently such individuals would be the most susceptible to the influence of recent experience. It follows, therefore, that assessors who report the lowest confidence should be the most susceptible to recent experience, and that the two quantities (confidence, and susceptibility to recent-experience) should be inversely related. Of course this supposition assumes that assessors possess insight into the adequacy of their judgement strategies, and their ability to be influenced; the existence of a different relationship between confidence and susceptibility to recent-experience may indicate that assessors lack such insight. Consequently we chose to study whether participants’ decision-confidence can predict their susceptibility to manipulation through recent experience. In so doing, we hope to understand whether assessors have insight into this susceptibility.

**Brief literature review on the role of confidence in judgements**

An extensive literature exists on confidence in decision making; some principles from this literature are worthy of brief consideration. The precise mechanism by which an individual judges their accuracy in a decision remains controversial, with the existence of a number of competing theories (Gigerenzer et al. 1991; Koriat 2008). Regardless, confidence seems to arise from a metacognitive appraisal by an individual of their ability to make a correct decision (Mitchum & Kelley 2010). This process relies on inferences made from a range of cues that are used during the decision process, rather than from any direct appraisal of the knowledge that the person has in their memory (Koriat 2008; Mitchum & Kelley 2010). Nonetheless, these inferences arise from the degree to which the individual’s existing knowledge enables them to solve the problem at hand. Inappropriate confidence (for example over-confidence) may come from individuals using cues that are not diagnostically useful (Gigerenzer et al. 1991; Prowse Turner & Thompson 2009). In so doing, an individual may think that they are relying on useful information (increasing confidence) but may be no nearer to the correct answer (therefore not improving accuracy). In such a scenario (as described) the individual would lack insight into the adequacy of their decision making.
The ability of a person to monitor their decision-making accuracy is potentially very useful; appreciating the trustworthiness of a decision can guide the degree to which the decision is relied upon for important behaviours. Low decision confidence can prompt a search for more information, or a better decision, whereas high confidence may halt information search, saving resources and allowing action to proceed (Koriat 1993; Prowse Turner & Thompson 2009).

Despite this, human-beings’ self-monitoring capacity is imperfect and decision confidence and the actual accuracy of decisions can be poorly related. In perhaps the clearest example of this disparity, DePaulo and colleagues (1997) performed a meta-analysis on 18 studies that examined the relationship between the ability of individuals to detect deception, and their confidence in their judgements. They showed that the relationship between the two did not significantly differ from zero. Confidence and accuracy, therefore, were not related.

More optimistically, a number of authors have shown a positive relationship between confidence and accuracy. Roberts et al (2002) showed that when asked to recall information about previously-viewed videos, statements that were made with high confidence were more often accurate than statements made with low confidence. Similar findings were shown by Sauerland et al (2009) and Kebbell et al (2009). In particular these latter studies show that individuals reporting very low confidence in their recollections did indeed show very low accuracy. Other investigators have found curvilinear relationships, with the two quantities showing a relationship for lower levels of confidence, but dissociating for higher confidence levels (Ames et al. 2010).
Whilst some general trend exists between individuals’ subjective confidence and objective accuracy, a further phenomenon is clear: that of “overconfidence”. Individuals’ estimates of the proportion of responses they have answered correctly are typically higher than the proportion they actually got correct (Lichtenstein & Fischhoff 1977). For example individuals who have estimated they were 70% accurate are typically only 60% accurate (Koriat et al. 1980). This tendency occurs for most levels of difficulty; individuals’ confidence in their decisions increases as their accuracy increases, but their confidence is persistently unduly high.

The consistent exception to this phenomenon occurs for very easy items, where the pattern reverses and individuals’ confidence is actually lower than their accuracy (Lichtenstein & Fischhoff 1977; Koriat et al. 1980). In view of this “Hard-Easy” effect, the previously described results by Tweed and Ingham (2010) (which showed overconfidence around threshold-of-competence decisions but under-confidence at greater extremes of performance) may indicate that assessors in medical education find decisions on the extremes of performance easy, but decisions near to the threshold of competence more difficult.

A few factors can be seen to mediate this overconfidence affect. When people receive more information relating to a decision they form richer representations of the decision problem. This increases their confidence without necessarily increasing their accuracy (Gill et al. 1998; Tsai et al. 2008). People recall information that is consistent with a chosen option more than information that contradicts their choice. Consequently (because they are less aware of the contradictory evidence) people feel more certainty than is warranted (Harvey 1997). Consistent with this, instructions to deliberately consider alternatives reduce over-confidence (Harvey 1997; Koriat et al. 1980).
Goodman-Delahunty et al (2010) showed that men were more prone to overconfidence effects than women, and (in contrast to the Koriat et al. 1980) also showed that instructions to consider alternatives did not reduce this effect. This suggests that this influence may operate inconsistently across different circumstances. A further contribution to overconfidence may arise from “option fixation” (Sieck et al. 2007) in which people search for highly familiar options, and then convince themselves that these are true. This makes them feel confident, but may not produce corresponding accuracy. As a result overconfidence occurs.

Confidence and the influence of context

Whilst these studies considered the way in which confidence related to accuracy, our interest concerned whether assessors’ confidence in their scores showed any awareness of their vulnerability to contextual manipulation – particularly whether some awareness of possessing well developed, fixed criteria would produce both confidence and resistance to contextual manipulation. Comparatively little literature has examined this possibility.

Friedlander and Stockman (1983) investigated assimilative anchoring effects in clinical judgements of psychiatric patients by manipulating the timing of information presentation, and whether participants would be held publicly accountable for their responses. They showed an anchoring effect in one of the two cases that participants rated. Confidence ratings were also obtained, and were universally high in all groups (mean of 5.4 out of 7). They subsequently performed a similar study using undergraduates (Friedlander & Phillips 1984), in which they failed to produce an anchoring effect. Confidence ratings by students were somewhat lower than in the previous study (mean 3.3 out of 7). Consequently, confident clinicians in the first study were more susceptible to anchoring than less confident students in the second. Importantly, as both the designs and populations varied between
studies, these results could have numerous explanations. None the less, these researchers speculated on the basis of these findings that confident individuals may be less willing to deviate from their initial judgements than less confident individuals, making them more susceptible to anchoring effects.

Jacowitz and Kahneman (1995) more directly investigated the link between judgement confidence and anchoring effects. Participants in their study were exposed to either a high or a low anchor (or no anchor in a control group) and then provided ratings of 15 quantities. Comparison of the two experimental groups’ responses with the control group allowed them to express findings as an “anchoring index (AI)” and pool results on the same scale. Confidence ratings were obtained for each item by each participant on a 10 point scale. Over the 15 quantities, items rated with lower confidence showed greater degrees of anchoring ($r = -0.68, p<0.05$). Consequently the two quantities were inversely related, and therefore contradicted the relationship suggested by Friedlander and colleagues. Jacowitz and Kahneman then isolated the 25% most confident participants for each judgement, and showed that they had a mean AI of 0.28, somewhat lower than the overall mean of 0.49. Consequently highly confident participants were less prone to anchoring than the overall study population.

Finally Jacowitz and Kahneman examined whether high confidence participants were more often accurate than low confidence participants. They showed that compared to the low confidence group, the high confidence group were nearer to the correct value on 8 out of 15 occasions, further from the correct value on 5 out of 15 occasions and equal on 2 occasions. They interpreted these findings as being unlikely to indicate superior performance by the high confidence group. The increased confidence of these participants did not reflect
increased accuracy even though they were less susceptible to anchoring; participants’ judgements were more “fixed” rather than more accurate.

In summary, the limited number of studies that examined the relationship between confidence and susceptibility to anchoring showed opposite results. The study by Jacowitz and Kahneman (1995), however, had a stronger methodology as the comparison was direct within the same study population and the anchoring manipulation was more effective. Importantly, confidence was not manipulated in these studies so the observed relationships cannot be assumed to be causal. None the less, the study by Jacowitz and Kahneman shows that susceptibility to context effects (in this case anchoring) can be predicted by participants’ decision confidence in some circumstances. Moreover, as illustrated here, it is reasonable to suppose that this occurred because of metacognitive monitoring of the judgement process by participants, with those who were more “fixed” apparently being aware that their judgements were better than those who were less “fixed”. In this respect our assertion that insight mediates this inverse relationship seems viable. Of course this relationship did not occur in the studies by Friedlander and colleagues (1983; 1984). This may have occurred either due to methodological weaknesses in the studies, or alternatively because participants lacked insight into their judgements and their susceptibility to manipulation.

On this basis it seemed useful to investigate whether assessors’ susceptibility to contrast effects (due to recent experiences) could be predicted by their decision confidence. In querying this relationship we wondered whether assessors possess insight into their susceptibility to manipulation through contrast effects.

Study 3 is described in the paper presented next. Participant information sheets and SPSS output tables relating to the study can respectively be viewed in appendices 9 & 10.
Study 3: is now described in the paper that follows:

You’re certainly relatively competent

The influence of recent experience, perceived rank and confidence on assessors’ scores

In submission to the journal Medical Education.
You’re certainly relatively competent

The influence of recent experience, perceived rank and confidence on assessors’ scores

Peter Yeates, MBBS, M ClinEd
University of Manchester, School of translational medicine / The Academy at UHSM, University Hospital of South Manchester
ATR4, ERC, University Hospital of South Manchester
Southmoor Road, Manchester, M23 9LT,
United Kingdom
peter.yeates@manchester.ac.uk
tel: 0161 448 8063
fax: 0161 291 4974

Paul O’Neill, MBChB, MD
University of Manchester, Manchester Medical School / The Academy at UHSM, University Hospital of South Manchester
p_a_oneill@btinternet.com

Karen Mann, PhD
Dalhousie University, Division of Medical Education
University of Manchester, Manchester Medical School
karen.mann@dal.ca

Kevin Eva, PhD
University of British Columbia, Centre for Health Education Scholarship (CHES), Department of Medicine
kevin.eva@ubc.ca
Abstract

**Background:** Recent study has suggested that assessors judge performance comparatively rather than against fixed standards. We extended investigation into assimilation effects (judgements become unduly similar to recent performances) or contrast effects (unduly different) in assessors’ judgements, examining these effects across a range of performance levels. Further, we investigated whether assessors possess insight into the effect (whether decision-confidence predicted susceptibility to manipulation) and whether stimulating memory (by considering typical performance at that level) lessens influence of recent experience.

**Methods:** Consultant physicians were randomised to groups within an internet experiment. Group A judged videos of F1 (PGY1) doctors in descending order of proficiency; group B judged the same videos in ascending order. Participants provided 1. competence scores, 2. confidence ratings, 3. “percent better” (the percentage of other F1 doctors who would perform better on the same task) ratings for all videos. Ratings were compared between groups using 2x3 ANOVA.

**Results:** Overall, Group A scored lower than Group B (2.97 vs 3.50, F(1,47)=9.80, p=0.003, d=0.52). Pair-wise comparisons showed differences were significant for good and borderline performances. PB ratings similarly showed worse scores for Group A than Group B (Group A mean PB: 57.4, group B mean PB 43.4, F(1, 46)=16.0, p<0.001, d=0.67). PB ratings substantially explained variations in assessors’ competence scores ($r^2$ 0.5-0.88). Confidence ratings did not vary by level of performance and showed no relationship with the effect of group.

**Discussion:** Assessors’ judgements showed contrast effects at both good and borderline performance levels. Findings suggest that assessors use normative rather than criterion-referenced criteria whilst judging, which are weakly represented in memory and easily influenced. Confidence ratings suggested a lack of insight into these phenomena. Judgements could be importantly influenced in ways that are unfair.

Abstract word count: 283
Background

Workplace based assessments of clinical trainees’ performance seek to provide educators and trainees with useful insight into clinical trainees’ authentic performance (1–3). Despite general support for the assessments’ use (4–6) and educational effectiveness (7,8) their utility are limited by high inter-rater score variability (9,10). As neither training interventions (7,8,11) nor changes in scale format (12,13), have produced the desired degree of improvement in inter-rater score variability, a novel field of enquiry has sought to understand assessors’ judgment processes, with a view to developing theoretically based interventions (14–16). This paper will contribute to this evolving understanding.

In contrast to recommended criterion-referenced, competence-based approaches to assessment within medical education (17), our recent work suggested that assessors frequently compare trainees whilst judging their competence (18). In an initial follow-up study we showed that assessors’ judgements of borderline performances were susceptible to a “contrast effect”: assessors who had recently observed and scored good performances gave lower scores than assessors who had recently observed and scored poor performances (18). This finding is important because it offers an unexplored source of variance in assessors’ scoring that illuminates the potential to bias assessment scores, making them unfair. In this study we explored this effect further, examining whether it occurred at other levels of performance and the extent to which it interacts with other influences.

The psychological literature suggests that when an assessor compares a currently observed “target” stimulus (e.g., a student performance in our context) with other “anchor” stimuli (e.g., other examples of student performance), one of two effects may occur: contrast effects or assimilation effects. In assimilation effects a target performance is viewed as having undue similarity to anchor performances. Consequently, scores for the target unduly reflect the performance of anchor performances. This effect arises in part due to “incomplete adjustment” from the original anchor performance (19,20), and partly because information concerning the anchor performance is “selectively accessed” causing similarities between the target and anchor to be preferentially noticed (21).

Conversely, the opposite occurs in contrast effects: differences between target and anchor performances are over-emphasised, leading to scores that unduly diverge. This occurs because humans make subjective judgements of an item’s magnitude by estimating the combination of its rank position in a distribution and its distance from the limit of the distribution (22,23). Information in the immediate context is used more readily than information in memory (24) and so contextual anchors skew the perceived distribution, producing contrast effects (22). Both effects are well demonstrated; understanding which effect dominates in medical education performance assessment judgements is necessary prior to planning theoretically-based interventions.

Assimilation effects have classically involved the provision of an overtly irrelevant number (e.g., from a “wheel of fortune”) which then (despite its obvious irrelevance) biases subsequent judgements on an estimation task. That is, peoples’ estimates unduly reflect the irrelevant number they were provided (25). The effect has occurred when participants are highly experienced at the task (26,27), and even when participants generate their own anchors (28). Anchors induced by one task can influence judgements on unrelated tasks (29). The effect may be mediated through emotion (it effects sad or fearful people more than angry or disgusted people) (20), and has been shown to be inversely related to participants’
confidence in their judgements: lower confidence participants showed greater degrees of assimilation (30).

Contrast effects were first demonstrated on perceptual judgements (e.g., when judging weight or temperature) (31), but have subsequently been shown within person perception (22), and for abstract concepts such as life satisfaction (32) and gratitude (33), as well as judgements of monetary value within economics (34,35).

A number of studies have sought to determine conditions that favour either assimilation or contrast. An initial overall impression of similarity between items tends to produce assimilation whereas an initial impression of difference produces contrast (21,36,37). Simultaneous presentation of anchor and target stimuli tends to promote assimilation, whereas sequential presentation favours contrast effects (38). Prompting participants to recall a prior anchor produced contrast, but exposure to the anchors without a reminder produced assimilation (39). Similarly presenting fluent, easy to read, anchors produced assimilation whereas disrupting ease of reading produced contrast (40). Consequently, assimilation may be the default, but tasks that involve more deliberate consideration may be prone to contrast effects (21).

Within assessments of performance, findings have varied. A number of studies from occupational psychology have shown contrast effects (41–43). Conversely Damisch et al (44) found that judges of athletic performance defaulted to assimilation in their judgements, only producing contrast effects when manipulated to do so. Consequently it is difficult to make general predictions about which effect will dominate within medical education.

In this study we sought to replicate the contrast effect that we observed in our previous study involving borderline performance. As work by Tweed and Ingham (45) showed that decisions on intermediate levels of performance may be particularly difficult for assessors, it is possible that contrast effects could be limited to borderline performances. Therefore we sought to determine whether a similar effect also occurs at other levels of performance (i.e., good and poor performances). Additionally we queried the relationship of two other parameters: one indirectly queried the degree of assessors’ insight into the effect; the other tested whether appeals to assessors’ memory could mitigate the effect.

Decision-confidence is known to arise due to a metacognitive appraisal of the adequacy of a person’s decision processes (46,47). Concordantly, people with higher decision confidence should be less susceptible to contextual manipulation, if, indeed, they possess insight. This relationship was shown by Jacowitz and Kahneman (30) who found that susceptibility to assimilation effects was inversely related to confidence. By measuring assessors’ decision-confidence and relating it to their susceptibility to the influence of recent experience we sought to understand their degree of insight into the contrast effect. Additionally as the relationship between confidence and accuracy varied by level of performance in Tweed and Inghams’ study(45), we determined whether confidence varied by level of performance.

Finally we considered the effect of explicitly drawing attention to past examples assessors have encountered. As stated previously, contrast effects occur because individuals are influenced more by information in their immediate context than by information stored in their memory (24). Some assessors’ in our prior exploratory study (14) spontaneously described differing perceptions of the rank positions of performances they observed (i.e., where the performance would rank if a cohort of typical trainees performed that task). In order to consider the rank of performance, an assessor must bring a distribution of performances to mind. Therefore, explicitly, asking assessors about their perceptions of the
The ranking of target performances is likely to draw more attention to memory-based cases and lessen focus on the immediate context than is intended by asking questions about the ratee’s competence. Consequently, we hypothesized that asking assessors about the ranking they perceived a performance to fit would mitigate their susceptibility to contrast effects. Additionally, we questioned the relationship between assessors’ perceptions of rank position and the competence scores they produced.

Research Questions

On this basis we posed the following research questions:

1. Do assessors’ recent experiences influence their competence scores for good, borderline and poor performances?
2. Do assessors’ recent experiences influence their ratings of perceived rank of performances?
3. Do assessors’ ratings of the “perceived rank” of performances explain variations in their competence scores?
4. Do assessors’ ratings of confidence in their scores vary by level of performance or relate to their susceptibility to influence by recent experience?

Methods

Participants:
The study population was consultant physicians from the United Kingdom (senior doctors working in specialities associated with general (internal) medicine, who are overseen by one of the United Kingdom Royal Colleges of Physicians). We also included emergency medicine physicians who had trained via the Royal College of Physicians’ exam system, as they are frequently involved in supervising trainees in the management of acute presentations of general internal medicine patients. To meet inclusion criteria participants had to be comfortable assessing general internal medicine case material and estimate that they assessed at least 5 mini-CEX’s per year (The mini-CEX is a commonly used direct-observation performance assessment used to assess consultation or physical examination skills (48,49)). As our prior study showed that the number of years that a participant had been a consultant had no influence on their scores (18), no minimum duration of consultant employment was required for individuals to be eligible to participate. Individuals who had previously completed studies by our group were allowed to participate, but their “prior participant” status was noted so that differences between them and new participants could be examined during analysis. Individuals were not allowed to participate more than once in this study.

Recruitment:
Email invitations were cascaded throughout the United Kingdom from the national UK Foundation Programme, to regional Foundation Programme directors to local foundation tutors and then to individual educational supervisors (consultant physicians with responsibility for supervising foundation doctors (PGY1-2)). Additionally we used publicly available email addresses to directly mail consultant physicians throughout the UK. Interested individuals were asked to email the research team and were provided with study information, and a link and password to the study’s internet site. Reminder emails were sent to non-responders. Consent to participate and an indication that they met inclusion criteria were provided on-line by participants prior to starting. Participants were not paid for participating.
**Design:**
The study used an internet-based, experimental design. Participants were randomised to one of two groups: Group A viewed 6 videos of performances of foundation doctors in descending order (2 good, 2 borderline, 2 poor); Group B viewed the same 6 videos in ascending order (2 poor, 2 borderline, 2 good). Consequently, Group A always saw videos that were worse than previous when the level of performance changed, whereas Group B always saw videos that were better than previous when the level of performance changed. See figure 1 for a schematic. Within groups the order of videos was counterbalanced within each level of performance.

**Videos:**
Videos featured scripted performances by real foundation year 1 doctors (PGY1) in consultations with simulated patients. In each case the simulated patient depicted a new presentation to hospital with an acute medical problem. Videos featured two cases: unexplained loss of consciousness in a 34 yr old man and probable upper gastrointestinal bleeding in a 44 year old man. For each case, 1 video depicted “good” performance, 1 video depicted “borderline” performance and 1 video depicted “poor” performance. To avoid the potential for race or gender to bias judgement (50), all foundation doctors were white females, and were dressed uniformly in blue hospital scrubs. The same SP played the patient through each case. The videos appeared to occur in a realistic clinical environment. To avoid placing participants under undue cognitive load, videos were approximately 4 minutes long and depicted the “history of presenting complaint” and “explanation of diagnosis and management” sections of the consultation. More routine data gathering sections were omitted. Videos had been developed and validated as part of prior research (14,18).

**Response formats:**
After each video concluded, participants were asked to answer three questions.

First, **competence scores** were collected using the standard United Kingdom foundation programme mini-CEX assessment format (49). This required participants to allocate scores for each of 7 domains: history taking, physical examination, communication skills, critical judgement, professionalism, organisation / efficiency and overall clinical care. Scores were given on a 6 point Likert scale with options of 1: well below expectations for foundation year 1 (F1 / PGY1) completion, 2: below expectation for F1 completion, 3: borderline for F1 completion, 4: meets expectations for F1 completion, 5: above expectation for F1 completion, 6: well above expectations for F1 completion, as well as an “unable to comment” option. Scores were averaged across the 7 domains to give a single score. This gave the dependent variable “score”.

To obtain **confidence ratings**, participants were asked to answer the question, “Balancing things out, please rate overall how confident you feel in the scores that you have just provided”. They were asked to respond on a 7 point Likert scale. This was anchored at point 1 with the text, “I am not at all confident in the scores I have given” and “Examples of this might be: I have substantial doubts in my scores, the ‘correct’ scores could easily be different to the scores I’ve given”. At point 7 the scale was anchored with “I am very confident in the scores I have given” and “Examples of this might be: I have almost no doubts in my scores. The ‘correct’ scores are most probably not different to the scores I’ve given”. Participants could also indicate that they felt unable to comment in response to this question. This produced the dependent variable “confidence”. Seven point confidence ratings have previously been used in other research on judgemental confidence (51).
To obtain ratings of perceived rank, participants were asked, “What proportion of F1 doctors would you expect to perform better, worse, or about the same as this doctor if they were to undertake the same task?” They then used drop-down menus to respond to three questions: “Percentage of F1s that would do better than this performance (5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%)”, “Percentage of F1s that would do worse than this performance (95%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, 5%)” and “Percentage of F1s that would do the same as this performance (0% to 95% in 5% increments). Participants could also indicate that they felt unable to comment in response to this question. If the three responses did not total 100%, participants were alerted and asked to re-consider their responses. The response to “percentage of F1s that would do better” gave the dependent variable “percent better”.

Prior to inclusion, both new response formats (i.e., confidence and percent better) were pre-tested and developed using consultant physicians. Respondents indicated that these response formats were unambiguous, and could be meaningfully interpreted in the context of assessing foundation doctor performance.

Upon completion of the 6 videos, demographic data were collected. This included gender, amount of time the participant had been in consultancy, and the number of mini-CEX ratings assessments the individual typically provides (across any grade and within F1 specifically) per year. Participation took roughly 30 minutes. Ethical approval was issued by University of Manchester ethics committee.

**Analysis:**
Data were analysed using SPSS 15 software, by 3 (level of performance) by 2 (group) repeated measures ANOVAs. We used “competence scores” as the dependent variable for RQ1, later individually adding demographic data as covariates to check for confounding. For RQ2 we used “Percent Better” as the dependent variable. Adjustments for non-sphericity were made where required with the Greenhouse-Geiser correction.

For RQ3, we firstly calculated the within-participant correlation using the method of Bland and Altman (52). Next we calculated individual Pearson’s product-moment correlations for score and “Percent Better” for each video within each group. This gave 12 correlation coefficients. We then calculated the median and range of these correlations.

For RQ4 we firstly examined whether “confidence” varied by group or level, or was influenced by demographic covariates. Next we examined the interaction of group by confidence in an rANCOVA model (“scores” as dependent variable). Finally we split participants based on their mean confidence ratings into high and low confidence groups (median split). We then separately analysed the influence of level and group on participants’ scores for the “high” confidence and “low” groups. Throughout analyses, pair-wise comparisons were adjusted with the Bonferroni correction opting to adjust p values rather than α levels (53). Effect sizes were calculated using an adaptation of Cohen’s d in which the pooled standard deviation is estimated from the square root of the mean-square error term.

**Results**

**Baseline data**
Forty-nine consultants participated in the study, from 14 clinical sub-specialities, and from 12 regional Deaneries covering most of the UK. Groups were similar in terms of gender and
duration of consultancy. Group B showed slightly higher mean values for numbers of assessments completed annually by assessors. These data are shown in table 1.

**Dependent variables**
The distributions of all three scores were examined (skewedness / kurtosis / mean-median similarity) and deemed to be acceptable for parametric analyses. Between-participant scores for each video showed considerable variation, with one video showing a score range of 1-6 on the six point scale (video 4), thus replicating the anticipated degree of inter-assessor variability.

**The influence of experimental group on competence scores:**
Participants’ competence scores by level and group are presented in figure 2. Scores given to the 3 levels of performance differed (Good mean score: 4.6, borderline mean score: 3.1, poor mean score 2.0, F(1.69, 94)=328.1, p<0.001), and consequently were in the intended range. The mean score on the Good videos for Group A and on the Poor videos for Group B provide the best estimate of the actual scores those videos deserve as those ratings are not biased by the experimental manipulation. Overall, participants’ mean scores differed by group, with group B scoring higher than group A (Group A mean score: 2.97, Group B mean score: 3.50, F(1,47)=9.80, p=0.003, Cohen’s d=0.52). Consequently these results indicated a contrast effect, with the group who observed videos in ascending order (Group B: poor to good, i.e. current video was always better than preceding videos) showing higher scores than the group who viewed videos in descending order (Group A: good to poor, i.e. current video was always worse than preceding videos).

Examination of pair-wise comparisons showed that the group effect was significant for “good” and “borderline” levels of performance, but insignificant for “poor” performances though the overall interaction of group x level was non-significant (F (1, 94)=1.6, p=0.22). When demographic variables were included in analyses as covariates duration of consultancy showed a trend towards significance (β range -.03 to -.05, p=0.076), but did not interact with group indicating that this variable did not confound the main comparison of interest. No other covariates (gender, participation in prior studies, estimated numbers of mini-CEXs assessed by participants) approached significance.

**The influence of experimental group on perceived rank:**
Participants’ Percent Better (PB) ratings by level and group are presented in figure 3. Ratings varied by level of performance (good mean PB 18.2, borderline mean PB 55.6, poor mean PB 77.4, F(2, 92)= 287.9, p<0.001). Again, Good videos in Group A and Poor videos in Group B provide unbiased estimates of assessors’ opinions. Overall, participants’ PB ratings varied by group. Group A ranked performances lower than Group B (Group A mean PB: 57.4, group B mean PB 43.4, F(1, 46)=16.0, p=0.001, Cohen’s d=0.67). It should be noted that for ranks, higher numbers indicate lower rank positions. Consequently, similar to competence scores, PB ratings also showed a contrast effect.

The interaction of level x group was significant (F(2, 92)= 4.58, p=0.013) indicating that the influence of group varied across different levels of performance. Pair wise comparisons showed that at the poor level of performance, differences between groups were only marginally significant.

**Relationship between Scores and Perceived rank of performances**
The within-participant correlation between PB and score was -0.94, p<0.001 giving a corresponding $r^2 =0.88$. 11 of the 12 Pearson’s correlations (and associated $r^2$ values) were
statistically significant and ranged from -0.52 ($r^2=0.27$) to -0.87 ($r^2=0.76$), with a median of -0.71 (median $r^2=0.50$). The one non-significant correlation was $r=-0.41$. Consequently, regardless of method, there was a strong negative relationship between PB and scores, with between 50% and 88% of variance in scores explained by differences in perceived rank of the performances.

The influence of confidence on scores and group effect
Confidence was not significantly related to gender, previous participation, duration of consultancy, or frequency of conducting Mini-CEX assessments. Scores were generally high; the median confidence level was 6, with scores of 3 or less accounting for less than 3% of responses. Participants’ confidence by level and group is shown in table 2. Differences were not significant by either level or group.

The interaction of confidence x group was not significant when a median split was performed on confidence and competence scores were re-analyzed as the dependent variable. The high confidence group showed a significant main effect of group ($F=6.468$ (1,23), $p=0.018$), whilst the low confidence group did not. This difference appears to have been driven by differences in the unbiased conditions (i.e., Good performances in Group A and Poor performances in Group B) and, as such, confidence was not considered to be an important mediator of the contrast effect.

Discussion

Summary of results
In this study we replicated the contrast effect which we had previously demonstrated (18). We examined the influence of assessors’ recent experience on scores given to performances at three different levels: good, borderline and poor. We demonstrated an overall contrast effect, although this effect was only significant at “good” and “borderline” levels of performance. The contrast effect produced an average score difference between groups of 0.54 on the study’s 6 point competence scale showing a moderate effect size (0.52). Differences at the poor level of performance may have been dampened by a floor effect.

Additionally we showed that appealing to longer term memory by asking assessors to indicate the rank they perceived the performances to maintain was similarly susceptible to a contrast effect. On this occasion, as well as being significant at good and borderline levels, the poor level of performance approached significance. Moreover, we showed that assessors’ perceptions of perceived rank of performances correlated highly with their individual variations in competence ratings (accounting for 50 to 88% of the variance depending on the analytic method used).

Finally, we showed that the previously reported negative relationship between confidence and assimilation effects did not occur in this study. The contrast effect that occurred due to assessors’ recent experiences appeared to be independent of their confidence in their decision.

Theoretical implications of findings:
These findings have a number of theoretical implications. Contrast effects appear to be a replicable finding, which can influence more than just borderline performances. We previously suggested that the influence of a contrast effect on borderline videos indicates that assessors may lack a clearly defined mental representation of the assessment criterion.
Tweed and Ingham’s study (45) showed that assessors were particularly over-confident of their judgements around the threshold of acceptable performance, but were under-confident at extremes of performance. Thus their results tended to suggest that threshold decisions are particularly difficult. Our confidence ratings did not replicate theirs and our finding that good performances were also susceptible to contrast effects shows that assessors’ criteria for other levels of performance may be equally poorly represented. Decision confidence is understood to arise due to metacognitive inferences by a person about the adequacy of their ability to make the decision (47). The lack of a negative relationship between confidence and group effect indicates that participants were unaware of their susceptibility to manipulation; they lacked insight.

As described in the background section, comparisons between students could equally have produced an assimilation effect. Although no clear pattern emerged from previous investigation on performance judgements in other arenas, a few features of assessment in medical education may favour contrast effects in judgements. Firstly, assessors may be attuned to differentiate between candidates: this would promote contrast effects (21). Candidates are typically observed and rated sequentially, which also promotes contrast effects (38). This latter observation may cause us to query whether an assimilation effect may occur in tutorial-based settings in which observation of all candidates occurs simultaneously.

Requesting assessors to judge the perceived rank of performances was intended to draw on their longer-term memories of F1 (PGY1) doctors. We anticipated that this might lead PB ratings to be less susceptible to the contrast effect than competence scores. Contrary to this hypothesis, both competence scores and PB ratings were equally susceptible to the contrast effect. This may indicate that assessors’ mental representations of the range and typicality of F1 (PGY1) doctor performance are just as weakly represented as competence criteria. The finding that perceptions of rank also varied substantially between assessors and significantly explained variations in competence scores indicates that assessors’ competence criteria may be normatively-based on their varied (and weakly represented) representations of typical performance. Of course whilst observing this relationship, our data can neither determine whether this relationship is causal, nor its direction.

Limitations of study:
First, our study enrolled consultant physicians and involved asking them to judge performances by F1 (PGY1) doctors on consultation skills; as such, our findings do not implicitly generalise to wider contexts although there is little a priori reason to suggest that these findings should be specific to this domain. Second, given the recruitment strategy employed there is no way of calculating an accurate response rate. The fact that this strategy enabled recruitment from a range of practice locales offsets this limitation to some extent. Further, the most likely dimension across which respondents are likely to differ from non-respondents in this study is that respondents may be more interested in either assessment or education. If anything, that should make them more conscientious/experienced in completing the type of task required for participation than non-responders are likely to be, a difference that has been shown to increase consistency between raters (54) and that we would anticipate to mitigate the effects observed. Finally, the choice of an experimental design maximised the study’s internal validity at the cost of ecological validity. The realism of the study videos and recruitment of genuine assessors, however, should help to lessen concerns about ecological validity.

Implications for practice:
Contrast effects in assessment in medical education have an important potential to bias score judgements of junior doctors and medical students’ performance in a way that is unfair to candidates. The magnitude of the effect is sufficient to cause classification errors near to the threshold of competence, but equally could misclassify students into or out of “honours” or “distinction” categories, unfairly altering their career opportunities. For example, data reported by Mitchell et al (55) showed that real average Mini-CEX scores from a cohort of F1 (PGY1) doctors had a standard deviation of 0.38. Consequently the effect we have reported would dramatically alter the position of a performance within a cohort. Whilst we conducted this study in the context of performance assessments, the effect may be more important in objective structured performance exams (OSCEs) in which candidates’ progress sequentially around examination stations. As each student typically follows the same student around the exam circuit, contrast effects would represent an uncontrolled influence on candidates’ overall scores.

The variation in assessors’ perceived rank of performances suggests that their mental distributions of the range of typical performance at F1 (PGY1) level also vary. Due to competitive entry recruitment processes, hospitals most probably vary in the degree of excellence that their F1 doctors display, with more popular hospitals potentially recruiting higher calibre trainees. Consequently, the standard of the real trainees to which assessors are exposed in their work may vary between hospitals, potentially accounting for varying perceptions of what is typical. As assessors’ competence judgements relate to these perceptions of typical performance, this could allow concerning disparities to arise between hospitals in the standard of competence that their assessors apply.

**Recommendations for further research:**
Having demonstrated a contrast effect twice in a laboratory situation, investigation to examine this influence in real assessment data would be useful to determine the applicability of these findings to real practice. Exploration of factors that may promote assimilation effects rather than contrast effects would also be useful. In due course, with further theoretical understanding of this phenomenon, further research may test interventions to mitigate this effect.

**Conclusions**

Contrast effects appear to be a replicable and robust phenomenon impacting on the judgments of medical education assessors, which are not confined to borderline or threshold cases. This suggests the notion of a true criterion referenced assessment to be farcical in that the competence of any given candidate would appear certainly to be judged relative to other candidates. The confidence ratings collected indicate that assessors lack insight into this susceptibility. These findings have a range of theoretical implications, and in practical terms represent an important challenge to assessment fairness that is currently unaddressed.

Words: 4677

**References**


Tables:

Table 1: Participants’ demographic data at baseline displayed by group

<table>
<thead>
<tr>
<th></th>
<th>Group A</th>
<th>Group B</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>25</td>
<td>24</td>
<td>–</td>
</tr>
<tr>
<td>Participated in previous study</td>
<td>8 (32%)</td>
<td>9 (38%)</td>
<td>$\chi^2(1)=0.16$, p=0.69</td>
</tr>
<tr>
<td>Gender (proportion male)</td>
<td>64%</td>
<td>58%</td>
<td>$\chi^2(1)=0.17$, p=0.68</td>
</tr>
<tr>
<td>Duration of consultancy yrs (s.d.)</td>
<td>10.9 (7.0)</td>
<td>8.0 (5.3)</td>
<td>t(47)=1.63, p=0.11</td>
</tr>
<tr>
<td>Estimated MiniCEXs assessed per year for any grade of trainee (s.d.)</td>
<td>13.1 (6.3)</td>
<td>19.1 (12.0)</td>
<td>t(47)=-2.21, p=0.03</td>
</tr>
<tr>
<td>Estimated MiniCEXs assessed per year for F1s (s.d.)</td>
<td>6.4 (4.1)</td>
<td>10.2 (9.7)</td>
<td>t(47)=-1.80, p=0.08</td>
</tr>
</tbody>
</table>

Table 2: Participants’ Confidence ratings by level and group

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Level of performance:</th>
<th>Mean score (95% C.I.s of mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Good</td>
<td>Borderline</td>
</tr>
<tr>
<td>A (Good to poor)</td>
<td>5.4</td>
<td>5.3</td>
</tr>
<tr>
<td>B (Poor to good)</td>
<td>5.8</td>
<td>5.6</td>
</tr>
<tr>
<td>Bonferroni p-value</td>
<td>0.276</td>
<td>0.969</td>
</tr>
</tbody>
</table>

Please note: values shaded in grey represent initial (unbiased) estimates
Figures

Figure 1: Schematic of experimental design

![Schematic of experimental design](image)

**G** = video of good performance  
**B** = video of borderline performance  
**P** = video of poor performance  
Numbers 1 & 2 indicate different performances within the same level of performance

Figure 2: Participants’ mean competence scores by level and group

![Bar chart of participants' mean competence scores](image)

* = significant at p <0.05  
Unbiased estimates at extreme left and right (Good, group A, Poor, group B)
Participants' "Percent Better" ratings by Level and Group

* = significant at p <0.05
§ = marginal significance, p <0.1
Unbiased estimates at extreme left and right (Good, group A, Poor, group B)
Chapter 6
General Discussion

Summary of findings

This thesis began with a consideration of the role and purpose of assessing authentic performance within medical education. Assessments of this nature have been shown to suffer from high inter-assessor score variability; further, attempts to ameliorate this problem through either training or alterations to response formats have not succeeded in adequately overcoming this issue. In response, the aim of our research program was to:

- develop an understanding of the judgement processes used by assessors as they reach judgments and allocate scores whilst undertaking clinical performance assessments

Within the programme of research described in this thesis we have demonstrated a number of novel findings that contribute to this aim. These are summarized below.

Differential Salience

By exploring assessors’ judgement processes we have shown that when different assessors judged the same performances, they viewed the relative saliency of different aspects of the performances differently; the aspects that one assessor viewed as important were not the same as the aspects of the same performance that a different assessor viewed as important. Some assessors emphasised a particular aspect greatly, whilst others didn’t mention it. In other instances assessors both mentioned the same aspect of the performance, but for one assessor it held great weight, whereas for another it was comparatively unimportant.
As a result, despite observing the same performances, assessors appeared to carry differing observations into the judgement process. The methodology we used did not allow us to distinguish between attention (or “noticing”) and subsequent weighting of observations. As a result assessors may have noticed more than they described. None the less, the information which they treated as salient to their decisions varied.

**Criterion Uncertainty**

Having attributed differing salience to different aspects of their observations, assessors then judged their observations in comparison to mentally held conceptions of competence. Assessors’ conceptions of competence (both their generic understanding of it, and the way they operationalised it in the context of the performance they were judging) varied. Moreover, their conceptions of competence were prone to uncertainty. Assessors frequently compensated for this uncertainty by making comparisons against the performance of other trainees, and also by relating their observations to what they believed to be typical of performance at that stage of training.

**Information Integration**

Assessors rarely expressed their evolving judgements in terms of the assessments’ rating scale or descriptors, but instead usually described their evolving judgements in narrative terms. Importantly these narrative descriptions differed between assessors, reflecting an individualised approach to expressing (and possibly thinking about) the degree of competence that they observed.

Within our initial exploratory study we also found some replication of prior work (Kogan et al. 2011; Govaerts et al. 2011) showing the importance of assessors’ inferences and the influence of contextual “non-performance factors” on the assessments: inter-personal interactions, the assessors’ prior opinions of a trainee, their awareness of the trainees’
appearance, the purpose of the assessment and assessors’ justifications for deliberate leniency.

**Idiosyncrasy within judgements**

Collectively the findings of the first study suggested a picture of idiosyncrasy in assessors’ judgements: areas of comparative individuality or uniqueness of focus, individuality in their criteria and individual ways of expressing the magnitude of responses, as well as the further influence of making comparisons against other trainees who will inevitably vary between assessors. Collectively these varied influences resulted in unique, different or individual judgement approaches: in effect, assessors saw the same thing differently.

**Influence of comparative judgements**

Having demonstrated these initial findings through an exploratory methodology, we then chose to focus on assessors’ use of comparisons within their judgements. From the psychology literature, we identified two potential effects that comparisons might have on assessors’ judgements: assimilative-anchoring effects and contrast-relativity effects. In assimilative-anchoring, current scores become unduly similar to prior anchors; in contrast-relativity effects they become unduly different. We tested this possibility in two studies: together they advance our theoretical understanding in number of ways.

**Influence of assessor characteristics on scores**

Similar to many of the findings by Kogan et al (2010), participants’ baseline characteristics had no relationship with their overall scoring tendency. We examined relationships between scores and: participants’ gender, their duration of consultant-grade experience, and the numbers of Mini-CEX assessments they estimated they conducted on doctors at F1 grade or at any grade. None of these showed significant correlations, although “duration of consultancy” showed a trend towards significance in study 3 (β range -.03 to -.05, p=0.076).
Whilst these beta coefficients appear small, they indicate that after 20 years experience as a consultant, an experienced individual might score as much as a whole point lower on the mini-CEX scale than a newly appointed consultant. This observation was only marginally statistically significant and occurred inconsistently across the two studies. As such it should be viewed as circumspect at this stage, and should ideally be replicated in more highly powered studies. None the less it is a potentially notable finding. Importantly, adjusting for these variables in the model did not alter any of the main findings.

Contrast effects on borderline performances

Recent experience of other performances influenced the scores that participants gave. Across two studies we repeatedly demonstrated contrast effects in assessors’ judgements: assessors who judged successively better performances gave higher scores than assessors who judged successively poorer performances. In study 2 we demonstrated this finding for borderline performances, showing a difference of 0.67 points on the assessments’ 6 point scale (Cohen’s d of 0.63). In study 3 we showed that this effect wasn’t confined to borderline performances but also occurred for good performances. The overall between-group difference on this occasion was 0.54 on the assessments 6 point scale (Cohen’s d of 0.52). Between group differences were insignificant for poor performances, although (as discussed in chapter 5) this finding may represent a type II error due to a floor effect.

Influence of assessors’ “Hawk / Dove” differences

In study 2 we examined assessors’ “Hawk / Dove” tendency: their tendency to score either consistently stringently or leniently. We initially showed that assessors had a moderately consistent tendency to be either Hawks or Doves (Hawk/Dove reliability of between 0.51 and 0.58). We then showed that whilst Hawk / Dove differences between assessors accounted for 18% of observed score variance, the influence of assessors’ recent experiences (whether they had just viewed good or poor videos) accounted for an additional
24% of observed score variance. As a result assessors’ recent experience accounted for a greater proportion of score variance than the competing hypothesis of fixed Hawk / Dove differences. Collectively the combination of Hawk / Dove differences and the influence of recent experience accounted for 50% of observed score variation. Consequently they explained a substantial proportion of observed variance, but further unexplained variation remained.

Perceived Rank of Performances

In contrast to our stated hypothesis, judgements of the perceived rank of performances neither showed a lesser susceptibility to the influence of recent experience, nor mitigated the influence of recent experience on competence scores. Instead these judgements also showed a contrast effect, accounting for an average between-group difference of 14% on the percentile ranking scale (Cohen’s d = 0.67). Consequently they appeared to be similarly susceptible to the influence of recent experience as competence scores. Again this effect was demonstrated for both good and borderline levels of performance, but this time differences at the poor level of performance were marginally significant. Of further note, variance in participants’ competence scores was substantially explained by variance in participants’ perceived rank ratings with $r^2$ values ranging from 50% to 88% by different methods.

Confidence in judgements

Participants’ confidence in their judgements did not vary by level of performance or by group. Overall, participants’ ratings of confidence in their judgements were high: median confidence was 6 on the 7 point scale. The interaction between group (the influence of recent experience) and level of confidence was not significant, and examination of “high” and “low” confidence groups indicated that there was no tendency for lower confidence individuals to be more susceptible to the manipulation. Importantly, therefore, the expected
inverse relationship between confidence and susceptibility to contrast effects was not found.

Theoretical implications of findings

Cognitive schemas and idiosyncrasy

Collectively these findings have a number of theoretical implications. We have described idiosyncratic differences in assessors’ focus, their use of criteria and the way they described the magnitude of their judgements. From this we can reasonably propose that assessors use different categorical systems (or schemas) whilst judging performances. At first glance, this assertion may appear to rely on an assumption, but it can be supported by a series of logical deductions from our data.

The literature reviewed in chapter 1 described that people use contextual cues to determine the salience of potential categories, and then activate or inhibit the categories that seem most relevant to the judgment task (Macrae et al. 1995). Participants in our first study all viewed the same performances and so viewed the same series of performance cues. Despite this they attributed different degrees of salience to different aspects of the performances. From this we may deduce that (as the series of cues was constant) participants’ sensitivity to cues was different. Consequently we may infer that either assessors possess different categorical schemas, or that they possess the same schemas, but their schemas are activated by different cues. Either way we can assert that assessors had different categorical schemas active during the judgement process. As a result the described idiosyncrasy in assessors’ approach appears to have arisen because different assessors had idiosyncratically-different categorical schemas active whilst judging.
Our suggestion that assessors use idiosyncratically different categorical schemas whilst judging is not without precedent. Within occupational psychology, demonstration exists that untrained individuals possess varied, unique cognitive schema which they apply when judging performance. These unique schemas are described as “idiosyncratic”, and can be shown to influence which behaviours are considered relevant and the accuracy of subsequent ratings. (Gorman & Rentsch 2009; Uggerslev & Sulsky 2008). Idiosyncrasy has also been noted within assessment in medical education. Ginsburg et al (2010) described assessors combining different domains of an assessment in idiosyncratically different ways depending on the assessment context. More recently (and subsequent to the publication of study 1), Govaerts et al (2012) showed that raters used differing performance schemas in idiosyncratically different combinations whilst judging performance.

This tendency to use idiosyncratically-different categories is important; once categories have been activated, they guide action (Fiske 1993). Category activation determines the information that is attended to, guides recall and recollection (Stangor & McMillan 1992) and enables inferences which may themselves distort memory (Tsujimoto et al. 1978). Consequently idiosyncratic differences in assessors’ categories or schemas may have been substantially responsible for the differences in the judgements that participants formed.

**Assessment criteria**

Further theoretical implications arise from our findings relating to assessors’ use of assessment criteria. Participants described their expectations of F1 doctors in various terms, or (put differently) participants described using differently-constructed assessment criteria. In that respect the content of assessors’ criteria differed; the behaviours in which they were interested varied between assessors. Participants described developing these expectations (or criteria) over time, through observation of numerous junior doctors. This latter observation is consistent with the findings of Govaerts et al (2011) who showed that more
experienced assessors have more complex mental representations than novice assessors. In addition to differences in content, assessors also constructed their judgement-criteria differently in terms of the magnitude or proficiency of the behaviours that they expected. We might term this the standard of their criteria. Findings from study 1 and study 3 suggest that assessors’ conceptions of standard may be (at least in part) normative: they are based on what assessors perceive to be typical for that stage of training, judging trainees’ competence by determining their position in a perceived distribution of performance at that stage of training.

A note of caution is necessary. Study 3 showed only an association rather than a causal link between perceived rank and competence scores. Therefore, this theoretical suggestion is currently speculative. Moreover, it may not be the only means by which assessors develop a sense of standard for their criteria; the study by Kogan et al (2010) suggested that assessors use their own clinical skills as a frame of reference whilst judging. There is no reason why assessors might not triangulate between their own practice and a mentally held distribution of performance of doctors at the relevant stage when determining the level of their assessment criteria. Moreover, this may explain why assessors have a moderately consistent tendency to be either “Hawks” or “Doves”; consistency in their own clinical approach may produce a consistent “hawk” or “dove” frame of reference whilst assessing. As a result, on the basis of our findings we may make a provisional suggestion that assessors’ criteria for performance assessments comprise idiosyncratically-different constructs to represent content together with a sense of standard that may be triangulated between self-reference, their perceptions of norm-referenced standards and potentially other as yet unexplored frames of reference. Notably this varied frame of reference is consistent with the work of Kogan et al (2011). It is important to note that our data do not preclude the co-existent use of additional categories which confer some absolute (rather than normative) sense of
standard. None the less, assessors’ assessment criteria appear to be complex entities that are not closely akin to the idea of criterion-referencing that is espoused in assessment theory (Wass et al. 2001).

Judgemental “fuzziness”

In addition to these theoretical developments, studies 2 and 3 add an important further concept: namely that assessors’ judgements could be contextually manipulated through the influence of recent experience.

Having observed this contrast effect, we suggest two ways in which it may have produced the observed change in scores. For illustration, let us consider the group in study 2 who initially saw poor performances, and who subsequently gave relatively inflated scores to borderline performances. This effect could have occurred for two reasons: either participants’ (assessors’) experiences of the performances themselves were relatively enhanced; or, alternatively, the experience of the performances was unchanged but the criteria against which performances were judged were reduced in standard.

Notably both of these explanations already exist within the literature on assimilation and contrast effects. “Selective accessibility” (Mussweiler 2003) is suggested to distort observation by causing individuals to preferentially notice behaviours similar to the prior performance. Conversely range-frequency theory (Parducci 1965) suggests that changes in context alter individuals’ perceptions of the distribution and range of items; in essence that the criteria are altered. Importantly selective accessibility is understood to produce assimilation effects whereas range-frequency theory is believed to produce contrast effects. If we accept the validity of this prior literature, it seems likely that the observed contrast effect was mediated by alteration to participants’ criteria.
How then were assessors’ criteria manipulated? Schwartz (1999) suggests that the less well an item is represented in memory, the more it is susceptible to contextual influences. Consequently it appears that the mental representations that assessors use to infer the standard of a performance (whether a distribution of trainees’ performance, the ability to relate observed performance to their own practice or categorical schema which represent an absolute sense of standard) are not represented clearly enough in memory to resist manipulation. Consequently the presence of a concrete example of performance is able to distort assessors’ pre-existing criteria, producing the observed contrast effect.

In their seminal review of research on cognitive categorisation, Mervis and Rosch (1981) concluded that cognitive categories, rather than having well defined or determined boundaries have “fuzzy” edges. In coining this phrase they indicated that “fuzzy” means the opposite of “well defined”. Understood this way, the categories that assessors use to determine the standard of performances appear to have a central “fuzziness” and are consequently susceptible to contextual manipulation, at least in terms of other standards of performance.

Consequently we see that assessors’ criteria were idiosyncratically different in content, that their representations of standard may have been influenced by both self reference and norm reference and that they were “fuzzy” in nature, enabling them to be distorted by concrete exemplars.

This “fuzziness” is perhaps surprising considering that in both studies 2 and 3, participants had a median duration of consultant experience of roughly 10 years. Moreover, from Govaerts et al’s (2011) work we may presume that participants’ assessment-schemas had, on average, developed some increased level of complexity compared with novice assessors.
Possibly, the context of the experiment produced a stronger contrast effect than might occur in real life; it would be useful to confirm the finding in a more ecologically valid setting. None the less, it appears that despite 10 years of experience, perceptions of competence can be manipulated by just two or three performances. A few explanations for this are possible.

Firstly, if we may temporarily assume that “standard” is indeed substantially influenced by norm-referenced perceptions, then one possible explanation becomes apparent: we may suggest that fuzziness arises due to the limits of human working memory. A distribution of performance by trainees at a particular level involves ordering the ability of numerous trainees; conversely human working memory can usually only actively process approximately seven items (Miller 1956). The capacity of working memory can be increased by grouping items into chunks (Baddeley 1994); for example an assessor might compare against a “chunked” distribution of trainee performance by considering whether a performance was “well above average”, “slightly above average”, “average”, “slightly below average” and “well below average”. None the less, deliberately inducing participants to consider typical performance by a cohort of trainees (through requesting the “Percent Better” rating in study 3) did not reduce the “fuzziness” of judgements. Perhaps assessors do not chunk in this way, or perhaps an alternative explanation of this observation is required.

Secondly, despite extensive experience of numerous trainees, assessors may have never observed a range of trainees perform the same task they are currently judging (i.e. they have not seen a range of trainees perform a consultation with the exact patient, presenting with the exact complaint that they are currently judging). Consequently, in order to make a norm-referenced judgement about a current performance, they must infer from a sense of the general ability of each trainee in their distribution to the specific case they are judging. Fuzziness may arise from this process of inference.
Finally, assessors’ experience of trainees over their careers is most likely non-systematic. Different trainees will have been experienced as part of different cohorts, with each cohort being progressively further removed in time from the assessor’s memory. Changes in practice, in education, and societal expectations may have contributed to real differences in trainees’ performances across different cohorts, as may movement between hospitals or healthcare settings by the assessor. Consequently their experience may not have sampled a single distribution of trainees, but instead lead to a mental distribution that is fragmented and incomplete, and therefore poorly defined or fuzzy.

Having established that assessors’ categories are represented weakly enough to enable contrast effects, we might reasonably presume (from a theoretical perspective) that they might also be influenced by other contextual influences – for example any pre-existing beliefs or stereotypes that are held by the assessor (Kunda & Spencer 2003). Moreover, such vulnerabilities are likely to be exacerbated by conditions that promote over-reliance on categorical thinking such as cognitive load (Dijksterhuis & Van Knippenberg 1995) or emotions such as anger (Bodenhausen et al. 1994) envy (Hill et al. 2011) or anxiety (Wilder 1989).

The final theoretical implication of these findings concerns the lack of an inverse relationship between participants’ decision-confidence and their susceptibility to contrast effects. As discussed earlier, this tends to imply a lack of insight by participants into their own vulnerability to manipulation. Consequently it appears that assessors are unaware of being manipulated at the time that it occurs. Therefore they may lack the capacity to make deliberate efforts to reduce the effect of contextual influences when they occur.
Limitations

In chapter 2 we discussed our choice of methodological strategy; the approach was initially exploratory, using a method that was comparatively subjective but broad in scope, followed by more robust experimental studies that tested specific aspects of the findings from the initial exploratory study. Consequently we triangulated between studies, producing consistent findings by different methods.

In all three studies, our investigative approach favoured internal validity over ecological validity. In the two experimental studies this was necessary in order to establish causality; in our initial exploratory study (study 1) we introduced control by using video stimuli which enabled us to compare between participant groups. This enabled us to attribute observed differences in approach to the assessors rather than to variations in case or setting. Consequently this methodological choice produced one of the study’s strengths: its explanatory power with respect to assessor differences. None the less all three studies were artificial to a degree, and generalisation of their results to real assessment practice currently requires some extrapolation. A number of aspects of the design of the studies helped to ensure their realism: the use of real assessors as our study population, the use of a real assessment format which was in current use by the study population, and the careful creation of video stimulus material which depicted plausible performances in a highly realistic environment. None the less participants in our studies were aware that they weren’t performing a real assessment and there was no inter-personal interaction between the assessors and trainees during studies. Consequently it would be beneficial as this research proceeds to supplement our findings with either naturalistic data or field observations.

In chapter 2 we discussed the role of subjective “Think-Aloud” data in determining cognitive processes, concluding that it gives a window into cognition that is useful for exploration but
which is inadequate for hypothesis testing. In study 1 we advanced several findings that helped to explain assessor variance; in studies 2 and 3 we sequentially supported a number of these findings relating to the influence of recent experience on assessors’ judgements, showing that they were replicable. None the less other findings from study 1 remain to be tested, and further explored. The findings in the “Differential Salience” theme should be tested, and more specifically investigated to determine if differences in salience arise from differences in attention by assessors or differences in weighting of what is noticed. Similarly the findings in “information integration” should also be tested, particularly to determine the extent to which score variations arise from judgmental differences between assessors, and how much they arise from differences in translating judgements onto assessment scales.

Whilst the above issue should be noted, we do not believe that it invalidates any of the claims that we have made in this thesis; the findings that we have described were rigorously established within the limits of the methods used. Rather, we are observing that the degree of supporting evidence for “differential salience” and “information integration” is currently less for than for the repeatedly demonstrated contrast effect of recent experience. Further this reduced support is not because any contradictory findings, but simply because we have not yet had an opportunity to test these findings through an alternative method. Consequently, as this research progresses it will be important to seek to either further support or falsify these findings.

Our research was conducted within a particular population, consultant physicians within the United Kingdom, and in the context of a particular assessment format, level of learners and type of clinical scenario: mini-CEX assessments on F1 (PGY1) doctors whilst they performed consultations with (simulated) patients who presented with new general internal medicine problems. As noted in each of the three studies, participants were not paid and were
recruited based on their personal willingness to participate. This may have introduced a selection bias as participants may have been unusually interested in either education or assessment. As previously discussed, if anything, this is anticipated to have made participants more similar to each than would be expected for general members of the study population. As a result, we might anticipate that without this possible bias, differences between assessors could have been greater than those observed. We believe that the way that we sampled from this population is a strength of the programme of research. Participants were recruited from geographically dispersed locations and a diverse range of clinical specialities and hospitals. Participants varied by gender and duration of their appointments as consultants. Consequently there is good reason to believe that the samples used in each study were a fair representation of the study population.

Findings from this population and context do not implicitly generalise to other populations of assessors or other assessment contexts. We are not, however, aware of any theoretical reason why the same issues might not be expected to recur in other populations or contexts; plausibly we might expect them to differ in their extent. None the less this assertion requires empirical verification.

**Remaining issues and future research**

**Replication and extension of findings**

Our aim with this programme of study was to explain variations in assessors’ scores whilst they undertook performance assessments. We believe that the findings and theory presented here represent a significant novel contribution to understanding this topic. None the less (consistent with the early stage of development of this field) the literature reviewed and the research conducted have suggested an array of further topics which warrant further investigation.
A few issues have already been mentioned. It would be useful to test the findings from study 1’s “Differential Salience” and “Information Processing” themes, by a more objective method. In particular, as described, this research might determine whether differences in salience arise due to differences in attention, or difference in weighting of observations. Investigation into the “information processing” theme might determine the degree to which assessors’ score differences arise due to judgemental differences between assessors, and how much due to variations in translation of their judgements onto assessment scales.

It would be useful to replicate these findings in other contexts. In contrast to the situation for consultation skills, generalisability D-studies for assessments of procedural skills have indicated the need for as few as 3 assessments to achieve a G-coefficient of 0.8 (Beard et al. 2011; Wilkinson et al. 2008). As a result we may wonder whether assessors’ judgement processes are more consistent for procedural assessments. Consequently seeking replication of our findings for assessments of procedural skills is important. Additionally seeking replication of our finding for other clinical specialities and for trainees at other levels of training is important to establish the generalisability of these findings.

Beyond these issues of skill-type, speciality and trainee-level it is also important to determine whether the findings that we have described occur for different formats of assessment. In chapter 1 we described our focus on workplace based (or performance) assessments, but other assessment formats within medial education also rely on assessor judgements. Objective structured clinical examinations (OSCEs) are a more standardised format of assessment that are often used for high stakes summative assessment. In OSCEs, candidates rotate through the same series of clinical simulations (“stations”) and are judged
by examiners who use standard checklists to guide their observations and scores (Newble 2004).

Whilst the original intention of OSCEs was to provide an “objective” view of candidates’ performance (Harden et al. 1975), it is clear that their reliability is imperfect (Brannick et al. 2011; Cohen et al. 1997), being importantly influenced by inter-examiner variability (Swanson et al. 1995; Pell et al. 2008). Item checklists seek to standardise assessors’ observations (Newble 2004), and so it is possible that examiners’ judgements are consequently neither idiosyncratic nor fuzzy. None the less, as (in view of their imperfect reliability) checklists must logically fail to completely standardise observations, it is likely that the issues we have demonstrated for performance assessments will similarly occur in OSCEs. Determining whether this is indeed the case, and the comparative extent of the effects across formats would also be useful.

As discussed in the limitations section of this chapter, demonstration of the effects we have observed in a more authentic setting would be useful. This could utilise field observations by researchers, making comparisons between assessors. To enable comparisons, it would be helpful if assessors were examining the same clinical cases, so it may be more feasible to conduct this research using OSCEs than using workplace based assessments. Alternatively, a quantitative approach similar to that used by Damisch et al (2006) might examine whether numerical relationships (positive or negative) exist between successive candidates.

Additionally it would be useful to look at other variables that might mediate the contrast effect that we observed; it would be useful to determine for how long the effect persists, or (put differently) whether the effect occurs if the target is encountered sometime after the initial anchor, for example after a day or a week. Both Murphy et al (1985) and Becker and
Miller (2002) showed that contrast effects could be achieved with a 24 hour delay between initial anchor and target.

It would be equally useful to understand whether there is a mediating effect due to change in person between anchor and target, or a change in clinical case. It is also possible that a cross-modality influence might occur, for example: does observation of good performance on a skill influence subsequent judgements of intermediate performance in a communication task? Whilst this might seem counterintuitive it would be consistent with the cross-modality assimilation effect demonstrated by Oppenheimer et al (2008).

In studies 2 and 3 we demonstrated that the influence of recent experience influenced assessors’ scores; it would be useful to understand whether recent experience additionally influences assessors’ observations, for example whether the content of the prior performances influence what assessors subsequently notice or comment upon in later performances. From a theoretical perspective this might suggest that the “Selective Accessibility” which has been demonstrated in assimilation effects (Mussweiler 2003) also mediates contrast effects. From an educational perspective this would indicate that recent experience not only influences trainees’ scores but would have implications for the formative message that is provided to trainees through assessment feedback.

Finally, our findings repeatedly demonstrated contrast effects; this does not preclude there being circumstances within medical education when an assimilation effect might occur. Investigation of the parameters that switch contrast effects to assimilation effects would further our understanding of this field. Assimilation is known to be favoured by simultaneous observation of anchor and target (with delayed rating of both) (Tanner 2008). Consequently assessment of learners’ contributions to a tutorial setting (for example) may be prone to
assimilation rather than contrast. Equally there may be other variables capable of inducing this switch.

**Social and interpersonal influences on judgement**

Thus far we have considered replication, confirmation and direct extensions of the work presented in this thesis. Beyond this there are numerous other related issues that our work and the associated literature suggest may be useful to study. As a feature of our methodological approach, we deliberately excluded the influence of interpersonal factors; our participants had no interaction with the F1 doctors’ whose performances they judged. This approach was useful as it produced methodological control, allowing us to isolate relevant aspects of cognition. None the less there is good reason to believe that social or interpersonal issues might comprise another layer of factors which all have the potential to influence judgements (Govaerts et al. 2007; Fletcher 2001; Ilgen et al. 1993), and could even interact with the findings we have so far described.

The additional results we presented in chapter 3, and the findings of Kogan et al (2011) both indicated that assessors are aware of the influence of social pressures whilst judging: the presence of a pre-existing relationship between the trainee and the assessor, not wishing to damage the trainee’s confidence, apprehension about the trainee’s response to negative feedback and various other reasons for leniency have all been described. Additional to these, literature reports of performance appraisal in occupational psychology have described the influence of: inter-personal affect or liking (Lefkowitz 2000), impression management strategies (Wayne & Liden 1995) and group-membership (within “leader-member exchange”) (Duarte et al. 1994) on performance appraisal judgments. Not only is it relevant to investigate how these factors influence assessors’ judgements in medical education, but also further investigation could determine whether interactions exist between these influences. For example, might contrast effects be either increased or decreased based on
the leader-member exchange status of a trainee, or conversely might a trainee’s “in” or “out” group status determine whether an assessor assimilates or contrasts their performance with that of another trainee?

A number of other assessor-related factors including assessors’ emotional state (Bodenhausen et al. 1994; Hill et al. 2011), their degree of cognitive load (Bodenhausen 1990), their level of motivation (Neuberg & Fiske 1987), or more fixed traits such as their need for cognition (Crawford & Skowronska 1998) all have the potential to influence assessment judgements or to interact with judgement processes (such as contrast or assimilation effects). Consequently investigation into the influence of these factors would also be beneficial.

As well as excluding interpersonal interaction, our methodology held a number of other variables constant between performances. The clothing, gender and ethnicity of the F1 doctors were kept constant in each video to avoid the potential for confounding influences of these variables on assessors’ scores. The literature we reviewed in chapter 1 indicated that such variables could have an important influence on judgement processes, namely through stereotype priming. If, for example, aspects of the trainees’ gender, appearance, or ethnicity were to prime particular stereotypes (organised, knowledgeable, chaotic, poor communicator) then we might expect, based on the stereotyping literature, that these activated stereotypes might influence assessors’ observations and cause them to believe they had observed behaviours which had not in fact occurred (Feldman 1981; Srull & Wyer 1989). Whilst this might influence scores, it could also influence the content of educational feedback. Consequently the effect of these factors on assessors’ judgements also merits investigation.
Ginsburg *et al.* (2008) showed that attending faculty members’ concepts of professionalism vary between contexts; for example they espoused that lying is wrong, but admitted that in some circumstances students should realise the need to withhold the truth. We have already demonstrated variability *between* assessors in their conceptions of competence (see also [Thammasitboon et al. 2008]). It would be useful to determine whether conceptions of competence (and therefore the assessment criteria) vary *within* assessors due to changes in the clinical context. For example, might clinical assessors espouse that exploring a patient’s care-related expectations is part of conducting an excellent patient-centred consultation, but actually expect that trainees should dispense with such “non-essential” features as soon as the list of waiting patients starts to lengthen? If this were so, it would have important implications for the ability to define and specify competence.

**Organisational culture in which assessments occur**

Finally, it is important to note that assessments do not occur between a trainee and an assessor in a vacuum; the assessment occurs within a healthcare or educational organisation, in which there is an institutional and indeed national culture. This culture is determined in part by policies set by the organisational administration, and in part by the beliefs and attitudes of individuals or groups within the organisation. A number of reviews within the performance appraisal literature have considered the influence of organisational culture on performance appraisal ([Ilgen et al. 1993; Fletcher 2001; Levy & Williams 2004; DeNisi & Pritchard 2006]). These highlight the importance of the managerial culture within the organisation, the external economic climate, human resource strategies, the degree of “power-difference” between managers and subordinates and the degree to which the organisation espouses a continuing learning culture ([Fletcher 2001; Levy & Williams 2004]). Other work has suggested that perceptions of fairness, accuracy and adherence to due process within the appraisal system mediate the ways that staff react to feedback, which is,
in turn, an important determinant of whether they use the feedback to develop or simply ignore it (Jawahar 2010).

In chapter 1 we described a number of studies that have explored trainees’ satisfaction with workplace based assessments, and their perceptions of its educational effectiveness. As described, perceptions have been mixed, but generally supportive of performance assessments. Overall, however, there has been comparatively little investigation into the influence of organisational culture on performance assessments within medical education. Moreover, the study we described in chapter 1 by Daelmans et al (2006) indicated that a range of organisational issues can create a culture which makes assessments both difficult to conduct and ineffective. More recently Bindal et al (2011) indicated that a proportion of trainees view assessments as a “box ticking” exercise, with little educational value. These authors concluded that a cultural change is required to aid assessments success. This conclusion resonates with the findings of a recent large scale review of Foundation training in the UK (the “Collins report” (Collins 2010)) which queried whether workplace-based assessments have the support of either trainees or trainers, citing concerns around validity and variations in assessment conduct. Whilst these concerns may have very valid foundations, their expression in high-level policy could further impact the culture in which assessments are conducted, for example by legitimising cynicism or poor practice.

The aim of this programme of research was to understand assessors’ judgement processes within clinical performance assessments. Consequently studying the assessments’ cultural context seems beyond the intended scope of our research. None the less, the degree to which assessments are culturally accepted may be enhanced if the perceived fairness or uniformity of assessors’ judgements can be improved. Equally, organisational culture may have a bearing on whether assessors are sufficiently engaged with their role to be able to
make optimal judgements. As a result the two may be inextricably linked. It therefore seems reasonable to suggest that as research on assessment judgements progresses, further investigation should address the cultural meaning of assessments and the attitudes and reactions that users hold towards them.

**Implications for practice**

**Idiosyncrasy versus error, and different theoretical perspectives**

The findings of this programme of study have a number of implications for practice. Classical test theory (Streiner & Norman 2008, p170) views inter-assessor variations as error which reduces the reliability of the measurement of a singular construct. From this perspective it makes logical sense to calculate an average of assessors’ scores, as this reduces overall error and approximates candidates’ true scores. Having demonstrated that assessors’ judgements are idiosyncratically different we may challenge this assumption, suggesting instead that assessors are all “measuring” different constructs. Calculating an average of their scores will therefore tend to create a reductionist view of a complex issue which may be less rather than more “accurate”. This assertion is consistent with the social-constructivist position of Govaerts *et al* (2007) who plead for assessors’ Idiosyncrasies to be viewed as a plurality of true scores rather than faulty measurements. Whilst their position clearly aligns with our current understanding of the phenomenon, no-one has yet suggested a manageable means by which to amalgamate assessment results in this way.

From a social constructivist perspective perhaps the most authentic approach would be simply to present candidates and administrators with all results relating to each candidate without any form of averaging or compilation. Unfortunately this is unlikely to be workable for summative purposes, as (if all scores are equally valid) how would pass / fail decisions be made when assessors disagreed?
From a formative perspective, presenting trainees and their educational supervisors with all assessment data may be a feasible approach. Up until August 2012 trainees within the United Kingdom Foundation Programme (UKFPO) kept the scores from all workplace based assessments in their training portfolios. Variations in trainees’ scores across the course of a year were generally attributed to case specificity (FP 2007); they were seen as indicative of the trainees’ performance. This view may need to be challenged if we are to encourage trainees and their educational supervisors to triangulate between multiple assessors’ perspectives.

The richness of different idiosyncratic views may be more usefully expressed in free text or verbal feedback than by scores (Brutus 2010). The UKFPO have recently abandoned scores on their version of the Mini-CEX form (UKFPO 2012), instead only asking for free text comments. This is hoped to increase the provision of constructive feedback. Whilst this approach may be beneficial, it is too early to judge its success. Importantly it will depend on the degree to which assessors provide rich feedback rather than bland or brief comments. Brutus (2010) suggested that providing narrative comments will require greater cognitive effort by assessors than providing numerical ratings. Consequently success of this approach will require adequate motivation by assessors.

**Criterion or norm referencing of cut scores**

Within the medical education community, criterion referencing is recommended as best practice, with its inherent appeal to fairness and to patient safety by ensuring that trainees are measured against a fixed standard of competence rather than against each other (Wass et al. 2001). Whilst this idea is attractive our findings challenge its viability for assessments that rely on assessors’ judgements. Assessors, due to their inherently fuzzy criteria, may be unable to avoid comparing trainees with each other whilst judging their performance.
Moreover, assessors’ mental representations of the “standard” of assessment criteria may depend on normative comparisons (they are at least clearly associated). Consequently “competence” judgements may be substantially norm- rather than criterion- referenced.

If true, this inability to make criterion-referenced judgements may be particularly important for the borderline regression technique of creating assessment cut-scores. In this approach linear regression is used to relate assessors’ global judgements to their checklist scores for each station, and station pass marks are subsequently calculated by interpolation (Downing et al. 2006; Pell et al. 2010). If the assessors’ global judgements are substantially influenced by norm-referenced comparisons, then the resulting checklist cut-scores will be fully or partly the result of a norm-referencing. Consequently the borderline regression technique may create an unjustified impression of objectivity in the cut scores that it produces.

Variations in competence-criteria between hospital settings

If assessors’ performance judgements are both comparative and norm-referenced, then their criteria will be influenced by the quality of the cohort of trainees that they assess. Hospitals in the UK recruit F1 doctors on a competitive basis. For a variety of reasons hospitals differ in the numbers of applicants they receive for each available position: due to geography, the range of available clinical training opportunities, variations in job plans and differences in perceived prestige. Consequently some hospitals are heavily over-subscribed whilst others struggle to recruit. If we assume that the laws of supply and demand operate, then over-subscribed hospitals will generally recruit higher calibre trainees than under-subscribed hospitals.

In this scenario, assessors in different hospitals may be exposed to very different calibres of trainees over successive years. Consequently, assessors’ mentally-held distributions of “typical” trainee performance may be somewhat higher in standard in over-subscribed
hospitals than in under-subscribed hospitals. As a result, competence decisions may similarly vary. Importantly, this may tend to create undue leniency for the group of trainees most likely to have deficiencies: by virtue of their low competitiveness, they find work in a hospital where assessors’ conceptions of competence may also be lower. Equally, moderately able trainees who find themselves at the lower end of a talented cohort within a prestigious hospital may be led to believe their performance is deficient when it is not. Clearly both processes have the potential to be damaging. Similar processes could occur within any hospital where small numbers of trainees are consistently paired or grouped, so that any given individual tends to be consistently compared against a small number of other trainees. Notably this assertion relies on a logical extension of the theory that we have developed, and makes some assumptions; it should be tested in due course.

Potential implications for OSCE exams

As stated earlier, we cannot necessarily assume that the contrast effect which we observed in these studies will also occur in OSCE exams; the use of structured checklists may mitigate the effect. None the less, if it does occur, it may have a potentially serious influence. Hawk / dove differences between examiners in OSCEs are well recognised (Swanson et al. 1995). One of the advantages of the OSCE design is that these fixed differences tend to counter-balance as students progress around the OSCE circuit. Moreover (within a given OSCE circuit) all students are equally exposed to different examiners and so (unless there are particularly large examiner-by-student interactions) the process tends to be fair.

Conversely, the contrast effect that we have demonstrated has no innate counter-balancing in OSCE exams; students typically progress around the circuit in a fixed order, with each student following the same student throughout the exam. Consequently, the scores of a student who follows a colleague who is either very good or very poor may be unduly inflated or deflated in every OSCE station. Standard OSCE design will not counterbalance this effect,
and as a result the student may receive undeserved scores on every station. Moreover, under current psychometric procedures, this effect will appear to be due to student variance and will be unnoticed. Notably, our findings occurred after assessors were exposed to a number of performances that were either consistently good or poor (study 2), or which consistently either increased or decreased in the standard they portrayed (study 3). We don’t yet know whether contrast effects will occur if assessors experience a mixture of performances. None the less, this is a potentially highly important implication for assessment practice.

Implications for improving uniformity of assessors’ judgements

At the outset, we hoped to develop our understanding of assessors’ judgements with a view to recommending theoretically based interventions that might reduce their variability. As this discussion has highlighted, there is much yet to be done before our theory can be considered complete. However, a few initial suggestions can be offered.

Perhaps the most important insight from our work, and the other studies we have summarised, is to begin to appreciate the sheer complexity of assessors’ judgment processes. In view of this complexity, it seems unlikely further modifications to assessment mark sheets will achieve substantial gains. This is consistent with most studies in medical education (Cook & Beckman 2009; Donato et al. 2008) and with the views expressed in the performance appraisal literature (Ilgen et al. 1993; Landy & Farr 1980). The paper by Crossley et al (2011) that showed an improved reliability with their “construct aligned” scale seems at odds with our findings. From the theoretical perspective we have developed, there is little reason to suggest that assessors’ constructs of either entrustability or progressive independence will be any more uniform than their conceptions of competence. In chapter 1 we discussed some methodological issues which could potentially account for Crossley et al’s findings. As such we recommend that (especially in view of the counter theoretical
implications of their findings) further exploration into how the construct aligned scale achieved increased reliability should be undertaken.

The complexity of assessors’ judgement processes also has implications for assessor training. Our theory may be too elementary to explain why assessor training (particularly frame of reference training) was effective in occupational psychology (Woehr & Huffcutt 1994) but has not shown an effect of the desired magnitude in medical education (Cook et al. 2009; Holmboe et al. 2004a). It is appreciable, however, why it might be difficult to effectively train assessors who have built up idiosyncratically different cognitive representations of the problem over long experience, but who despite that long experience are still fuzzy in their approach. This combination of difference and fuzziness may make it very difficult to reach a reproducible shared understanding. Perhaps the most encouraging feature of our work with regards to training is the demonstration that assessors’ judgements can be influenced by small number of exemplar performances. Potentially, therefore more effective training may be achieved through an exemplar-based training method that creates a common pool of reference performances. Of course, this would require further understanding of the case specificity of exemplar performances, and the duration over which any effect persists. As a result much further investigation is required into this possibility before it can be asserted with degree of certainty.

On a more immediately practical note, if subsequent investigation confirms that contrast effects appear to occur in OSCEs, then it may be prudent to devise procedures to alter the order in which students progress through stations. Thus, rather than student “x” always preceding student “y” as they progress around stations, students might be allocated stations in a more complex order that seeks to counter-balance student order across the
examination. Unfortunately this is likely to make the logistics of running OSCEs more difficult than they already are.

Conclusions

In this thesis we have sought to develop an understanding of the processes that lead to variations in assessors’ judgements whilst undertaking performance assessments in medical education. Through a series of three studies, and the integration of their findings with existing theory, we have provided empirical support for the theoretical advancement that we propose: namely that a substantial portion of variation in assessors’ judgements can be explained by “fuzzy idiosyncrasy”.

Fuzzy idiosyncrasy involves a series of processes: assessors appear to use different mental representations (cognitive schema) whilst judging, focusing on comparatively unique combinations of observations and comparing observations against individually constructed notions of competence. These notions of competence may find their sense of standard (in part) through reference to the assessors’ perceptions of a distribution of typical performance by trainees at that stage of training, although other factors such as self-reference may also influence these criteria. Having reached judgements, assessors express their judgements in unique individual terms which then require translation to be expressed as scores. All of these processes contribute to judgements being unique or idiosyncratic. Despite extensive experience assessors’ idiosyncratic judgements are based on mental representations that are imprecise – they are “fuzzy”. As a result they can be contextually manipulated through recent experience of exemplar performances of different levels of proficiency. Consequently, due to their fuzzy imprecision, assessors’ already idiosyncratic judgements are most likely further influenced through the haphazard experience of other
trainees. Moreover, assessors appear to be unaware of these influences on their
judgements.

Having advanced this explanation for variations in assessors’ judgements, we note that it is
both partial and provisional. These findings represent a beginning not an end to the study of
assessors’ judgements, and considerable further research is suggested by our findings and
theoretical developments. None the less our work has numerous implications for practice:
our findings challenge the notion of criterion-referencing in judgement-based assessments;
they help to challenge the existing theoretical basis on which assessments are based (in
particular the perspective provided by classical test theory); they have implications for the
fairness of exam systems and the comparability of performance assessments across different
healthcare settings; and lastly they provide initial theory from which to consider attempts to
improve the uniformity of assessors’ judgements.

As a result we assert that our findings provide a useful initial insight into assessor
judgements in performance assessments in medical education. We hope that from this
insight we can develop theory which may lead to important developments in assessment
practice. At the outset we focused on workplace-based or performance assessments,
describing their uniquely important role within medical education to guide trainees’
development through the provision of feedback on their real performance. Our findings may
turn out to have broader implications that are applicable to other forms of assessment.
None the less it is our aspiration that the theoretical insight we have developed will
contribute to making workplace based assessments more effective in fulfilling their intended
educational role.
References


161


163


Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2), pp.81–97.


172


Appendices

Appendix 1: Search terms used for literature review

Search terms used for direct observation performance assessments:

Appendix 2: MiniCEX mark sheet from the UK foundation Program

### 4.4.a Mini-Clinical Evaluation Exercise (CEX)

Please complete the questions using a cross: X  Please use black ink and CAPITAL LETTERS

<table>
<thead>
<tr>
<th>Doctor’s Surname</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Forename</td>
<td></td>
</tr>
</tbody>
</table>

**GMC NUMBER MUST BE COMPLETED**

<table>
<thead>
<tr>
<th>GMC Number:</th>
<th></th>
</tr>
</thead>
</table>

**Clinical setting:**

- A&E
- OPD
- In-patient
- Acute Admission
- GP Surgery

**Clinical problem category:**

- Trauma/Ortho
- Resp
- CVS
- Gastro
- Neuro
- Psych/Behav
- Other

**New or FU:**

- New
- FU

**Focus of clinical encounter:**

- History
- Diagnosis
- Management
- Explanation

**Number of times patient seen before by trainer:**

- 1-4
- 5-9
- >9

**Complexity of case:**

- Low
- Average
- High

**Assessor’s position:**

- Consultant
- SASG
- SpR
- GP

**Number of previous mini-CEXs observed by assessor with any trainee:**

- 0
- 1
- 2
- 3
- 4
- 5-9
- >9

Please grade the following areas using the scale below:

<table>
<thead>
<tr>
<th>Below expectations for F1 completion</th>
<th>Borderline for F1 completion</th>
<th>Meets expectations for F1 completion</th>
<th>Above expectations for F1 completion</th>
<th>U/C*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 History Taking</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2 Physical Examination Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Communication Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Clinical judgement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Professionalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Organisation/Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Overall clinical care</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*U/C Please mark this if you have not observed the behaviour and therefore feel unable to comment.

**Anything especially good?**

**Suggestions for development**

**Agreed action:**

**Trainee satisfaction with mini-CEX**

- Not at all
- Slightly
- Fairly
- Mostly
- Highly

**Assessor satisfaction with mini-CEX**

- Not at all
- Slightly
- Fairly
- Mostly
- Highly

**What training have you had in the use of this assessment tool?**

- Face-to-Face
- Have Read Guidelines
- Web/CD rom

**Assessor’s Signature:**

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
</table>

**Assessor’s Surname:**

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
</table>

**Assessor’s GMC Number:**

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
</table>

**Acknowledgements:** Adapted with permission from American Board of Internal Medicine

**Please note:** Failure of return of all completed forms to your administrator is a probity issue

8781.358628

178
Appendix 3: Characteristics of “Good” and “Poor” performance, and qualitative performance descriptors used in scripting cases

Contrasting features

**Poor Performance**
- Limited or absent introduction
- Leaves purpose of consultation unclear
- Does not develop rapport well
- No / poor use of open questions
- Ignores / dismisses patient’s cues
- Unfocused / incoherent use of closed questions
- Inadequately characterises relevant features of condition
- Uses leading questions
- Inaccurately summarises
- Doesn’t adequately check accuracy with patient / allow them to correct
- Misses important RFs / red flags
- Partial, incomplete, or inappropriate PMH / FH / DH / Allergies
- Incomplete or inappropriate SH.
- Incomplete or inappropriate poorly focused ROS
- Little or no consideration of impact on patient’s functioning
- Dismissive of / insensitive to ICE
- Not empathic or sensitive
- Poor regard for patient privacy or dignity
- Makes poor / limited diagnostic inferences
- Poor or no discussion of diagnosis
- Insensitive to the patient’s desire for information
- Does not discuss or negotiate plan.

**Good Performance**
- Introduces themselves politely / clearly
- Clear opening to consultation
- Develops a rapport
- Open questions to begin interview
- Responds to patients cues
- Focuses with closed questions following patient’s description
- Chooses important +ve / -ve symptoms to attend to
- Clearly characterises important features
- Summarises part way and checks patient’s agreement
- Addresses relevant RFs / red flags
- Covers PMH / FH / DH / Allergies
- Relevant, appropriate SH
- Relevant ROS
- Considers the impact of patient’s life / functioning
- Elicits and discusses ICE
- Empathic / sensitive
- Makes good diagnostic inferences
- Discusses relevant diagnoses (sensitive to patient’s desire for info)
- Discusses and negotiates plan

The choice of relevant symptoms, risk factors, red flags and other details pertinent to a case can be developed on a case by case basis. This may be informed by clinical guidelines, textbooks or clinical experience.

Definition of “Good Performance”
A history that contains most of the points described in the “good performance category”. A “good performance” history is not expected to be perfect. A few items may be missed, or within individual features (for example “chooses important symptoms to attend to”) there may be small omissions. The overall impression must be that the history is thorough, whilst focused, and contains many of the elements listed above.

Definition of “Poor Performance”
A history that contains many of the points contained in the “poor performance” category. The history will be substantially flawed, and would raise significant concerns in many ways.
Not every feature of “poor performance” needs to be present, but the combination of poor features must combine to give a sense that the degree of thoroughness, focus, safety and concern for the patient are poor and that the nature of the presentation has been poorly characterised.

Definition of “Borderline Performance”
A history that contains a number of elements of poor performance, and is as such flawed in ways that are important. The history may have some elements that are good, but these stand in contrast to some important flaws. Alternatively, it may display mediocrity across a wider number of features which combine to constitute a flawed clinical encounter. Either way, the performance is below satisfactory, but only to a limited degree.
Appendix 4: Example scripts for performances

Script case 1 good

FY: Good morning, my name is Dr ____. I’m one of the junior doctors here and I’ve just come to talk to you about why you’ve come in. Is that all right.

SP: yes, that’s fine.

FY: Would you mind if I checked your name and age to start with?

SP: I’m Julia Wallace, I’m 54

FY: Thank you. Now, your GP has sent you up here to the medical admission unit because you’ve been having some pain in your chest – is that right?

SP: yes it is.

FY: Could you tell me about that please

SP: Well, it’s on the left here. It is quite sore really. It’s sort of worse when I take a deep breathe. I thought at first I’d just lifted something awkwardly in the garden, but it didn’t go away

FY: what were you doing in the garden?

SP: I was gardening. I had a big bag of compost, and I was lifting it around the garden for most of the day.

FY: was that when the pain started?

SP: No, it wasn’t straight away. It was the next day sometime. About 24 hours ago now.

FY: so how did it start?

SP: I can’t tell you exactly how it started. It was just sort of there.

FY: Could you say if it started suddenly or gradually?

SP: I’m not sure. I suppose it crept up on me.

FY: Has it been there constantly since it started, or has it come and gone?

SP: It’s been pretty constant. It’s got a bit worse if anything.

FY: You said that it is quite sore. How bad is it?

SP: quite bad. Not agony, but uncomfortable. It’s quite sharp. It stops me wanting to take a deep breath.
FY: Quite sharp, and worse when you breathe in... Is it all in the same place, or can you feel it anywhere else?

SP: No, just here.

FY: so, you’ve never felt it in your arm or your jaw?

SP: no.

FY: Does it hurt when you move your arm, or twist. Can you get dressed alright?

SP: No, those things are ok

FY: Have you felt breathless?

SP: Um... a bit. Walking up the stairs. I don’t normally feel breathless going up stairs.

FY: how breathless were you?

SP: By half way up I felt a bit puffed, and that made it hurt more. By the top I stood and held the banister to catch my breath for a few moments.

FY: and that’s not normal for you?

SP: No. it’s not. I thought “that’s a bit odd”

FY: were you breathless any other time?

SP: no, just on the stairs. I suppose after that I avoided things that might make me breathless.

FY: have you been up the stairs since.

SP: yes, today. Before I went to the GP. It felt the same.

FY: Have you been coughing, or wheezy. Have you coughed anything up?

SP: No.

FY: not coughed up any blood?

SP: No

FY: or felt feverish or unwell?

SP: No.

FY: So, for about 24 hours you’ve had a sharp pain on the left, that is uncomfortable and is worse when you breathe in. It doesn’t go anywhere else. You’re not sure how it started, although you had been lifting the day before. You’ve also felt a bit breathless on the stairs which isn’t usual for you. You’ve not been wheezy or coughing or unwell.
SP: yes, that’s about right.

FY: have you ever had anything like this before?

SP: No, I haven’t

FY: Now, this might seem unrelated, but have either of your legs swelled recently?

SP: No, they haven’t. I had that years ago, though, when I was pregnant.

FY: What happened?

SP: I got a blood clot in this leg, and it swelled up.

FY: so, you’ve had a blood clot in the past?

SP: yes, I have. They said it was because I was pregnant. I was on tablets for a few months, and it went away. I never had it again.

FY: Can I ask you a few other questions?

SP: Yes, that’s fine

FY: Have been on any flights or long journeys recently?

SP: No

FY: Have you had any operations in the last few months? Or had your leg in plaster?

SP: No

FY: Do you have varicose veins?

SP: (Smiles) No

FY: Have you ever had a tumour or a cancer of any sort?

SP: (looks worried) No I haven’t

FY: Sorry, that’s just a routine question... Have I upset you?

SP: It’s just... the doctor said I might have blood clots in my lungs. And.. they’re what my aunty had when she had cancer...

FY: I’m sorry. Is that a possibility that has been worrying you?

SP: I... I don’t know. Yes. I suppose I’m worried.

FY: You might turn out to have a blood clot in your lungs. It’s possible. Equally you may not. I don’t know yet. But there are lots of reasons why people get blood clots. Cancer increases
the chance of having one, but there are lots of other reasons too. If we do find a blood clot, it doesn’t need to mean that you have cancer.

SP: ok, that’s reassuring. ok

FY: Are you alright to continue, or would you like to ask more?

SP: No, that’s ok, keep going.

FY: Do you have any further problems with your health, either now or in the past.

SP: High blood pressure. That’s all

FY: Do you take any medications.

SP: Yes, um, bendroflumethiazide. One tablet each day

FY: Is it 2 and a half milligrams

SP: I think so.

FY: we can check form your GP. Are you allergic to any medications?

SP: No

FY: Does anything run in the family?

SP: I’ve mentioned my Aunt. My Brother had a heart attack, but he’s alright now. My parents are both still alive and well.

FY: Do you smoke or drink?

SH: I drink a glass of wine at the weekend. I smoke about a packet a day – I have done since I was 20.
FY: Have you ever thought of stopping?

SP: I’ve tried the patches in the past. I keep meaning to have another go. I suppose that’s why the thought of cancer plays on my mind.

FY: perhaps you could see the smoking cessation nurse at your GP’s surgery

SP: Yes. Perhaps it’s time to think about that again.

FY: do you have a partner, or other family

SP: Yes, I’m married. My husband is in pretty good health. And two kids. They’ve both grown up and left home now.

FY: Do you work?

SP: Yes, I’m an administrator in a school
FY: can I just check for any other symptoms you might have?

SP: Yes, that’s fine

FY: Do you get any palpitations? Do you get any other chest pains

SP: No

FY: Do you have a regular cough? Do you wheeze or bring up phlegm?

SP: No

FY: Do you get tummy pains, or bloating, or heartburn?

SP: No

FY: Has there been any change in your bowels – diarrhoea, constipation, bleeding?

SP: No

FY: Any bladder or gynae problems?

SP: No

FY: Headaches, fits, blackouts?

SP: No

FY: do you get depressed or anxious?

SP: No, I don’t

FY: nothing extra I’ve not asked about?

SP: No.

FY: Good. Would you like me to discuss my thoughts with you?

SP: Yes I would.

FY: At the moment, it is too early for me to say for definite what this is. Some of the features that you’ve described might point to a blood clot in your lungs. Some of them don’t. It could, like you said, be something simple like a pulled muscle in the side of your chest. I think that we might need to do some tests to look further. To start with that will just be some simple blood tests, and a heart tracing and an x-ray – as long as that is alright with you. I’ll have a look over you as well if I may. Depending on what those show, we might need to look a bit further, which would be with a scan, but we’ll decide on that later.

SP: and if it is a blood clot?

FY: They are very treatable – you would need treatment with blood thinning tablets for 6 months
SP: Like when I had the clot in my leg

FY: that’s right. If you have a blood clot, it won’t necessarily mean that there is anything else going on. Some people just get them – especially if they’ve had them before. But if anything does seem out of line, we will look into it and we’ll discuss that all with you at the time. Does that seem like a reasonable plan to you?

SP: yes it does

FY: can I help with anything else

SP: No, that’s fine.

FY: Ok. I’ll come back and examine you in a moment. I’ll book those tests, and you’ll be reviewed by the consultant later.

SP: Thank you

FY: Thank you

CONSULTATION ENDS
Case 2 – Borderline

FY: Hi, my name is Dr _____ and, if I can just check your name (reading), you’re ______

SP: Yes, that right

FY: and you are 34 years old, and you’ve been sent up here from A&E because you blacked out?

SP: Yeah, that’s right

FY: The A&E sheet says that you looked like you were fitting

SP: Er, yes. Apparently (Looks concerned)

FY: Do you remember landing on the ground?

SP: No, I don’t

FY: Did you have any chest pain?

SP: Not that I can remember

FY: and was your heart going quickly or irregularly

SP: No

FY: And did you wet yourself?

SP: Yes

FY: And did you bite your tongue?

SP: No

FY: And the actual fitting –were you shaking your arms and legs

SP: I think so. I don’t remember that bit

FY: No, I guess you wouldn’t. But that is what you were told by... bystanders

SP: Yes, my colleagues

FY: Oh right, so you were at work?

SP: Yeah

FY: and did they say anything else that you were doing. Any strange movements or kind of doing strange things

SP: Uh, I don’t know
FY: Or maybe one arm jerking more than the others to start with

SP: Sorry, I can’t remember, they didn’t say anything about that

FY: Did you see any strange things or maybe smell something or hear something before you blacked out?

SP: No, I just felt quite dizzy

FY: How long you were drowsy for afterwards

SP: I’m not sure. Once I came around, I felt really sick

FY: But did you take quite a long time to come around?

SP: They said I was out for about a minute and then I woke up. And I just felt really sick.

FY: And you said you felt dizzy

SP: Yes. Really dizzy. And sweaty

FY: and was that before or after

SP: that was mostly before

FY: Would you say that that was a spinning type of dizziness, or a light-headedness?

SP: What do you mean

FY: Was it like the room was spinning? Like you’d been on a merry-go-round

SP: Possibly. It was, everything started to swim and then blur and kind of went distant, and I felt really sweaty and then the next thing I remember I was waking up and I felt really sick.

FY: And were you stressed or anxious or emotional?

SP: Well, My boss had been putting us under a lot of pressure, so maybe I was a bit stressed

FY: was it too hot or were you feeling unwell

SP: It was quite warm I think

FY: And had you just stood up quickly

SP: No

FY: Or turned your head suddenly to one side or the other

SP: No

FY: Or rubbed your neck
SP: No

FY: Hmm, ok and was this the first time ever that this happened?

SP: Yes, well, no. I haven’t blacked out before. I felt very dizzy in the pub a few weeks ago, but I went outside and sat down and it went away.

FY: Um, ok. But no other episodes exactly like this. Where you blacked out and were jerking?

SP: No

FY: Ok, so, You were at work and you collapsed, and you felt really dizzy before hand – its not quite clear if that was a spinning kind of dizziness or pre-syncopal – and you were sweaty and then you were on the ground shaking, and you wet yourself but didn’t bite your tongue, and you were probably out for a minute. You can’t remember any aura as such but you felt sick afterwards although not all that drowsy. You hadn’t stood up quickly and you hadn’t twisted your neck and you didn’t have any palpitations or chest pains.

SP: Um, yeah.

FY: How’s your health been in the past

SP: Ok generally. I had my tonsils out when I was 7 and I think I had a fit when I was a baby

FY: Hmm. That’s not usually a part of this. But no more fits after that?

SP: No

FY: and no other health problems?

SP: No

FY: No diabetes, or heart attacks or angina,

SP: No

FY: or epilepsy

SP: No (Looks worried again). So is this epilepsy?

FY: Well, it’s a bit early to say. It can be difficult to work out what these kind of things are due to, but there are lots of things that can cause blackouts like just fainting or your heart going into funny rhythms.

SP: Oh

FY: Do you take any regular medicines?

SP: No

FY: Are you allergic to any medicines?

SP: No
FY: Do any problems run in the family?

SP: My cousin has epilepsy

FY: Oh, I’m sorry to hear that. What do you do for a job?

SP: I work in direct sales, as a team leader

FY: Do you smoke or drink?

SP: I don’t smoke. I drink a couple of times a week. Maybe 6-8 pints in total

FY: Do you have a partner?

SP: Yeah, and we’ve got a 2 year old son

FY: And just a few more questions to make sure I’ve not missed anything out

SP: Ok

FY: Do you get any chest pains or palpitations

SP: No

FY: Or breathlessness or coughing or wheezing

SP: No

FY: Or bloating or heartburn or indigestion

SP: No

FY: Or abdominal pains

SP: No

FY: Or diarrhoea or bleeding or mucous

SP: No

FY: Or urinary frequency or trouble getting started when you are passing urine?

SP: No

FY: or blood in your urine

SP: No

FY: Or any weakness or numbness in any of your hands or feet

SP: No
FY: Or any sight problems

SP: No

FY: Or co-ordination problems

SP: No

FY: Or getting very depressed or anxious

SP: I did feel a bit anxious today I suppose, but not generally

FY: Or seeing things or hearing things that might not be there?

SP: (looks surprised) No

FY: Ok, good, I think that about covers everything. Do you want to ask me anything?

SP: Well, so you can’t say yet if it was a fit

FY: No, not yet, we need to find out more

SP: Ok

FY: Ok

CONSULTATION ENDS
Appendix 5: Consent form and participants information sheet for study 1

Consent Form – Assessor participants Study 1

Assessor Judgments in WPBA

Researcher: Dr Peter Yeates

1. I acknowledge that I have read the information form for Consultant assessors participating in stage 2 of the study (dated 08/01/10, version 2). I have had the opportunity to consider the information and ask questions and have had these satisfactorily answered.

2. I understand that I am under no obligation to be involved in this research study and that there are no consequences to choosing not to be involved.

3. I understand that I can withdraw from the research study at any time, without consequences.

4. I understand that once data from my investigation has been anonymised, it will not be possible to remove that data from the study if I choose to withdraw.

5. I understand that direct quotes from what I say may be used in research results, including being published, but I understand that they will be anonymous if this occurs.

6. I understand that the interview will be audio recorded and that the recordings will be sent to an external company to transcribe.

7. I give my consent to researchers retaining the mark sheets that I produce during the assessment and using these within the study.

8. I understand that the copyright and any intellectual property which may arise from the videos will belong to the researchers.

Please initial to indicate
Prior to the interview:
I agree to taking part in the above study:

Name……………………………………

Signature……………………………….                                  Date……………

Statement concerning Foundation Doctors role in video clips:

The videos that you are about to see are scripted performances. The performance that the foundation doctors in these videos show are due to the scripts that they were given and not due to their individual abilities as a doctor. You should in no way consider these videos of the foundation doctors to reflect their level of competence in their clinical work. In the event that you meet any of these foundation doctors in the future, you should in no way base any opinions or judgements on them as a result of the performance you see in these videos.

I have read and acknowledge the above statement

Signature……………………………….                                  Date……………

Consent Form – Assessor participants Stage 2

Assessor Judgments in WPBA

After the interview:
Now that the investigation has concluded, I continue to agree to take part in the above study as outlined in the information sheet (dated 02/11/09, vs 1), including giving my consent for video material to be used as described above.

Name……………………………………

Signature……………………………….                                  Date……………
Information Sheet

Assessor Judgements in WPBA

Information for Assessors in study 1

Background:
Workplace based assessment was introduced with the intention of being able to assess the actual clinical practice of junior doctors in a way that was authentic and represents how they perform in their clinical work.

There has been quite a lot of study of various aspects of WPBA – with some mixed results. In many types of assessments (like OSCEs for example), there are very specific descriptors to guide assessors in allocating scores. In WPBA, assessors have to judge what they see compared to “the expected level for completion of stage”. We know very little about how assessors reach these judgements, or what they are based on.

The purpose of this research is to investigate how assessors make judgements in WPBA.

How will the study work?
The study will have two stages. In the first part, we’ll make some videos of foundation doctors performing MiniCEXs. These will be artificial, but we’ll make them as life-like as possible. There will be a foundation doctor, a consultant acting as the assessor, and a simulated patient.

In the second part we’ll show these videos to other consultants, and get them to score the assessments and give feedback. Whilst the assessors are watching the videos we’ll study the way they come to decisions about how the foundation doctor has performed, and how to give them scores.

If I choose to take part, which bit would I be involved in?
We would like you to take part in the second part of the study – by watching the videos as an assessor.

What would taking part involve?
The principle researcher would arrange to meet you at a time that is convenient for you, in your normal place of work. This must be in a room in which we won’t be disturbed – an office would be ideal.

You will be asked to watch a video clip on a laptop computer, listening to the sound through headphones. You will be taught how to perform what is called a “think aloud protocol”. In essence, this is a way of speaking your thought processes out loud as you have them, whilst performing a task. You will be asked to practice this approach until you feel comfortable with it, using a sample video clip.
You will then be asked to watch 3 video clips which show foundation doctors performing simulated miniCEX assessments of new patients being worked up on a medical assessment unit. In each case you will be asked to watch each video clip – thinking aloud as you watch it. At the end of each clip you will be asked to score a mark sheet, and state what feedback you would like to give to the foundation doctor.

Whilst you are watching the video, and whilst you are giving feedback at the end, what you say will be audio recorded.

After watching both video clips, you will be asked some questions by the researcher.

**How long will it take?**
The expected duration of the investigation is approximately 1 hour.

**How many times would I have to do this?**
You would be asked to perform this once as part of this study.

**Will this be an assessment of my ability as an assessor?**
No. The purpose is not to study individuals’ ability or performance, it is to learn more generally about the thought processes that lead to assessors judgements. Results pertaining to individual assessors will not be fed back to the Deanery or any other body. All data used from the study will be anonymised.

**What will happen to the audio recording of my think aloud protocol?**
This will be treated as confidential information. The audio recording itself will be numbered so researchers can identify it, but no list will be retained of which assessor produced which audio recording.

The audio recording will be sent to an external company for transcription. The transcription company will also abide by a confidentiality agreement.

Transcripts of the audio recording will be fully anonymised. Any information within the transcripts that is judged as capable of identifying any individual will be removed. They will then be analysed by researchers to inform the study’s research question.

**Might what I say during the study be quoted?**
It is expected that direct excerpts from the think aloud protocols will be used in dissemination of study results. These will be fully anonymised.

**Are there any conditions under which confidentiality would be breached?**
Yes. In the unlikely event that information arises during an investigation that indicates serious misconduct or a serious threat to the individual’s personal health or wellbeing, researchers will breach confidentiality. This would only be done in liaison with the project’s supervisor and the chief investigator. Confidentiality would be breached by approaching an individual in the Deanery responsible for foundation doctor training, and then may be taken forward within the trust within which the assessor works.

**Do I have to take part?**
No. Participation is completely voluntary.

**What if I want to withdraw later?**
You are free to withdraw; there would be no consequences to withdrawal. Once audio transcripts have been anonymised, it will not be possible to remove the data that an individual assessor provided, as their transcript will not be traceable.

**Will I receive any payment?**
No, you will not receive payment for the study.

**Can anyone take part?**
Assessors in stage 2 will all be consultant physicians who have at least 2 years experience of conducting WPBA, and who have a reasonable level of involvement with foundation program training. Anyone meeting these inclusion criteria can participate.

**Will I need to sign anything?**
Yes. If you agree to take part, you would have to sign a consent form. This will include giving permission to be audio recorded. It will also include a statement about copyright and intellectual property of the videos.

**What if I have any further questions?**
Feel free to contact one of the individuals below if you would like to discuss any aspect of this

Peter Yeates  
Research Fellow in medical education  
ATR4, Education Research Centre  
UHSM, Southmoor Road  
Wythenshawe  
Manchester M23 9LT  
[Peter.yeates@manchester.ac.uk](mailto:Peter.yeates@manchester.ac.uk)  
Tel: 07932 675066

Patient Advice Liaison Service  
UHSM,  
Southmoor Road  
Wythenshawe  
Manchester  
M23 9LT
Appendix 6: Sample transcript from study 1

I:  Okay, so this is the start of assessor seven and I will just start you there.

1st video: Good performance

Concurrent section:

A: 0:05  Good introduction. ... Yeah, that’s good. ... Good, I like it.
A: 0:18  She’s letting the patient talk, that’s good.
A: 0:23  She hasn’t interrupted. Yep.
A: 0:33  This is good, yeah. I like the way she’s wanting him to a...talk. She’s not closing the questions.
A: 1:04  Okay, here’s the closed questions coming now. That’s alright. So good introduction, good scene setting. Now she’s moving in a bit.
A: 1:32  Oh right. So she’s picking up the fact that this may just be a simple vasovagal. She’s doing well here.
A: 1:43  And she li...I like the way she asks it. Yep, she’s thinking. Yep. She’s feeding him some lines but I think she’s able to do that because she’s let him talk before this.
A: 2:21  Okay, that’s good, she understands the difference between sitting and standing.
A: 2:31  Good. She’s asking good corroborative stuff.
A: 2:39  Yeah, good. She asks what they mean. ... Yeah.
A: 2:51  So she’s...she knows what we’re dealing with here: fit and faint, fit and faint, and she’s going through it well.
A: 3:04  He won’t know that.
A: 3:14  Oh, absolutely brilliant. That’s exactly what she should do and she’s done exactly that and asked permission.
A: 3:45  Good, she’s gonna recap now. Let’s see if it sounds anything like what he said. Good. Yep. I mean they’re great at recapping now, the foundations. Sometimes they recap so much no bloody progress ever gets made, but I think she’s doing this well.
A: 4:25  She’s very good, this girl. She’s got a good grasp of this. Yep.
A: 4:50  Yep, she’s looking into the effects on him now. This is so good. Oh yeah. She’s already getting into why he thinks what he thinks is coming.
A: 5:06  Yeah, she’s got the worries coming out about the job. She’s sympathetic, empathetic.
A: 5:14 Yes, absolutely, she’s right. Yep, good, she’s gonna run some tests. She’s taken a good history.
A: 525 Good. She’s not telling any lies. But she can’t…she’s not giving false reassurance. Yeah. Yeah. Yeah.
A: 5:43 Good. She’s gonna go into family history now I guess. Yeah. So she’s now going…she’s gone away from the presenting complaint and she’s going back through the, er, background. I mean possibly she should have done this before she started telling what she thought it was, but he was desperate to know.
A: 6:12 Good. She’s all asking the routine stuff now. Yep.
A: 6:36 Good.
A: 6:45 Well he did say that before. She didn’t have to ask that again.
A: 7:01 Good. Bit of chance now for a bit of preventive stuff.
A: 7:07 Is she gonna give any advice?
A: 7:14 Okay. She’s thinking it through. Okay. Oh, golly, she’s really going into the whole fitness thing, which is good. [Pause]. Yeah.
A: 7:38 Throughout this she’s had good eye contact, she’s been pleasant, she’s been empathetic. I mean golly, it’s very difficult to fault any of this. She’s just going through a review of symptoms now. Systems, sorry. A review of systems. [Pause].
A: 8:01 Yeah, he keeps telling you that. He’s anxious at the moment. Like there is now. Yeah. Good. Yeah.
A: 8:19 Oh this is excellent; she’s going back to get his understanding. Yep. That’s right. Yeah. Okay. Yeah… Yep… Very nice. Okay, we can stop it now. Is that it? Do I stop?

Retrospective section:
I: Yes, tha…that’s fine. Y…
A: I shut up now do I?
I: …you don…n…well no, no…not at all.
A: Alright.
I: Um, you, you can take the earphones out if you like to…
A: Thank you.
I: …ju…ju…just for your own comfort.
A: Yeah.
I: Er, but what I’d like you to do, yeah, please, is if, er…now I’ll just move that…
A: Okay.
I: …so we’ve got enough space, make sure it’s not [inaudible 09:10 voices overlap].
A: Alright, well the clinical setting I think you told me was acute admissions. Th…
I: Yeah. A j…w…w…just, just focus on, on these boxes…
A: Oh, just want to focus on the clinical encounter.

I: ...and, and, and the free textboxes underneath if you...

A: Alright. The history taking for an F1 I thought was above expectations for an F1. I thought she was very, very good indeed. The physical examination we didn’t do so we wouldn’t be able to, er, comment on that. Her communication skills, again, were above expectations. I mean I’d be tempted maybe to put even well above a...um, but, er, cos it’s the first one I’ve seen it’s hard sometimes to know isn’t it? But for an F1, especially if she was in her first month or whatever that would be well above expectations. I thought her clinical judgment was above expectations, I thought her professionalism was above expectations. Her organisation and efficiency, well, it’s a bit difficult to see that in this particular encounter, but in terms of that encounter, um, she got to the crux of the matter quickly, she had a system that she worked by, so that was, again, above expectations. And her overall clinical care, if she was working for me I’d be absolutely delighted. So anything especially good? Well, she made good eye contact with the patient, she was empathetic, she gave him time to talk. She started with the open ended questions. She went into the closed questions to get a bit more information. She then went through the routine things about previous illnesses, et cetera. The whole time she, er, summarised things. She gave him a chance to say what he thought. She got into the bottom of what he was worrying about, which turned out to be his job and then if he lost his license because of it being epilepsy and she found out that he was worried it was epilepsy. I mean it was an absolutely all round almost perfect performance. I would be very happy with that. Thank you.

I: Great. Thank you very much. Tha...that’s, er, that’s good to have that, that kind of feedback. Um, er, er, I normally at this point a...ask the, er, the, the, er, the assessor to, to state what feedback they would give but I think you’ve, you’ve probably...

A: Yeah.

I: ...ju...just done that.

A: No, I would tol...tha...I thought that was excellent. It would be hard to fault any of that really.

I: Uh-huh. Okay. As you’re watching that do you have an awareness of the way that you’re making the judgment o...of your thought processes as you do it?

A: Well, I guess you pick up very early in the interview the clinical scenario. Okay?

I: Uh-huh. Okay. As you’re watching that do you have an awareness of the way that you’re making the judgment o...of your thought processes as you do it?

A: Uh-huh.

I: Uh-huh.

A: So you’re picking up very quickly from her questions and his answers that this is a...is it a fit, is it a faint. Now, that is a very common and very important clinical encounter that happens many, many times. Um, so you’ve already got...being an experienced clinician you’ve already got your system. So a lot of what I suppose I’m judging her on is is she doing it as well as I would do it, so I suppose that’s...

I: Uh-huh.

A: ...a bit big headed and selfish, but that’s really what I’m doing. Now, that wouldn’t mean that if she did something slightly different to me but I felt it was valuable or I’d learned from it that I wouldn’t give her extra points for that. I wouldn’t be ridged in the fact that she had to do it my way, but she certainly
covered all what I would perceive if I was dealing with the patient the important things to cover.

I: Uh-huh. So in terms of, of what she covered you, you were quite happy...
A: Yeah.

I: ...with the kind of factual coverage.
A: Yeah.

I: When you were giving her feedback you mentioned quite a number of different things that you thought she’d done well.
A: Yeah.

I: You, you mentioned her communication style...
A: Yes.

I: ...the way she was questioning...
A: Yes.

I: ...um, the information that she covered, the way she looked at the patient’s...
A: Yes.

I: ...ideas, concerns.
A: Yes.

I: Um, do you think any one of those points, or indeed any of the other points that you focused on, is more important than any of the others in the way that you make the judgment?
A: [Laughs]. Yeah, that’s a good question. I guess, being an old fashioned guy, I’m more interested in the fact that she gets to the crux of the clinical situation. Um, that’s the important thing, that she doesn’t say anything that is false to the patient, that she doesn’t do anything that would be dangerous to the patient, that she doesn’t give false hope or tell the patient he’s got epilepsy when he obviously hasn’t. Um, so I guess me personally, the clinical thing is the crux. You would then expect after that that she’d be pleasant in that, that she’d be nice, she’d be empathetic, but that, I suppose, would be slightly less important to me. I would forgive her if she was a bit brusque with the patient provided she didn’t do him any harm. The fact she seems to have the full package.

I: Okay. So clinical details first and then, um, niceness later.
A: Yeah, I think so.

I: [inaudible 13:44 voices overlap].
A: But, but remembering of course that if you’re nice to the patient you’re more likely to get the correct clinical details.

I: Sure.
A: You know, I’m not forgetting that. But, er, yes, I suppose simplistically put that...what you said is right.

I: Wha...wha...and what about the other way round? What if she’d, um, been nice but, but hadn’t asked all the right things?
A: Well I’m afraid, er, she’d have done badly with me.
I: Okay.
A: Not as badly as someone who’d been horrible and done all the wrong things. You know what I mean?
I: Right. Yep.
A: There is a grading to it.
I: Sure. Okay. Now - and I’m very interested in this - you talked a little bit about…you said above expectations, might even be well above.
A: Yeah.
I: Um, how do you judge the difference between those two things?
A: Purely by the number of times you’ve seen F1s in action. I mean I’ve supervised F1s for years. I have F1s working with me every day, all day, so I’m able to watch them closely. I’m…I do mini CEXs with them. I get them to present their patients. So I have a good feel for what an F1 can do.
I: Uh-huh.
A: And you’re very forgiving to your F1s aren’t you, for god’s sake. They’re only just starting. But the…so I, I thought that girl was excellent.
I: Okay. Good. Thank you. Er, well if, if you’re happy with that we’ll, um, we’ll go on to the next one.
A: Yep.
I: Um, so, yeah, if you just pop those back in, I’ll give you that form again.
A: Have I filled that other one in alright for you and…?
I: Er, you’ve...yeah, that’s, that…um, um…
A: I’ve not put my name or anything on.
I: I’m, I’m, I’m, I’m happy with that. No, no…
A: Okay.
I: …that’s fine. I’ll just, um, I’ll just...oh where’s the cursor gone? There it is. This will probably start playing straight away…
A: Okay, I’m ready anyway.
I: …so I’ll just have to pause it and get it back to the start for you. Just so I’ve got...so I can get the screen big enough. Okay. So, again, if you’re happy, I’ll start.

Anything especially good?

(blank)

Suggestions for development?

(blank)

Agreed action?
2nd Video: Borderline performance

Concurrent section:
A: 0:02 She’s introduced herself nicely. Good. Again, good eye contact.
A: 0:09 Good, she’s checked the patient. Yeah.
A: 0:19 Oh dear. She’s already used the word ‘fit’ you see, which I don’t think she should do.
A: 0:29 Oh my god, she’s gone straight into the closed questions. She’s not done an open question at all.
A: 0:38 Oh. Did you wet yourself? No. She doesn’t care about his embarrassment. She hasn’t picked it up at all.
A: 0:48 She keeps saying fitting to him. She’s already made her mind up it’s a fit. She’s all over the place basically.
A: 1:07 Oh, good grief. It’s dreadful to watch. She’s gone straight into the closed questions. They’re haphazard, they’re all over the place. She’s talking fitting, she’s...
A: 1:18 oh good god, she’s gonna scare him to death. She’s asking about did you do any more strange movements, so he’s gonna wonder what’s going on.
A: 1:32 Look at the poor man now, he’s totally broken hearted. He’s not gonna be listening to anything she says after this.
A: 1:53 They’re not just closed questions, they’re leading questions. She’s almost telling him what she’s thinking in the question and then she doesn’t understand that he’ll pick that up. She doesn’t know what dizzy means. You know what I mean? He says she feels dizzy. That can mean all sorts of things. She doesn’t explore it at all.
A: 2:18 Oh, wait a minute. Forgive me, she’s gone back to it.
A: 2:32 Okay, that’s not a bad question.
A: 2:42 Yeah, so it wasn’t vertigo.
A: 3:08 Again, it’s all leading questions.
A: 3:24 She’s just going through all the rigmarole of questions, but she has no concept of how they all come together into one. I dread to think what she’s gonna tell him.
A: 3:44 What you might do with this girl is you’d let her go on a bit longer but you might even interrupt this now and sort of say well, come on, let’s, er, let’s not go any further for the moment, because I think she’s actually doing irreparable harm here.
A: 3:59 Presyncopal. For god’s sake. As if he’s gonna know what that means. Yeah, what we’ve got here is the example of how not to do it.
A: 4:13 Look at the proportion of the number of times she’s spoken to his. It must be about 90 percent her and 10 percent him, whereas you look at the difference with the other one.
A: 4:28 Look at him, he absolutely hates her. Have you any other medical problems, like she’s already telling him this is a medical problem.
A: 4:51 All jargon.
A: 5:16 Irregular heartbeat. Why bring that up? What she should say when he asks is I haven’t got all the information yet, I’ll come back to that, I can see you’re anxious.
A: 5:48 Oh, she keeps saying that. Instead of convulsive activity with a faint. Sh...now she’s stuck. She keeps saying...asking the same questions over and over again. She’s lost now. Oh, she’s just remembered another question. New.
A: 6:16 Cousin has epilepsy.
A: 6:25 He told you that.
A: 6:46 Yeah, she’s gonna ask a few other questions.
A: 7:11 Just going through a rote system of questions now. There seems to be no pattern to them. She comes up with anything that comes into her head. She obviously doesn’t know what she’s doing.
A: 7:26 Have you any coordination problems? What does that mean?
A: 7:54 Oh, here we go, she’s asked him if he’s any questions. Here we go. Is it a fit? Okay, fair enough, she doesn’t know for definite it’s a fit. ... More tests. Could be a fit, ... she said. ...Yep.
A: 8:25 So she thinks it could be a faint as well.
A: 8:32 He’s asking about a car. She doesn’t know the answer to this but I bet you she makes an answer up.
A: 8:44 Okay, fair enough, she hasn’t done so bad there. She hasn’t given a definite answer. Okay, so she’s finished. Now, what do I have to do now, remind me [inaudible 24:21 voices overlap].

Retrospective section:
I: So if you could just score the, er...
A: I’ll just score this.
I: ...the mark, mark...the mark sheet.
A: You don’t want me to do anything else.
I: And then, then again I’ll, I’ll get you after that to state what, what, what...
A: I’ll talk while I’m doing it then.
I: ...feedback you’d give.
A: It’s...you’ve got to be careful with your F1s, I’ll be honest with you. You’ve got to be reasonably charitable to them because they’re fairly new, so it’s very har...very rare that I would put below expectations on somebody I’d only just met. So I would say borderline to let them know I wasn’t happy with it. I’d say borderline. Because then one would look to improve in that performance while they’re with you and you’d make sure you went back and saw them again. And obviously if they’re not making any progress then you’re gonna start to put them below expectations, but on the first meeting I think it’s hard to put below. So
I’d…borderline with history taking. Obviously the physical examination she didn’t
do. Communication skills, again, I’m afraid were borderline. The, the F1s, to be
fair, are usually very good at the communication part of it, so she’s even bad at
that. And clinical judgment, well, again, wasn’t too bad, but the end of it she
didn’t say it was definitely a fit. She said it could be a faint. She was quite good in
saying we need new tests, and she didn’t give definite answers. So I would
probably put meet expectations for that bit. Professionalism is a difficult one
isn’t it? Um, she, you know, she said all the right things, looked the right way
even though what went on, so I don’t think I can knock her professionalism. Her
organisation and efficiency’s below expectation. She…even though she didn’t
even allow the patient to give the open ended answers, even going through the
closed questions they were haphazard and all over the place. She was repeating
it, she kept asking the same questions, so that would be low. And her overall
clinical care would therefore have to be borderline. I mean I’m a bit of a dove I
suppose. Some hawks would put her below expectations and I wouldn’t have
too much to argue about there. But I would…because of the system I play where
I like to give them a second chance before we start judging them then I would, I
would bring it to their attention I wasn’t happy by saying borderline, but then
we’d have a plan as to how they could put it right.

I: Uh-huh.

[interruption]

I: Um, okay. Er, well, tha…thanks for doing that. Um, could you now state what
feedback you would give to her if she was...

A: Feedback.

I: ...sat facing you in this room.

A: Okay. The feedback I’d give to her, first of all we would go over the case and
summarise the case. I’d ask her well what do you think we were dealing with
there and get her in the end to decide it’s the fit/faint scenario. I would then say
to her what did you think from the clinical symptoms it could be, and eventually
we’d get her to talk around and hopefully convince that it was more likely to be
a faint than a fit. I would then get very much into, um, how she could improve
her history taking by asking the open ended questions first, and I would give her
quite stringent advice on that and saying in future what you do is you go in,
introduce yourself, which she did very nicely, make sure it’s the right patient,
and then you ask the question and you shut up. Um, there’s been this study
hasn’t there, how long do doctors, er, not interrupt when they say they’re not
gonna interrupt? Eight seconds. So I’d explain that to her, so tell her that she’s
not alone in this, and try and prove to her how much better she’d have done and
then, um, say to her…and then having gone through her history taking then say
to her and you also paid no attention to what that man was going through, and I
would say to her what do you think he was going through, what do you think he
got out of that, what do you think he worried about, what do you think he thinks
was wrong with him, what was the big things he was worried about, and I’m sure
she won’t be able to answer that question and you’d be able to then explain to
her look, he’s worried that he’s got epilepsy. His cousin’s got epilepsy. He’s
worried that he’s gonna lose his license and lose his job. You kept using the word
‘fitting’. You were, um, if you like, reinforcing his worries by that. So I would go
through first of all the clinical scenario, how she could improve her history taking
and then think more about what messages she’s giving to the patient, and I think
I could convince her. It would be nice to have the video to show to her wouldn’t it [electronic beep] but we don’t have that, but we could convince her that it can be done better.

I: Okay. Good. Thank you. So i...if I come to the scores that you’ve given her, you’ve given...mostly you’ve given three, er...

A: Four threes, two fours and a UC, yeah.

I: Yeah. Um, bu...but three more than anything else. But then you’ve said you think you’re being a little bit dove-ish.

A: Yeah, I think I am.

I: So what you’re saying really is that, that a score of three, a score of borderline is actually better tha...than you think she did there. W...is it...

A: Yeah, I think that’s probably...

I: ...is that true? I, I’m not...

A: ...true, yeah. I think that...

I: ...trying to lead you but...

A: You are leading me a bit, but yeah...

I: [Laughs]. Okay.

A: ...I’ll be honest, I think it...it’s probably better than she was, but I don’t want to wreck our relationship having only met her for the first time...

I: Okay.

A: ...and we are gonna go back and revisit it.

I: And so, so you...you’ve mentioned that, it, it being the first time y...you’re willing to let the first interaction with someone alter the score that you give.

A: Yeah. I think so. Because she hadn’t actually done anything to, er, harm the patient...

I: Uh-huh.

A: ...um, although she wasn’t, er...ob...obviously she’s worried him psychologically. We can put that right can’t we by eventually telling him the good news.

I: Uh-huh.

A: Um, so it’s a temporary, er, worry that she’s put him through, so I think we can do that. I did say it at the time though that, you know, in real practice I might have intervened in that and said...

I: Mm.

A: ...now, I just want to interrupt [phone rings], this young doctor is telling you that it could be, um, er...

I: If you need to take that that’s fine.

A: ...it, it could be a, a fit, but tha...tha...that’s her opinion, not ours. Are we able to stop for a minute?

I: Yeah, no, that’s fine. Yep.

[A answers phone]
I: Okay. Um, and, and the other thing I’m interested in in response to that is, is you said after the first video that it’s the, it’s the kind of the factual, the core kind of medical content that’s the important bit. How do you feel that she did on the core medical content in this video?

A: Um, I don’t think she did well. I mean...

I: Okay.

A: ...I really do not believe she did well. There were a lot of unanswered questions. Um, she may have asked all the questions in a bad order but there’s no doubt that she could not have assimilated them. I mean what we need to go through is the exact sequence of events and if you remember her...she used to ask things like how did you feel afterwards, had you stood up quickly before, did you...what were you doing when you were fitting. Er, you know what I mean, there wasn’t any impression from me that she’d got a full story about the actual complete sequence of events that happened, so she didn’t do well clinically.

I: Okay. Okay. So...and if...cou...cou...could I infer from that...again, I’m...

A: Yeah.

I: ...trying to say this without leading you, but...

A: Yeah.

I: ...just trying to make sure that I understand you correctly.

A: Yeah.

I: A sort of ticking off each question isn’t enough I, I think is what you’re saying isn’t it?

A: Yes, absolutely right.

I: You have to put it together.

A: Oh got, yeah.

I: Okay.

A: And she didn’t go back over it with him like the first girl did, so she didn’t test out her hypotheses, she didn’t ask him to, um, fill in any information that she didn’t have. I mean it was an incredibly scanty, poor history I’m afraid.

I: Mm. Okay. Right, we’ll move on to the third one...

A: Okay.

I: ...and then we’ll have a, a bit more chat about, about all three of them...

A: Alright.

I: ...after that if that’s okay. Er, so, again, same thing as last time. I’ll just shut that and then I’ll open this and I’ll have to fiddle around with it again. So I’ll just make it bigger.

A: You need to give me one of these don’t you?

I: Um, yeah, so it’s, so it’s...

A: That’s this one.

I: ...that one. And let me just...okay, now...er, could I have my...oh tha...that’s my pen there isn’t it, cos I, I need it whilst I’m...okay, brilliant.
A: Yeah.
I: I’ll start there then.
A: Here we go.

3rd Video: poor performance
Concurrent section:
A: 0:02 She hasn’t introduced herself.
A: 0:11 Oh lord, here we go again.
A: 0:20 This one’s...she’s not looking at him at all, so she’s the least of the empathetic. Again, she’s straight down the epilepsy thing. No open questions at all.
A: 031 And she’s...oh god [laughs]. She’s the real non-sympathetic girl. She’s shaking her head from back to forth. She’s obviously bored, uninterested. She’s giving very bad signals. Did you hurt yourself? Oh, ye...biting on her tongue, shaking herself about.
A: 0:59 She’s leaning back.
A: 1:07 God, I actually believe this is the worst of the three, cos not only is her history taking appalling but she looks as if she couldn’t care less anyway. Think...so the second pat...the second girl was poor at gathering the history but did look as if she cared. This girl’s poor gathering the history and looks as if she couldn’t care less, like he’s a, a nuisance.
A: 1:39 [Laughs]. Well how can he answer that question? He, he didn’t see any of it. She’s feeding him lines here.
A: 1:54 She’s ignored the sickness and dizziness. She’s ignoring that.
A: 2:07 Right, he’s trying to tell her something. She keeps pulling faces. Doesn’t let on that she’s listened to him.
A: 2:26 Strange how she talks to him. She talks to him like she’d talk to another doctor. She’s assuming that he knows what these questions are about.
A: 2:44 Oh, she’s ignoring...[laughs] absolutely ignored that.
A: 3:16 Oh lord, she’s saying it’s a fit. She’s made up her mind it’s a fit.
A: 3:27 God, she looks bored out of her brain. This is the worst of the lot because this is appalling both medically and in terms of attitude.
A: 4:06 Oh, no explanation. Using the word EEG. See what the cons....s...yeah, here we go, he’s gonna ask a question now.
A: 4:29 Here we go. He’s asking about the driving license.
A: 4:50 God, that was absolutely diabolical.

**Retrospective section:**

A: I mean truly diabolical. I mean she looked like she was bored to death, she didn’t care less about him, she hardly looked at him. She’d made up her mind already it was a waste of time, she wasn’t gonna take any interest in him, she was just gonna wait for the consultant. She hardly looked at him, she looked bored. Her body language was awful. She was leaning back. She dismissed anything he said. Um, she took an appalling history. I, I’m not even sure what she thinks it is. I think she’s decided it’s a fit but she didn’t even make that totally clear. Um, I mean that is diabolical. I mean I know I was fairly…[laughs] looking back now, possibly I wasn’t too dove-ish with the second one, er, but this one, you know, you would definitely have to mark her down. That is definitely below expectations on every level really. So do you want me to go through me marking?

I: Yeah, please, if you would.

A: Alright. Her history taking was below expectations. Now, you know, I’ve a wondering…I’m wondering if a lay person in a pub couldn’t have taken a better history and shown some interest. You know, your bloody aunty would ask you more questions. So you might even think about well below, but I suppose I’m a bit dove-ish, so definitely below. Communication skills definitely below. Clinical judgment definitely below. Professionalism definitely below. Organisation and efficiency definitely below. Overall clinical care definitely below. Physical examination I didn’t see. Um...

I: Thank you.

A: I…it would be hard, um, to come up with anything good about that. Um, if you were feeding back to her, which is what you’re asking me next, I suppose you’d have to sit down and say well, tell me how you think that went. Um, and she’d say oh yeah, I think it went reasonably well or whatever. And then you’d have to say to her well, you describe to me exactly what happened to this guy, and then I’d get her to describe that, and when there were huge gaps in it I’d then say to her well did you ask him about this, did you ask him about that, what do you think about that. And I guess over...god, the feedback here would take some time, but over about ten minutes you might be able to feed in. You might then with the history start saying to them where do you think this history’s taking us, what clinical scenario we, we...are we in. And I’m not even sure she would understand this thing about fit and faint and, and how important it is. And then I’m not even sure we could go any further. Y...they often say when you feedback you, you, you should only feedback a bit and I may have to bring her back. Mind you, that won’t work because she’ll have forgotten. But almost with a heavy heart having gone through the clinical scenario you then have to go through with her what about...what’s he feeling, did you ever ask him how he felt, what do you think he’d feel. I’ve a horrible feeling this girl is so un-empathetic she wouldn’t even know he’d be bothered. Um, and then you might ask about what she’d done to reassure him or what information she’d given to him or what she’d told him or how she’d summed it up at the end, how she’d put the plan in. So, oh, I could go on and on, Peter. I mean it was just diabolical. I mean you, you’d be very hard not to be a little bit angry when you fed back to her. You really would have to be a little bit angry. I can’t believe anyone...there can’t be
many coming out of medical school that bad. Alright. Any specific questions [laughs]?

I: No, no. Tha...tha...tha...tha...that's great, thank you. Um, so I mean the...it's, it's very useful to have that, and there are a number of things I could ask you about, about, about the three videos. Um, but I'm interested, in relation to that one, one thing that you said was looking back probably the second one I wasn't being as dove-ish as I thought.

A: Yeah.

I: And, and you sort of almost, um...forgive me if I'm...

A: Yeah.

I: ...if I'm leading here...

A: Yeah.

I: ...correct me if I'm wrong...

A: Yeah.

I: ...but, but you almost seemed to change your mind of the second one...

A: Well...

I: ...a little bit.

A: ...I think...I examine as well don't I, for the Royal College of Physicians, and I...if you ask any examiner, um, it's easier if one of the candidates comes in and is brilliant first cos then you've got your standard.

I: Uh-huh.

A: Um, and you're quite right, it is hard sometimes, um, when you've had the three things come at you. Um, what you try and do is you try and assimilate all of them three, which I don't suppose you do in real practice do you, but, um, it's sometimes hard to know exactly where they stand. I mean I know I gave you some big spiel before didn't I about my mass of experience with foundation officers and how I know they should be, um, but I suppose all of us can be a little bit, er, shaken by that.

I: So, so you think recent experiences may, may influence your perceptions or...?


I: Okay. Um, so then if I come to...er, on the form it asks you to judge against this thing which says meets expectations for F1 completion.

A: Yeah, yeah, yeah.

I: How, how fixed a thing, how clear a thing do you think that is in your mind?

A: Well, it's not is it? I think it'll be very hard to define won't it? Um, and the educationalists no doubt will have a huge list of boxes that you must tick. But, um, you tend to go on a gut feeling don't you? You tend to know your team. I've got a team of F1s, F2s, ST1s, ST3s, ST4s, so I've got some idea of the hierarchy. I've got some living, breathing organism in front of me that lets me know where they should all be in this, er, hierarchy. So I've got a gut feeling about what an F1 should be like, so I suppose what you do at the end of a thing like that is you think to yourself if that was one of my family coming in would I be happy that
that was the first contact they had, knowing that they’re going to get more senior contact anyway. So, yes, you’re right, I’m afraid, like a lot of things in life, it’s gonna be based a bit on gut feeling and, er, your own personal prejudices. I’m afraid there’s no getting away from that.

I: Well I… I’m not, I’m not asking you to apologise for it, I’m just, er, I’m just exploring the ways that…

A: Yeah.

I: …the way that you think. Um, so you, you sort of describe…you, you, you’ve…you kind of described a couple of things there possibly because you’ve described in a way comparing them to more senior doctors.

A: Yeah.

I: Um, and, and you…but you’ve also compared…er, described comparing them to the other videos.

A: Yes.

I: I mean do you think you compare an F1 with other F1s or do you think you compare them with more senior doctors?

A: I think you do both don’t you? There’s a bit of horizontal comparison and a bit of, er, vertical comparison, and of course, you know, their…the way you judge them is going to be influenced to some extent by the rest of the F1s round them, so it…they might be a reasonably good F1 if they were in the company of other reasonably good F1s, but if all your other F1s are brilliant and that one’s only reasonably good then it’s hard not to see them as not as good as they should be. Do you see what I mean?

I: Okay. Yeah.

A: So you have to figure that in I think.

I: So you think the, the cohort that they’re in…

A: Yeah, it will.

I: …might…and do you think the cohorts change year on year or do you think they’re fairly standard every year or do you think, do you think it varies between hospitals or…?

A: Yeah, it’s a good question. I, I don’t think the cohort varies a great deal year by year. I don’t think it’s varied even from the time I became a house officer. I mean these kids come out fairly green behind the ears despite all they’ve learnt and it’s not just what they’re learning from their medical training that you look for, it’s all a bit to do with what they are as people. You know, how do they deal with stress, are they good to come and ask for help? I don’t know if these are things…I suppose you’ll tell me that these are things that can be taught. Um, but sometimes it’s, er, you know, have they got the common sense, er, something that can’t be taught, you know. So in general I don’t think they’ve changed. The poor kids that come out of medical school into their first house job or whatever it is, F1, are all going to be much of a muchness in terms of experience.

I: Uh-huh. Okay. How…if you’ve decided that, that they’re not…if you’ve got an idea in mind, albeit based on comparing with other people, um, of what meets expectations means, how do you judge when someone is either above or
below that, how far away from that they are? Whether they're not meeting it but fairly close to it or not meeting it and a long way from it?

A: You mean both above and below?

I: Mm.

A: Yeah. It is difficult, that. That is difficult. But if you think about what is the point of these mini CEXs, they're not exams are they? They're not pass/fail exams that we're doing with these kids. What we're doing is we're using mini CEX to improve their performance. So it, you know, it doesn't matter that it's way below or a bit below, it's below isn't it, and we've got to get it up. What's more important is how they progress over the attachment.

I: Mm.

A: And y...mean, I don't mind too much if they arrive terrible but by the end of it they've shown a willingness to learn and you can prove when you go through their portfolio that their mini CEXs to start with were ropey but now they're good. In fact, you know, I think that says a lot about a person. But if they're ropey to start with and they don't show any, er, improvement then that is absolutely a cause for concern isn't it?

I: Mm. You, you mentioned earlier, um, er, actually after the first video something about, um, if it had been done in August it would be well above expectations.

A: Yes. Yes.

I: Um, so I, I wondered do, do you think your, your standard, the standard that you judge them against, shifts through...

A: Yes.

I: ...through the course of the year?

A: I think it does, yeah. I think it does.

I: So, so when the box says meets expectations for F1 completion...

A: Ahh.

I: ...do you think you are marking against that or do you think, um...

A: No. Do you know, I'm, I'm sorry to say that I'm not marking about that, no.

I: And I'm not, I'm not, I'm...but I'm not trying to, bu...

A: I hadn't even read that by completion, no.

I: I, I, I'm not trying to pull you up or...

A: No, but you're right, yeah.

I: ...embarrass you there, but this is something I'm exploring with [inaudible 47:22 voices overlap].

A: Yes, well you’re right. I’m not judging it by completion, I’m judging it by...at that time, yeah. Yes.

I: So where you think they ought to be right now.

A: Yes. Yeah.
I: Okay. Okay. That’s, that’s interesting. Um, okay, um, now there are, there are seven different domains on here: the history taking, physical examination and communication skills and, and so on.

A: Yes.

I: And I’m, I’m interested that in each one you’ve...well, actually no, not in each one, because in this one you’ve, you’ve separated those into, um, into threes and fours...

A: Yeah.

I: ...but in the other two you gave the same...

A: Yes.

I: ...the same scores...

A: Yes.

I: ...for that one and, and this one...

A: Yes.

I: ...the same scores for all of them.

A: Yeah.

I: Do those things...do you think that’s because, er, these things get lumped together in your mind, do you think it’s because they did exactly the same with them in your mind, do, do you think they’re hard to separate out, do you think they influence each other so much you can’t separate them?

A: Yeah, I think you’re right. I think it’s hard to separate them out to be absolutely honest with you, and, um, you tend to give a gut score at the end don’t you? If you think they were good then you tend to mark them good in everything don’t you? It’s…I suppose A, it’s hard for me as the assessor to tease them out, and B, they do quite naturally tend to go together. I know that the examples that you’ve given us there have been quite clever haven’t they? One’s been very good, one’s been getting the history, er, badly but being a nice person, the other one’s been bad history and bad [laughs] person if you like. So it has been a bit, um, barn door-ish, but I...life isn’t as easy as that. So you’ll find the one that was bad in everything I’ve marked equally bad in everything, the one that was good in everything’s more than...and then the other one has got a little bit of differentiation because her history taking was bad but she was empathetic, so...but they are hard actually to, to, um...they’re hard to tease out both objectively and because just by natural...the way humans are, if you tend to be caring and, er, listening you do tend to get the best history; they do tend to go together.

I: Mm. Uh-huh. Okay. Um, how shall I put this? Um, in, in the two latter ones you, you described, you know, you...you’d want to, you’d want to make it clear to them that they weren’t meeting...

A: Correct. Yeah.

I: ...the, the standard that you expected.

A: Yeah.

I: Have you ever found in a face to face situation, um, that that can be a difficult message to, to, to...
A: Yes.

I: ...communicate to [inaudible 50:01 voices overlap].

A: It is a difficult message. I think it is a difficult message, but, um...and you’ve got to be careful how you do it. That’s why I think you can’t break their hearts, er, and be as...too tough on them to start with. I think you’ve got to set the ground rules that you’re telling them this for their own good and that this isn’t the final judgment, this isn’t the way I’m going to think about you for the rest of your life. You were bad on this occasion but together we’re going to put this right, and what’s more...

I: Mm.

A: ...I’m going to take an interest in you and come back to you and hopefully tell you by the end of your attachment, after I’ve seen another few, that yes, you have put it right, you know.

I: Mm. Um, d...would that ever make you feel tempted to not score them as harshly as you might instinctually feel the performance warrants?

A: Er, no. That particular thing wouldn’t. But I think, as I said to you originally, I, I think, I think if you go too hard on them I think you’re going to first of all...remember you’ve got...you are the appraiser for these kids. You are mentoring and them and, er, appraising them and supervising them throughout the attachment. You, you don’t want them to bloody hate you from the very start because [laughs] the chances of you getting any input to improve things are gonna be harder aren’t they? So it’s got to be things like yeah, you know, I didn’t like that, you were poor today, but, you know, I excuse you cos you’re, you’re new, er, and god knows I was as bad as you when I started. So you’d always do that sort of thing with them and, and give them the hope that we can put this right, you know, and what’s more, not only can we put it right but you’ll see it being put right and you’ll know that it’s been put right and you’ll enjoy the fact that you can do it better by the end, you know.

I: So i...if I...again, if I’m summarising you correctly, um, you’re willing to sort of withhold poor scores...

A: Yes.

I: ...to protect the sort of...

A: Yeah.

I: ...educational relationship...

A: Yes. Correct.

I: ...between...

A: Yeah.

I: Okay. Okay. Um, are you aware of any other factors that might influence the way that you score when you’re doing these things?

A: Yeah. Well the other thing is you tend to work a lot with other doctors don’t you...

I: Mm.

A: ...so you would go and talk to her peers...
I: Uh-huh.
A: ...um...they were all girls weren’t they? So you’d go and talk to the peers of the foundation doctor and say how do you feel she’s doing and do you think she’s struggling or whatever. You’d talk to your consultant colleagues, you’d talk to your middle grade.
I: Mm.
A: So that would definitely affect you. And, for instance, if you reported back to her...say you weren’t her educational supervisor but you reported back to them and said I did the mini CEX and god, she was terrible and they said well god, she’s been brilliant with me, then you’d be even more likely to think oh well, maybe it was a bad day, you know.
I: Mm.
A: Whereas if she...if they say oh god, she’s always terrible then I’m afraid that would influence you, yeah.
I: If it was your, our F1, someone who you’d worked with quite a lot...
A: Mm.
I: ...and your general impressions had been positive...
A: Yeah.
I: ...and then they performed like the person in either the second...
A: Yeah, yeah.
I: ...or third videos, how would...do you think that might influence your scores?
A: Um, influence my scores. No, I don’t think so. I don’t think it would influence my scores. It would influence...I would go and maybe even talk to them about, you know, I’ve had a report that that you, you didn’t do so well in the mini CEX with so-en-so and you’ve always been quite good with me, can you explain why it didn’t go so well. You might do that.
I: But, but if you were doing the mini CEX that day and you knew they were generally very good...
A: Yeah.
I: ...and then they do badly in front of you that day, w...
A: Yes.
I: ...ha...ha...how would you score in that situation?
A: Er, you’d score higher wouldn’t you? Yeah, you would.
I: I...I’m not telling you.
A: No, I know you’re not telling me, sorry.
I: But...
A: Yeah, me personally, Doctor _____ _____, yes, I would be more likely if they’ve previously been good to not score badly if they had...and I would put it down as oh they’ve had a bad day.
I: Okay.
A: Yeah. I’m being honest. That’s right.
I: Okay. Okay. Okay. Right. Um, has anything else ever occurred to you? Any other types of factors that, that might sort of sway that global judgment? Maybe, um...well, no, I’m not gonna, I’m not gonna lead you there.

A: [Laughs].

I: Er, I...er, have you ever been aware of anything else?

A: Um, I’m not sure what this question’s leading to. Do I score badly all the black doctors you mean or...

I: That’s...

A: ...the women doctors or anything like...not at all, no.

I: One of the...I suppose one of the things, one of the hints I’ve had coming out...

A: Yes.

I: ...um, from some of the other assessors is that, for example, um, if they find someone irritating it might influence their scores or, er, possibly aspects of their appearance [inaudible 54:37 voices overlap]...

A: Oh yeah, okay. So, sorry, nothing as gross...I’m sorry, I thought you were trying to trap me into, um, er, inappropriate behaviour. Let me think. Yeah...oh god. I mean I don’t think this will surprise you will it, but yes, if, if they’re tattooed and pierced then I’ve already taken a slight, um, against them if you like and they would have to perform a little bit better wouldn’t they? I’ll have to be honest there. But at the same time I would address that. You know, I would address, I would explain to them why. Er, you know, I’ve told junior doctors sometimes that there’s...er, some of what they do is irritating, they use, er, whatever so many times in the interview or they keep going back to, um, the recap so many times it just drags out and you get irritated, but as long as you’re up front and honest with them about it I don’t think there’s a problem with that.

I: Uh-huh.

A: The worst thing is isn’t it if you mark them down, er, for reasons that they irritate you but don’t tell them why you’re marking them down. That’s the worst thing.

I: Okay. Um, now I, I noticed, again, one of the things you commented on was, was the way that especially the middle doctor was thinking, um, and her ability to sort of put things together. How, how much do you feel able to judge those types of skills from, from watching a performance like this?

A: Um, from actually watching the performance you will start to make a judgment, but then at the end when it’s feedback time you will then explore with them what they were thinking...

I: Okay.

A: ...er, and you’d be prepared to modify...you know, you’d be prepared to mo...you might say to them well, I didn’t think you’d picked that up but now that you’ve explained it to me yes, I understand you had picked that up, well done, you know.

I: Okay.

A: Like it didn’t come across as I watched you and I’m glad I’ve asked you, yes, you know exactly what we’re dealing with there, that’s good. So, yeah, you
would…you do make judgments on whether they’re picking it up or not but you’re prepared to be surprised and pleasantly surprised in the feedback section to the mini CEX.

I: Okay. Okay. So if the final doctor there had turned round to you in the feedback and said well I think this is vasovagal syncope precipitated by stress...

A: Yeah [laughs].

I: ...wha...wha...wha...how would that have changed things?

A: Yes. Er, it would have changed things, yeah. Um, I would have said well, okay, yes, you’ve got there but I, I believe you probably got there by luck rather than judgment. Um, tha...she was so bad that I don’t think I could...you know, I wouldn’t be put off by the fact...I wouldn’t change my mark for her even if she came out with the right thing, I promise you...

I: Right.

A: ...because she was so bad. But I suppose, um, yeah, you will be changed a little bit by their feedback, but, again, it’s a matter of judgment. I would still be able to say to her well, you know, the attitude you gave to the patient was shocking and you never explored the patient’s fears and concerns or addressed them, so, yeah, okay, you got the diagnosis but I think you stumbled on that by luck and there’s still an awful lot of improvement needed.

I: Okay. What stage in the videos do you think you start to make your mind up about the scores you’re going to give?

A: Oh very quickly. Yeah, very quickly. Yeah.

I: How much did those scores then change, as the video progresses?

A: Impossible to answer that isn’t it? I suppose if they start badly and then suddenly bring it together at the end you, you would modify it. I mean I’m...you, you have started to make a judgment almost immediately but th...you will change that judgment. I’m not gonna stick by that throughout if something quite obviously changes. But yes, it’s the same when you...membership exam. If you talk to the membership guys you, you can almost tell within 30 seconds who’s gonna pass and who’s gonna fail.

I: Mm.

A: Occasionally you get surprised. But, yeah, I can’t deny that fact.

I: To, to...

A: You’re making judgment almost immediately.

I: Just something that you recognise as being...

A: I think so.

I: ...one thing or the other.

A: Yeah.

I: Okay. Um, you described yourself as a bit of a dove. Um, or maybe you didn’t actually. What you said is I’m being dove-ish, which maybe means something different.

A: No, no. I, I...in terms of scoring and in feedback to junior doctors I, I think I am a bit of a dove, yeah.
I: Uh-huh.
A: Yeah.

I: I know that MRCP examiners get a hawk/dove index.
A: Yes.

I: Is that what you base it upon?
A: Um, my judgments in myself.

I: Mm.
A: I’m a bit of a hawk in MRCP funnily enough.

I: Are you?
A: Yeah. Um...

I: But you think you take a different approach with...
A: Yeah, with the younger ones, yeah. With the F1s.

I: Is it a conscious thing? Is that something you’ve thought about, or is that just how it’s worked itself out?
A: I don’t know. I think that’s just the way it’s worked itself out I think.

I: Okay.
A: Yeah.

I: And, and is that just a suspicion or have you had any way of...
A: A suspicion, yeah.

I: Okay.
A: It’s a suspicion, yeah.

I: Okay. Have you had any training in these workplace based assessments?
A: Yes.

I: Yeah.
A: We were...I was with John Miles when they first got set up...

I: Yeah.
A: ...and I think he used us as a sort of index site or whatever. I remember going on...it’s a long time I know, Peter, but I remember he came up here and we, we looked at videos, we went through it, we judged it. So, yeah, I’ve done a lot of face to face training with the deanery and obviously done a lot over the years, yeah.

I: Do you feel that those videos helped you to develop a clearer idea of the standard you were marking the...
A: Yeah. I quite liked those videos.

I: You did.
A: Yeah, I did. Well they were so graphically obvious weren’t they...

I: Yeah.
A: ...in some respects and it can often be a bit more subtle than that. But, yeah, I did like those.

I: Mm.

A: And, er, I must admit, I’ll have to be absolutely honest with you, having seen the first girl I will probably have higher expectations now of the Fs...

I: Mm.

A: ...because, um, her performance was near perfect. In fact y...you know, if she was in the history taking section of the MRCP paces she’d have passed because...

I: Mm.

A: ...it was so beautifully done, yeah.

I: Mm. Mm. Mm. Okay. Um, just for the tape, because this is anonymous, could I get you just to state your clinical speciality?

A: Er, acute medicine.

I: Okay.

A: Er, consultant physician with an interest in acute medicine.

I: Okay. It...it’s interesting the, the only other acute medicine consultant I’ve, I’ve done this project with thought that they observed trainees a lot more than other specialities.

A: Yes.

I: Do, do you think that’s in your experience as well?

A: That is absolutely tru...absolutely true.

I: Just because of the way the unit works.

A: It’s the way it works. They are taking histories, doing examinations and reporting back throughout the day.

I: Mm.

A: It’s not on a twice daily ward round. All through the day...

I: Mm. Mm.

A: ...they’re reporting back to you, so yeah.

I: Do you see them taking the history or do...

A: Yeah.

I: ...they just report it to you?

A: Well, normally they’ll report it but...

I: Mm.

A: ...because you’ve got such, um, er, um, an opportunity...

I: Mm.

A: ...then often what you can do is if you know you’re gonna have to see that patient anyway as the consultant...

I: Mm.
A: ...why not sit with the kid while she does the, um, the clerking...
I: Yeah.
A: ...and then I’m, I’m able to give my senior opinion and do a mini CEX at the same time...
I: Yeah, that’s fine.
A: ...which is magnificent, you know.
I: Yeah. Yeah, that sounds...
A: And then now we’re introducing now...Doctor ______, who’s one of our new consultant geriatricians, he’s very much into tandem clerking.
I: Uh-huh.
A: I don’t know if you’ve heard of this sort of teaching aid.
I: I’ve come across it, yeah.
A: Yeah. What they do is th...the, the house officer or foundation attaches themselves to the registrar...
I: Uh-huh.
A: ...and they go and clerk the patients together and they alternate, so the F1 clerks while the registrar watches and then the next patient the registrar clerks and the F1 watches, so...
I: Oh okay. And they write for each other as they go or...
A: Yeah. Yeah.
I: Yeah. Mm.
A: So it’s tandem clerking.
I: Mm. Very interesting. No, I’ve not come across that.
A: Yeah.
I: Um, okay, I think I’ve covered, er...one more thing actually.
A: Yeah.
I: And, and...just...and that would just be in, in, in your experience of F1s...
A: Yes.
I: ...over the last few years...
A: Yes.
I: ...how typical would each one of those three videos be that you’ve seen?
A: Yes.
I: How much would you have seen a...a...a...any of those?
A: Well, two...
I: Uh-huh.
A: ...is very typical.
I: Okay.
A: You know what I mean? Probably not as bad as that.
I: The, the one in the middle.
A: Yeah.
I: The second video.
A: Yeah. The...
I: Even though you felt it was quite a poor performance.
A: Yeah, she wasn’t great. Yeah, she wasn’t great. Um, well, there were things...alright, let’s talk about it this way. What the F1s tend to be very good at now is the communication. They’re very good at sitting down, introducing themselves and going into an empathetic history taking session. So the second one wasn’t typical because she didn’t allow the patient to speak. They’re usually very good at that. Um, the first one is probably...well, I’ve said this haven’t I, she’s better. Um, so there was somewhere between one and two would have been typical...
I: Okay.
A: ...but three is way off the mark, yeah.
I: And do you think you ever see ones like three?
A: I’ll have to be honest, I don’t think I’ve seen one as bad as three...
I: Okay.
A: ...for ages.
I: Okay.
A: Honestly.
I: No, that’s fine. I...um...wh...
A: Well, you know, but it may be that they put on a bit of an act with me, but she didn’t even s...make any effort knowing that, that she was being mini CEXed. Do you know what I mean? I’m not, I’m not saying that...
I: And you think they do put on an act...
A: Yeah, I do. I’m not saying that maybe...
I: Yeah.
A: ...she wouldn’t be like that when I wasn’t watching, but...
I: Yeah, okay [laughs].
A: ...to actually do that while you’re watching is quite...
I: Yeah, yeah.
A: ...remarkably bad, yeah.
I: Yeah. Er, what about the first one, do you think you see that in real life?
A: Um, no. Not as good as that.
I: No. No.
A: I mean you do occasionally. There are some very good ones, don’t get me wrong. But, but she was almost perfect...
I: Yeah.
A: ...I felt. I...there was very little she did wrong.
I: Yeah.
A: I couldn’t…I don’t think I could have said to her anything that could have improved that. That was just tremendous.
I: Okay. Good.
A: Alright.
I: That’s, er, that’s all I want to cover, so thank you very much.
A: Okay.

Anything especially good?
(blank)

Suggestions for development?
(blank)

Agreed action?
(blank)

*End of recording*
Information Sheet

Assessor Judgements in WPBA

Information for Participants

Background:
Workplace based assessments were introduced with the intention of authentically assessing the actual clinical practice of junior doctors in their normal place of work.

There has been quite a lot of study of various aspects of WPBA — with some mixed results. In many types of assessments (like OSCEs for example), there are very specific descriptors to guide assessors in allocating scores. In WPBA, assessors have to judge what they see compared to “the expected level for completion of stage”. However, we know very little about how assessors reach these judgements, or what they are based on.

The purpose of this research is to understand more about the way that assessors make judgements about foundation doctors performance. The research is being conducted as part of a course of study towards a PhD by one of the researchers.

How will the study work?
The study is internet based; as such it is very flexible as you can complete it at any time or anywhere that is convenient. The study will use an experiment to test one particular aspect of the way that assessors make judgements. The study works by showing participants a number of short videos of foundation doctors performing tasks, and asking them to judge and score the performances.

What would taking part involve?
If you decide to take part, please follow the link in the email, and enter the study password. You will then be asked to confirm some details and complete an on-line consent form.

Following this you will be supplied with brief further instructions about the task. The website will allocate you to one of the two study groups. You will then be shown 6 short videos (approximately 3-4 minutes each) of a foundation year 1 doctor performing a clinical task. You’ll be asked to make a judgement on the performance that you see as if it were a MiniCEX assessment. After each video you will be asked to give a score before moving on to the next video. After you’ve seen all 6 videos, you will be asked for some demographic details.
You will only do this once. The whole process should take roughly 30 minutes, and can be completed at any time that is convenient. If you decide to take part, please do it all in one go, and try to avoid being distracted.

What will happen to my data?
The group you were allocated to, the scores you give and your demographic data will all be stored. This will be anonymous.

Do I have to take part?
No. Participation is completely voluntary.

What if I want to withdraw?
You can withdraw any time before completing the study. After completion your data cannot be withdrawn as it is anonymous.

Can I have a copy of the results if I participate?
Yes. You’ll be asked to enter an email address into the website. This will be stored separately from your data, so the data you provide will still be anonymous.

Can anyone take part?
All participants will be consultant physicians, who have been consultants for at least 2 years, and who estimate they do at least 5 MiniCEXs a year. They should be happy to assess a general medical topic. Please do not ask junior colleagues to participate on your behalf.

Regardless of whether or not you choose to participate, if you know any colleagues that you think would be interested in participating in this study, please pass the invitation email to them.

Further information:
We have been a little vague about the purpose of the experiment so far. This is because prior knowledge of the study’s research question might bias your responses. If you decide to participate, we would be happy to give you further information afterwards, and to give you a copy of the study’s results.

For further information please contact:

Peter Yeates
Research Fellow in medical education
ATR4, Education Research Centre
UHSM, Southmoor Road
Wythenshawe
Manchester M23 9LT
Peter.yeates@manchester.ac.uk
Tel: 07932 675066
Appendix 8: Data tables from study 2

Descriptive data

<table>
<thead>
<tr>
<th>Group</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Av_BordSc</td>
<td>Mean</td>
<td>2.7255</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>.15473</td>
</tr>
<tr>
<td></td>
<td>95% Confidence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interval for Mean</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>2.4017</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>3.0493</td>
</tr>
<tr>
<td></td>
<td>5% Trimmed Mean</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>2.6950</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>.479</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>.69196</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.58</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>3.89</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2.31</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>.029</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>-.1176</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>3.3989</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>.11934</td>
</tr>
<tr>
<td></td>
<td>95% Confidence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interval for Mean</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>3.1499</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>3.6478</td>
</tr>
<tr>
<td></td>
<td>5% Trimmed Mean</td>
<td>3.4005</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3.5000</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>.299</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>.54689</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>4.39</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>-.374</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>-.390</td>
</tr>
</tbody>
</table>

Repeated measures ANOVA tests relating to group effect

Tests of Within-Subjects Effects

Measure: MEASURE_1

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>Sphericity Assumed</td>
<td>20.397</td>
<td>2</td>
<td>10.198</td>
<td>40.348</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Greenhouse-Geisser</td>
<td>20.397</td>
<td>1.845</td>
<td>11.057</td>
<td>40.348</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Huynh-Feldt</td>
<td>20.397</td>
<td>1.982</td>
<td>10.292</td>
<td>40.348</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Lower-bound</td>
<td>20.397</td>
<td>1.000</td>
<td>20.397</td>
<td>40.348</td>
<td>.000</td>
</tr>
<tr>
<td>Cases * Group</td>
<td>Sphericity Assumed</td>
<td>2.509</td>
<td>2</td>
<td>1.254</td>
<td>4.963</td>
<td>.009</td>
</tr>
</tbody>
</table>
Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1152.882</td>
<td>1</td>
<td>1152.882</td>
<td>994.356</td>
<td>.000</td>
<td>.962</td>
</tr>
<tr>
<td>Group</td>
<td>13.906</td>
<td>1</td>
<td>13.906</td>
<td>11.994</td>
<td>.001</td>
<td>.235</td>
</tr>
<tr>
<td>Error</td>
<td>45.218</td>
<td>39</td>
<td>1.159</td>
<td>.518</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Linear regression examining different sources of variance on scores

Group, Stringency index (described here as “Av_z_score”), duration of consultancy and gender

ANOVA(e)
Regression

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>2.213</th>
<th>7.328</th>
<th>.000(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>Change Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R Square Change</td>
</tr>
<tr>
<td>1</td>
<td>.052a</td>
<td>.003</td>
<td>-.023</td>
<td>.71021</td>
<td>.003</td>
</tr>
<tr>
<td>2</td>
<td>.052b</td>
<td>.003</td>
<td>-.050</td>
<td>.71948</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>.420c</td>
<td>.176</td>
<td>.110</td>
<td>.66260</td>
<td>.174</td>
</tr>
<tr>
<td>4</td>
<td>.670d</td>
<td>.449</td>
<td>.388</td>
<td>.54955</td>
<td>.272</td>
</tr>
</tbody>
</table>

Tests of interaction between group and SI

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1145.614</td>
<td>1</td>
<td>1145.614</td>
<td>1200.755</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>13.240</td>
<td>1</td>
<td>13.240</td>
<td>13.877</td>
<td>.001</td>
</tr>
<tr>
<td>HiLoTen_Group</td>
<td>9.401</td>
<td>1</td>
<td>9.401</td>
<td>9.853</td>
<td>.003</td>
</tr>
<tr>
<td>Group * HiLoTen_Group</td>
<td>.418</td>
<td>1</td>
<td>.418</td>
<td>.438</td>
<td>.512</td>
</tr>
<tr>
<td>Error</td>
<td>35.301</td>
<td>37</td>
<td>.954</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hierarchical regression, Stringency index 1st, group 2nd
Appendix 9: Participant information sheet for study 3

Information Sheet

Assessor Judgements in WPBA

Information for Participants

Background:
Workplace based assessments were introduced with the intention of authentically assessing the actual clinical practice of junior doctors in their normal place of work.

There has been quite a lot of study of various aspects of WPBA – with some mixed results. In many types of assessments (like OSCEs for example), there are very specific descriptors to guide assessors in allocating scores. In WPBA, assessors have to judge what they see compared to “the expected level for completion of stage”. However, we know very little about how assessors reach these judgements, or what they are based on.

The purpose of this research is to understand more about the way that assessors make judgements about foundation doctors’ performance. The research is being conducted as part of a course of study towards a PhD by one of the researchers.

How will the study work?
The study is internet based; as such it is very flexible as you can complete it at any time or in any place that is convenient. The study will use an experiment to test a few specific aspects of the way that assessors make judgements. The study works by showing participants a number of short videos of foundation doctors performing tasks, and asking them to judge and score the performances. We’ll also collect ratings of two other parameters after each video.

What would taking part involve?
If you decide to take part, please follow the link in the email, and enter the study password. You will then be asked to confirm some details and complete an on-line consent form.

Following this you will be supplied with brief further instructions about the task. The website will allocate you to one of the two study groups. You will then be shown 6 short videos (approximately 3-4 minutes each) of a foundation year 1 doctor performing a clinical task. You’ll be asked to make a judgement on the performance that you see as if it were a
MiniCEX assessment. After each video you will be asked to provide three different responses. These will include:

- scores for the performance
- a rating of your confidence in your scores
- an estimate of the proportion of foundation doctors that you think would have performed either better or worse on the task if they had undertaken it.

You’ll be asked to do this for each video before moving onto the next one. After you’ve seen all 6 videos, you will be asked for some demographic details.

You will only do this once. The whole process should take roughly 30 minutes, and can be completed at any time that is convenient. If you decide to take part, please do it all in one go, and try to avoid being distracted.

**What will happen to my data?**
The group you were allocated to, the scores you give and your demographic data will all be stored. These will be anonymous.

**Do I have to take part?**
No. Participation is completely voluntary.

**What if I want to withdraw?**
You can withdraw any time before completing the study. After completion your data cannot be withdrawn as it is anonymous.

**Can I take part if I’ve been involved in previous studies?**
Yes. You’ll be asked to indicate that you have been involved in previous studies, and this may alter the group that you’re allocated to. This data will be kept with your demographic data.

**Can I have a copy of the results if I participate?**
Yes. You’ll be asked to enter an email address into the website. This will be stored separately from your data, so the data you provide will still be anonymous.

**Can anyone take part?**
All participants will be consultant physicians, who have been consultants for at least 2 years, and who estimate they do at least 5 MiniCEXs a year. They should be happy to assess a general medical topic. Please do not ask junior colleagues to participate on your behalf.

Regardless of whether or not you choose to participate, if you know any colleagues that you think would be interested in participating in this study, please pass the invitation email to them.

**Further information:**
We have been a little vague about the purpose of the experiment so far. This is because prior knowledge of the study’s research question might bias your responses. If you decide to participate, we would be happy to give you further information afterwards, and to give you a copy of the study’s results.

For further information please contact:

Peter Yeates
Research Fellow in medical education
Appendix 10: Data tables from study 3

Descriptive data relating to scores

<table>
<thead>
<tr>
<th>Group</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>2.9660</td>
</tr>
<tr>
<td>Score</td>
<td>95% Confidence Interval for Mean</td>
<td>2.7567</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5% Trimmed Mean</td>
<td>2.9523</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3.0000</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>1.684</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.29756</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>5.00</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>-.115</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>-1.176</td>
</tr>
</tbody>
</table>

|       | Mean              | 3.4955     | .11565     |
| Score | 95% Confidence Interval for Mean | 3.2669 | 3.7241 |
|       | Lower Bound       |            |            |
|       | Upper Bound       |            |            |
|       | 5% Trimmed Mean   | 3.4847     |            |
|       | Median            | 3.6667     |            |
|       | Variance          | 1.926      |            |
|       | Std. Deviation    | 1.38781    |            |
|       | Minimum           | 1.00       |            |
|       | Maximum           | 6.00       |            |
|       | Range             | 5.00       |            |
|       | Interquartile Range | 2.50    |            |
|       | Skewness          | .017       | .202       |
|       | Kurtosis          | -1.087     | .401       |
Boxplot of data from study 3: Min, Max, Median and inter-quartile ranges
Tests of group differences, scores as dependent variable

Mauchly’s Test of Sphericity

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>159.798</td>
<td>2</td>
<td>79.899</td>
<td>328.114</td>
<td>.000</td>
</tr>
<tr>
<td>Level * Group</td>
<td>159.798</td>
<td>1.690</td>
<td>94.532</td>
<td>328.114</td>
<td>.000</td>
</tr>
<tr>
<td>Level * Group</td>
<td>159.798</td>
<td>1.784</td>
<td>89.575</td>
<td>328.114</td>
<td>.000</td>
</tr>
<tr>
<td>Level * Group</td>
<td>159.798</td>
<td>1.000</td>
<td>159.798</td>
<td>328.114</td>
<td>.000</td>
</tr>
<tr>
<td>Level * Group</td>
<td>.757</td>
<td>2</td>
<td>.379</td>
<td>1.555</td>
<td>.217</td>
</tr>
<tr>
<td>Level * Group</td>
<td>.757</td>
<td>1.690</td>
<td>.448</td>
<td>1.555</td>
<td>.219</td>
</tr>
<tr>
<td>Level * Group</td>
<td>.757</td>
<td>1.784</td>
<td>.425</td>
<td>1.555</td>
<td>.219</td>
</tr>
<tr>
<td>Level * Group</td>
<td>.757</td>
<td>1.000</td>
<td>.757</td>
<td>1.555</td>
<td>.219</td>
</tr>
<tr>
<td>Level</td>
<td>22.890</td>
<td>94</td>
<td>.244</td>
<td>.487</td>
<td></td>
</tr>
<tr>
<td>Level * Group</td>
<td>22.890</td>
<td>79.449</td>
<td>.288</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level * Group</td>
<td>22.890</td>
<td>83.846</td>
<td>.273</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level</td>
<td>22.890</td>
<td>47</td>
<td>1.051</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tests of Between-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1533.703</td>
<td>1</td>
<td>1533.703</td>
<td>1458.934</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>10.299</td>
<td>1</td>
<td>10.299</td>
<td>9.797</td>
<td>.003</td>
</tr>
<tr>
<td>Error</td>
<td>49.409</td>
<td>47</td>
<td>1.051</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Estimates**

<table>
<thead>
<tr>
<th>Level</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.581</td>
<td>.077</td>
<td>4.427</td>
<td>4.735</td>
</tr>
<tr>
<td>2</td>
<td>3.069</td>
<td>.120</td>
<td>2.828</td>
<td>3.310</td>
</tr>
<tr>
<td>3</td>
<td>2.042</td>
<td>.106</td>
<td>1.829</td>
<td>2.255</td>
</tr>
</tbody>
</table>

**4. Group * Level**

<table>
<thead>
<tr>
<th>Group</th>
<th>Level</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4.312</td>
<td>.107</td>
<td>4.097</td>
<td>4.528</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.718</td>
<td>.168</td>
<td>2.381</td>
<td>3.055</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.867</td>
<td>.148</td>
<td>1.569</td>
<td>2.165</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4.850</td>
<td>.109</td>
<td>4.630</td>
<td>5.070</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.419</td>
<td>.171</td>
<td>3.075</td>
<td>3.764</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.217</td>
<td>.151</td>
<td>1.913</td>
<td>2.521</td>
</tr>
</tbody>
</table>

**Adding Duration of consultancy “ConsYrs” as a covariate**

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>506.438</td>
<td>1</td>
<td>506.438</td>
<td>496.682</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>5.193</td>
<td>1</td>
<td>5.193</td>
<td>5.093</td>
<td>.029</td>
</tr>
<tr>
<td>ConsYrs</td>
<td>3.372</td>
<td>1</td>
<td>3.372</td>
<td>3.307</td>
<td>.076</td>
</tr>
<tr>
<td>Group * ConsYrs</td>
<td>.841</td>
<td>1</td>
<td>.841</td>
<td>.825</td>
<td>.369</td>
</tr>
<tr>
<td>Error</td>
<td>45.884</td>
<td>45</td>
<td>1.020</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Adding Gender as a factor**

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1457.845</td>
<td>1</td>
<td>1457.845</td>
<td>1330.855</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>9.427</td>
<td>1</td>
<td>9.427</td>
<td>8.606</td>
<td>.005</td>
</tr>
<tr>
<td>Gender</td>
<td>.091</td>
<td>1</td>
<td>.091</td>
<td>.083</td>
<td>.775</td>
</tr>
<tr>
<td>Group * Gender</td>
<td>.025</td>
<td>1</td>
<td>.025</td>
<td>.023</td>
<td>.881</td>
</tr>
<tr>
<td>Error</td>
<td>49.294</td>
<td>45</td>
<td>1.095</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Adding Prior participant “PrevPart” as a factor**

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1457.845</td>
<td>1</td>
<td>1457.845</td>
<td>1330.855</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>9.427</td>
<td>1</td>
<td>9.427</td>
<td>8.606</td>
<td>.005</td>
</tr>
<tr>
<td>Gender</td>
<td>.091</td>
<td>1</td>
<td>.091</td>
<td>.083</td>
<td>.775</td>
</tr>
<tr>
<td>Group * Gender</td>
<td>.025</td>
<td>1</td>
<td>.025</td>
<td>.023</td>
<td>.881</td>
</tr>
<tr>
<td>Error</td>
<td>49.294</td>
<td>45</td>
<td>1.095</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Adding Number of estimated annual Mini-CEX's assessed on any grade of trainee "CEXAny" as a factor

Measure: MEASURE_1
Transformed Variable: Average

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>304.295</td>
<td>1</td>
<td>304.295</td>
<td>292.330</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>.435</td>
<td>1</td>
<td>.435</td>
<td>.418</td>
<td>.521</td>
</tr>
<tr>
<td>CEXAny</td>
<td>.602</td>
<td>1</td>
<td>.602</td>
<td>.578</td>
<td>.451</td>
</tr>
<tr>
<td>Group * CEXAny</td>
<td>.518</td>
<td>1</td>
<td>.518</td>
<td>.497</td>
<td>.484</td>
</tr>
<tr>
<td>Error</td>
<td>46.842</td>
<td>45</td>
<td>1.041</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adding Number of estimated annual Mini-CEX's assessed on F1 doctors "CEXF1" as a factor

Measure: MEASURE_1
Transformed Variable: Average

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>519.199</td>
<td>1</td>
<td>519.199</td>
<td>475.017</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>3.204</td>
<td>1</td>
<td>3.204</td>
<td>2.931</td>
<td>.094</td>
</tr>
<tr>
<td>CEXF1</td>
<td>.100</td>
<td>1</td>
<td>.100</td>
<td>.091</td>
<td>.764</td>
</tr>
<tr>
<td>Group * CEXF1</td>
<td>.001</td>
<td>1</td>
<td>.001</td>
<td>.001</td>
<td>.972</td>
</tr>
<tr>
<td>Error</td>
<td>49.185</td>
<td>45</td>
<td>1.093</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tests of group differences, PB as dependent variable

Tests of Within-Subjects Effects

Measure: MEASURE_1

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Sphericity Assumed</td>
<td>2</td>
<td>42932.199</td>
<td>287.908</td>
<td>.000</td>
</tr>
<tr>
<td>Greenhouse-Geisser</td>
<td>85864.399</td>
<td>1.880</td>
<td>45667.912</td>
<td>287.908</td>
<td>.000</td>
</tr>
<tr>
<td>Huynh-Feldt</td>
<td>85864.399</td>
<td>2.000</td>
<td>42932.199</td>
<td>287.908</td>
<td>.000</td>
</tr>
<tr>
<td>Lower-bound</td>
<td>85864.399</td>
<td>1.000</td>
<td>85864.399</td>
<td>287.908</td>
<td>.000</td>
</tr>
<tr>
<td>Sphericity Assumed</td>
<td>1367.003</td>
<td>2</td>
<td>683.501</td>
<td>4.584</td>
<td>.13</td>
</tr>
<tr>
<td>Greenhouse-Geisser</td>
<td>1367.003</td>
<td>1.880</td>
<td>727.055</td>
<td>4.584</td>
<td>.14</td>
</tr>
<tr>
<td>Huynh-Feldt</td>
<td>1367.003</td>
<td>2.000</td>
<td>683.501</td>
<td>4.584</td>
<td>.13</td>
</tr>
<tr>
<td>Lower-bound</td>
<td>1367.003</td>
<td>1.000</td>
<td>1367.003</td>
<td>4.584</td>
<td>.38</td>
</tr>
<tr>
<td>Error</td>
<td>Sphericity Assumed</td>
<td>13718.848</td>
<td>92</td>
<td>149.118</td>
<td></td>
</tr>
<tr>
<td>Greenhouse-Geisser</td>
<td>13718.848</td>
<td>86.489</td>
<td>158.620</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huynh-Feldt</td>
<td>13718.848</td>
<td>92.000</td>
<td>149.118</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower-bound</td>
<td>13718.848</td>
<td>46.000</td>
<td>298.236</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Tests of Between-Subjects Effects

Measure: MEASURE_1  
Transformed Variable: Average

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>365452.563</td>
<td>1</td>
<td>365452.563</td>
<td>835.904</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>7001.522</td>
<td>1</td>
<td>7001.522</td>
<td>16.015</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>20110.935</td>
<td>46</td>
<td>437.194</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Estimates

<table>
<thead>
<tr>
<th>Level</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
</tr>
<tr>
<td>1</td>
<td>18.248</td>
<td>1.660</td>
<td>14.905</td>
<td>21.590</td>
</tr>
<tr>
<td>2</td>
<td>55.570</td>
<td>2.794</td>
<td>49.945</td>
<td>61.194</td>
</tr>
<tr>
<td>3</td>
<td>77.446</td>
<td>2.187</td>
<td>73.043</td>
<td>81.848</td>
</tr>
</tbody>
</table>

### 4. Group * Level

Measure: MEASURE_1

<table>
<thead>
<tr>
<th>Group</th>
<th>Level</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>22.800</td>
<td>2.299</td>
<td>18.173</td>
<td>27.427</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>66.900</td>
<td>3.868</td>
<td>59.113</td>
<td>74.687</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>82.500</td>
<td>3.028</td>
<td>76.405</td>
<td>88.595</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>13.696</td>
<td>2.397</td>
<td>8.871</td>
<td>18.520</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>44.239</td>
<td>4.033</td>
<td>36.121</td>
<td>52.357</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>72.391</td>
<td>3.157</td>
<td>66.037</td>
<td>78.746</td>
</tr>
</tbody>
</table>

### Tests of the influence of confidence on group effect:

Means scores for low (0) and high (1) confidence participants

<table>
<thead>
<tr>
<th>ConfHiLo</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.235</td>
<td>.123</td>
<td>2.988</td>
<td>3.482</td>
</tr>
<tr>
<td>1</td>
<td>3.216</td>
<td>.120</td>
<td>2.975</td>
<td>3.458</td>
</tr>
</tbody>
</table>

### Interaction of Group and Confidence:

Tests of Between-Subjects Effects

Measure: MEASURE_1  
Transformed Variable: Average
### Group by levels scores in low confidence participants:

4. Group * Level

<table>
<thead>
<tr>
<th>Measure: MEASURE_1</th>
<th>Group</th>
<th>Level</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>4.433</td>
<td>.145</td>
<td>4.132</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2.853</td>
<td>.176</td>
<td>2.488</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1.879</td>
<td>.152</td>
<td>1.564</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>4.926</td>
<td>.157</td>
<td>4.599</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3.245</td>
<td>.191</td>
<td>2.850</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>2.077</td>
<td>.165</td>
<td>1.735</td>
</tr>
</tbody>
</table>

### Tests of Between-Subjects Effects

Measure: MEASURE_1  
Transformed Variable: Average

### Group by levels scores in high confidence participants:

4. Group * Level

<table>
<thead>
<tr>
<th>Measure: MEASURE_1</th>
<th>Group</th>
<th>Level</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>4.182</td>
<td>.159</td>
<td>3.853</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2.573</td>
<td>.289</td>
<td>1.975</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1.855</td>
<td>.261</td>
<td>1.315</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>4.786</td>
<td>.153</td>
<td>4.469</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3.567</td>
<td>.278</td>
<td>2.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>2.335</td>
<td>.251</td>
<td>1.816</td>
</tr>
</tbody>
</table>

### Tests of Between-Subjects Effects

Measure: MEASURE_1  
Transformed Variable: Average
<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>774.582</td>
<td>1</td>
<td>774.582</td>
<td>557.772</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>8.982</td>
<td>1</td>
<td>8.982</td>
<td>6.468</td>
<td>.018</td>
</tr>
<tr>
<td>Error</td>
<td>31.940</td>
<td>23</td>
<td>1.389</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Slimy Fish