

Assessor grades and comments: private thoughts and public judgements

Catherine Elizabeth Scarff

ORCID identifier: 0000-0002-7717-950X

Doctor of Philosophy

January 2020

Faculty of Dentistry, Medicine and Health Sciences

Department of Medical Education

Submitted in total fulfilment for the degree of PhD

Abstract

Assessment of medical trainees' performance in the workplace aims to provide them with accurate and meaningful information and guidance on their learning and developing competence. However, in practice, such goals aren't always achieved. Sometimes assessors find it difficult to deliver clear and consistent assessment messages to a trainee, especially when the information or judgement they have to give is negative. While this can occur for many reasons, including disagreement with or uncertainty about assessment processes, the MUM effect — the widespread human tendency to keep mum about undesirable messages — may also have relevance for the situation. With reference to this framework, this thesis explores how and why reluctance to deliver negative assessment messages manifests in a medical specialty training setting in Australia.

Literature reviews on the MUM effect and trainee perspectives of assessment messages informed the design of a mixed methods study which explores the MUM effect in this context. The study involved four parts:

- a questionnaire study of assessor self-reports of discomfort and MUM behaviours in assessment;
- a questionnaire study of trainee perspectives of MUM behaviours by their assessors and their views of the clinical performance assessments;
- a review of a sample of previously submitted assessment forms comparing the messages sent by numerical ratings with those by written comments; and
- an interview study of assessors to further understandings of their experiences with and perspectives of these assessment formats.

The findings show that reluctance to deliver negative assessment messages — which can result in failure to give feedback, failure to fail and grade inflation — are real and continuing issues in medical education. The MUM effect offers one explanation for their persistence, despite the many methods which have been employed to date to address them. The study shows how the MUM effect permits an expanded view of the problem and that assessor reluctance can lead to behaviours beyond the commonly reported failure to fail and grade inflation. These include behaviours such as delay,

avoidance and distortion of assessment information. Further, the results show that reluctance can affect the comments part of an assessment in addition to the ratings, which have been the main focus to date. This study reveals the many pressures and dilemmas that assessors face in their role and in particular, that the amount of discomfort they experience can potentially affect their assessment behaviours and result in MUMing. This work also shows that trainees are aware that their assessors sometimes keep mum, meaning the judgement delivered may differ from the assessor's private thoughts on their performance. Potential solutions are seen to be multifactorial and include addressing perceptions about "failure" in clinical performance assessments and the responsibility that assessors feel for the assessment decisions.

Declaration

This thesis comprises only my original work towards the award of PhD except where indicated in the preface.

Due acknowledgement has been made in the text to all other material used.

The thesis is fewer than the maximum word limit in length, exclusive of tables, bibliographies and appendices.

Signed:

Full name:

Date:

Preface – noting collaborations and contributions to authorship

- (i) *Description of work towards the thesis that was carried out in collaboration with others, indicating the nature and proportion of the contribution of others and in general terms the portions of the work which the student claims as original:*

The work towards the thesis was carried out by the candidate under the supervision of Stephen Trumble, Neville Chiavaroli and Margaret Bearman.

The candidate claims the work as original with the exception of Chapter 4, Trainees' perspectives of assessment messages: a narrative systematic review. This is a multi-author publication and all supervisors and the candidate contributed as follows. All contributed to the design of the work and the student was involved in all stages of the work. All contributed to data analysis. Margaret Bearman contributed to the data synthesis and to the writing of the paper. Stephen Trumble contributed to the writing and editing of the paper. All authors critically reviewed the paper and approved the final version for publication.

- (ii) *A description of work towards the thesis that has been submitted for other qualifications:*

This work has not been submitted for any other qualifications.

- (iii) *A description of work towards the thesis that was carried out prior to enrolment in the degree:*

No work towards the thesis was carried out prior to enrolment in the degree.

- (iv) *Whether any third-party editorial assistance was provided in preparation of the thesis and whether the persons providing this assistance are knowledgeable in the academic discipline of the thesis:*

Third-party editorial assistance was provided in preparation of the thesis in accordance with University guidelines by a person who is not knowledgeable in the academic discipline of the thesis. This assistance was limited to editorial intervention in accordance with [The Australian Standards for Editing Practice](#) as follows:

- (a) Standard D: Language and Illustrations; and
- (b) Standard E: Completeness and Consistency.

- (v) *The contributions of all persons involved in any multi-authored publications or articles in preparation included in the thesis:*

All supervisors contributed to Chapter 4, Trainees' perspectives of assessment messages: a narrative systematic review, to a total of 30%. The candidate's contribution was 70%.

- (vi) *The publication status of all chapters presented in article format using descriptors:*

Chapter 4, Trainees' perspectives of assessment messages: a narrative systematic review, is published in Medical Education – Scarff, C. E., Bearman, M., Chiavaroli, N., & Trumble, S. (2019). Trainees' perspectives of assessment messages: a narrative systematic review. *Medical Education*, 53(3), 221–233. <https://doi.org/10.1111/medu.13775>.

- (vii) *An acknowledgement of all sources of funding:*

I gratefully acknowledge being the recipient of a Research Training Program Scholarship (formerly known as Australian Postgraduate Award) for the duration of the project.

Acknowledgements

There are so many people that have helped me in this PhD journey. Firstly, to my supervisors — Steve Trumble (Principal Supervisor), Neville Chiavaroli (CoSupervisor) and Margaret Bearman (External Supervisor) — “thank you” seems so inadequate. But nevertheless, thank you for your never-ending guidance, support and encouragement. You have each brought a unique set of expertise, experiences and perspectives which have helped me to develop this work. Thank you for your energy and commitment and I hope one day to be able to provide the same level of support to others that you have shown to me.

To my advisory panel—Robyn Woodward-Kron (Committee Chair), Robert Corderoy (Director, Educational Innovation, ACD) and Sam Zagarella (Fellow, ACD) — my heart-felt thanks for your guidance along the way.

To staff of the DME — especially Joanne Rao and Michelle Grainger — thank you for all the assistance you have provided. To Kate Reid — I am so grateful for your support and guidance with statistics, review, suggestions and mentorship.

To staff of the University of Melbourne — especially Sue Finch and all at the Statistical Consulting Centre for statistical support, Patrick Condron and the staff of the Brownless and Baillieu libraries for assistance with the systematic review, and the Academic Skills teams.

To the ACD — my thanks to Tim Wills, CEO for College’s permission to conduct the study and for your encouragement with this project, and especially to Brett O’Neill, Tony Moore and all at the ACD for practical support (including answering my many questions, providing the SITA summary forms).

My thanks to Jo Martin, Jo Martin Illustration, for the elephant illustrations and Sandra Dexter for editing assistance.

Finally, to my family. Thank you Christopher for all that you have done to help me along the way, including putting your own studies on hold (I promise I won’t do any more study!). And to our girls, Molly, Olivia and Matilda — thank you your support, encouragement and the sacrifices that you have made.

Table of contents

Abstract	iii
Declaration	v
Preface – noting collaborations and contributions to authorship.....	vi
Acknowledgements	viii
Table of contents	ix
List of tables	xiv
List of figures	xviii
List of abbreviations	xx
List of third party copyright material	xxii
CHAPTER 1 Introduction.....	1
Motivation for the study.....	2
Purpose and importance of clinical performance assessments	2
Challenges to clinical performance assessments	4
Statement of problem	5
Statement of the aim and scope of the study.....	6
Significance of the study.....	7
Overview of the study	8
Summary	9
CHAPTER 2 Background literature: work-based assessment in medical education	11
Introduction.....	12
Purposes of assessment	12
Assessment formats.....	14
Challenges with clinical performance assessments	18
Issues with assessment outcomes – problems in delivering the outcome.....	20
Reasons for the under-reporting of poorly performing medical trainees.....	24
Reluctance to deliver a negative assessment message.....	30
The elephant in the room	33
Summary	34
CHAPTER 3 The MUM effect and its relevance to assessment in medical education	35
Introduction.....	36
Background	36
<i>Manifestations of the MUM effect</i>	<i>39</i>
<i>Why does the MUM effect occur?.....</i>	<i>40</i>
<i>What can influence the way MUM manifests?</i>	<i>42</i>
<i>What does the MUM effect mean for medical education?.....</i>	<i>47</i>
Conclusion	50

CHAPTER 4 Trainees' perspectives of assessment messages: a narrative systematic review	51
Introduction	52
Abstract	53
<i>Purpose</i>	53
<i>Methods</i>	53
<i>Results</i>	53
<i>Conclusions</i>	54
Keywords	54
Introduction	54
Methods	56
<i>Definitions and explanations</i>	56
<i>Search strategy and information sources</i>	56
<i>Inclusion and exclusion criteria</i>	60
<i>Data extraction</i>	60
<i>Data analysis and synthesis</i>	61
Results	61
<i>Study Characteristics</i>	61
<i>Thematic analysis</i>	62
Discussion	70
<i>Summary of findings</i>	70
<i>Implications for practice</i>	71
<i>Limitations</i>	73
Conclusion	74
Implications of this study	74
Summary	75
CHAPTER 5 Research Methods	77
Introduction	78
Theoretical framework	78
Methods of data collection	80
<i>Questionnaire – assessors</i>	81
<i>Questionnaire – trainees</i>	82
<i>Comparison of comments to ratings in the Summative In-Training Assessment (SITA)</i>	82
<i>Interviews</i>	82
Setting for the study	83
Participants	85
Limitations of the methods	86
Ethical considerations	87
<i>Informed consent</i>	87
<i>Recruitment</i>	87
<i>Risk concern</i>	88
<i>Funding</i>	89
<i>Conflicts of interest</i>	90
Analysis of data	90

Relationship of studies to each other	91
Summary	91
CHAPTER 6 <i>To provide an honest rating of a poorly performing trainee is difficult: Assessor self-reports of discomfort with clinical performance assessments and how this impacts upon assessment behaviours</i>	93
Introduction	94
Background	94
<i>Research questions</i>	96
Methods	96
<i>Development of the questionnaire</i>	97
<i>Distribution</i>	99
<i>Data analysis</i>	99
Results	100
<i>Demographics</i>	100
<i>Performance Appraisal Discomfort Scale (PADS) results</i>	100
<i>MUM behaviours and self-report of failure to fail and grade inflation</i>	104
<i>Free-text comments</i>	123
Discussion	125
<i>Summary and interpretation of results</i>	125
<i>Discussion of results</i>	126
<i>The problem with not delivering the assessment message</i>	130
<i>Future directions</i>	130
<i>Limitations</i>	131
Conclusion	132
CHAPTER 7 <i>I have hardly ever received constructive criticism: Trainee impressions of assessor behaviours and the clinical performance assessments</i>	133
Introduction	134
Background	134
<i>Research questions</i>	136
Methods	137
<i>Development of the questionnaire</i>	137
<i>Distribution</i>	140
<i>Data analysis</i>	140
Results	141
<i>Demographics</i>	141
<i>The MUM effect scale</i>	142
<i>Comparison of trainee and assessor views on the occurrence of MUM behaviours</i>	150
<i>Instructional Feedback Orientation Scale (IFOS)</i>	157
<i>Experiences of clinical performance assessments</i>	163
<i>Comparison of trainee and assessor responses to statements about direct approaches</i> ..	166
<i>Other free-text comments</i>	167
<i>Comparison of trainee IFOS responses to trainee reports of MUM behaviours by their assessors</i>	169

Discussion	173
<i>Summary and interpretation of findings</i>	173
<i>Implications</i>	177
<i>Future directions</i>	177
<i>Limitations</i>	178
Conclusion	179
CHAPTER 8 A review of the use of the summative in-training assessment (SITA).....	181
Introduction	182
Background	182
<i>Research questions</i>	183
Methods.....	184
<i>The forms</i>	184
<i>Coding</i>	185
<i>Data analysis</i>	188
Results.....	188
<i>Review of the use of the SITA summary form</i>	189
<i>Comparison of the numerical ratings and written comments for each SITA</i>	193
<i>Review of forms displaying incorrect criteria for unsatisfactory performance</i>	197
Discussion	198
<i>Summary and interpretation of findings</i>	198
<i>Implications</i>	199
<i>Future directions</i>	202
<i>Limitations</i>	204
Conclusion	205
CHAPTER 9 We probably don't fill them out truthfully enough: Assessors' experiences with clinical performance assessments	207
Introduction	208
Background	208
Methods.....	209
<i>Qualitative approach and research paradigm</i>	209
<i>Research characteristics and reflexivity</i>	210
<i>Context, sampling strategy and participants</i>	210
<i>Transcription and trustworthiness</i>	213
<i>Data analysis</i>	213
Results.....	214
<i>Demographics</i>	214
<i>Synthesis and interpretation</i>	215
Discussion	239
<i>Summary and interpretation of findings</i>	239
<i>Implications</i>	242
<i>Limitations</i>	244
Conclusion	245

CHAPTER 10 Discussion.....	247
Introduction.....	248
Synthesis of findings.....	248
<i>How does reluctance to deliver negative assessment messages manifest?</i>	248
<i>Why does reluctance to deliver negative assessment messages manifest?</i>	253
Implications for practice	267
<i>General considerations</i>	267
<i>Implications for practice for Australian dermatology training</i>	284
Future research agenda – areas for further study	287
Limitations	289
<i>Reflexivity</i>	290
Summary	293
CHAPTER 11 Conclusions	295
REFERENCES	301
APPENDICES.....	331
Appendix A Quality level determination for systematic review articles	332
Appendix B Articles included in the systematic review	334
Appendix C Questionnaire – assessors	344
Appendix D Questionnaire – trainees	355
Appendix E Clinical performance assessment forms	364
Appendix F Interview schedule	376
Appendix G Summary of the SITA process	381
Appendix H Supplementary data – questionnaire – assessors.....	383
Appendix I Supplementary data – questionnaire – trainees.....	390

List of tables

Table 2.1: List of common workplace-based assessments in use (Google scholar 6 Aug 2019)	16
Table 4.1: Search terms used in database searches	57
Table 4.2: Inclusion and exclusion criteria for this systematic review	58
Table 6.1: Mean assessor self-reported level of discomfort and range of ratings by individual feedback giving situation across all respondents.	102
Table 6.2: Assessor self-report of MUM behaviours in relation to giving negative assessment messages to trainees	106
Table 6.3: Number of items in each subscale respondents agreed with by number and percent.....	107
Table 6.4: Reasons given by assessors for not recording their true assessment judgement on the assessment form	113
Table 6.5: Additional actions by assessors who did not record their assessment judgement on the assessment form	114
Table 6.6: Number and percentage of respondents who self-reported being approached by a trainee in relation to a clinical performance assessment result	115
Table 6.7: Percentage of respondents self-reporting being approached by a trainee in relation to a clinical performance assessment result by demographic and other features.....	117
Table 6.8: Comparison of mean MUM scores between those with and without self-report of failure to fail.....	118
Table 6.9: Respondent mean PADS score compared with agreement with at least one MUM statement by subscale items	119
Table 6.10: Pearson correlation of MUM behaviours self-reported and PADS mean.....	119
Table 6.11: Comparison of mean number of MUM behaviours self-reported by respondents for PADS scores above and below the median.	121
Table 6.12: Comparison between PADS score and self-report of failure to fail or grade inflation (t-tests).....	123
Table 6.13: Free-text responses to the statement “Do you have any comments you would like to make about performance assessments?”	124
Table 7.1: Trainee report of their assessors’ engagement in MUM behaviours in relation to needing to give negative assessment messages to them	143
Table 7.2: Number of items in each subscale respondents agreed with by number and percent.....	147
Table 7.3: Comparison of mean scores by subscale between assessors and trainees	150
Table 7.4: Comparison of assessor and trainee responses to MUM statements (Agree = agree or strongly agree).	151
Table 7.5: Explanations given by trainees who agreed that they do feel pressure at times to make sure that they are marked competent in a WBA.....	164
Table 7.6: Explanations given by trainees who agreed that they may look for an assessor who is more likely to mark them as competent in a WBA	165

Table 7.7: Comparison of self-reports of trainees asking assessors to give or change a rating in an assessment with assessor reports of being approached by a trainee in relation to an assessment	166
Table 7.8: Comments made by trainees in response to the question “Are there any comments that you would like to make about the WBA and/or SITAs as they currently are?”	168
Table 7.9: Comparison of percentage agreement with MUM avoid behaviours and IFOS subscales.....	169
Table 7.10: Comparison of percentage agreement with MUM distort behaviours and IFOS subscales.....	170
Table 7.11: Comparison of percentage agreement with MUM other behaviours and IFOS subscales.....	170
Table 7.12: Pearson correlation – IFOS scales compared with number of MUM statements respondents agreed with.....	171
Table 8.1: Summary of information recorded on the SITA summary form overall and by year of training (ratings and favourableness scores).....	189
Table 8.2: Number of ratings of “1” or “2” across reviewed forms.....	190
Table 8.3: Summary of information recorded on the SITA summary form overall and by year of training (characteristics of comments).....	191
Table 8.4: Mean total rating ranges across a range of favourableness score ranges.	192
Table 8.5: Favourableness scores and mean total ratings for SITAs where 1. At least one rating of “1” or “2” was recorded and 2. No ratings of “1” or “2” were recorded (favourableness score ≤ 3)	192
Table 8.6: Comparison of SITA variables between junior and senior trainees – independent samples t-tests	193
Table 8.7: Pearson correlations between mean numeric rating and characteristics of written comments.	194
Table 8.8: Comparison of mean number of borderline ratings, mean total ratings and mean favourableness scores between main and incorrect groups	197
Table 9.1: Demographic features of interviewees.....	215
Table 9.2: Themes and subthemes developed from the thematic analysis of the interviews....	216
Table 10.1: Summary of <i>how</i> reluctance to deliver negative assessment messages manifests from this study	252
Table 10.2: How the MUM effect literature explains barriers identified in the failure to fail literature	254
Table 10.3: Reasons <i>why</i> reluctance to deliver negative messages manifests in dermatology training in Australia (80,125).....	256
Table H.1 Assessor self-report of number of assessments performed each year	383
Table H.2: Mean PADS scores by demographics across all statements	383
Table H.3: Comparison of mean discomfort levels by assessor gender in relation to telling male and female trainees their performance needs to improve.....	384
Table H.4: PADS scores by demographics in response to statement – “Warning a poorly performing trainee that unless performance improves, you will support a recommendation for his or her removal from the training program”.	384

Table H.5: PADS scores by demographics in response to statement – “Supporting a recommendation that a trainee be removed from the training program”	384
Table H.6: PADS scores by demographics in response to statement – “A trainee’s accusing you of playing favourites in the rating of trainees”	385
Table H.7: Comparison of mean number of MUM items agreed with in subscale by respondent seniority as dermatologist.....	385
Table H.8: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by respondent seniority as dermatologist	385
Table H.9: Comparison of mean number of MUM items agreed with in subscale by respondent seniority as assessor.....	386
Table H.10: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by respondent seniority as assessor	386
Table H.11: Comparison of mean number of MUM items agreed with in subscale by assessor gender.....	386
Table H.12: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by assessor gender	387
Table H.13: Comparison of mean number of MUM items agreed with in subscale by respondent SoT status	387
Table H.14: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by respondent SoT status.....	387
Table H.15: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by seniority as a dermatologist.....	388
Table H.16: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by seniority as an assessor	388
Table H.17: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by assessor gender.....	388
Table H.18: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by SoT status	388
Table H.19: Comparison of mean PADS score with self-report of approach in relation to a clinical performance assessment or recently having to share negative assessment information with a trainee	389
Table H.20: Comparison of number of MUM behaviours self-reported between those assessors who reported by directly approached by a trainee in relation to an assessment outcome with those who did not	389
Table H.21: Comparison of percent of assessor self-reported engagement in at least one MUM behaviour between those assessors who reported by directly approached by a trainee in relation to an assessment outcome with those who did not	389
Table I.1: Trainee responses to MUM avoidance subscale items	390
Table I.2: Trainee responses to MUM sugarcoating subscale items.....	390
Table I.3: Trainee responses to MUM other subscale items	391
Table I.4: Mean number of items respondents agreed to in MUM subscales by trainee seniority.....	391
Table I.5: Percentage of respondents agreeing to at least one item in MUM subscales by trainee seniority	391

Table I.6: Mean number of items respondents agreed to in MUM subscales by trainee gender.....	391
Table I.7: Percentage of respondents agreeing to at least one item in MUM subscales by trainee gender.....	392
Table I.8: Trainee responses to statements in Usefulness of feedback subscale.....	392
Table I.9: Trainee responses to statements in Response/Sensitivity subscale	393
Table I.10: Trainee responses to statements in Confidentiality subscale.....	393
Table I.11: Trainee responses to statements in Retention of feedback subscale.....	394
Table I.12: Comparison of mean number of items in subscale agreed with by level of training	394
Table I.13: Comparison of percentage of respondents agreeing with at least one item in subscale by level of training.....	394
Table I.14: Comparison of mean number of items in subscale agreed with by trainee gender	395
Table I.15: Comparison of percentage of respondents agreeing with at least one item in subscale by trainee gender	395
Table I.16: Percentage of respondents agreeing to statements about WBA by level of training	395
Table I.17: Percentage of respondents agreeing to statements about WBA by trainee gender.	395
Table I.18: Percentage of trainee respondents agreeing to statements by IFOS and MUM responses	396

List of figures

Figure 2.1: Miller’s pyramid (4).....	18
Figure 4.1: PRISMA (<i>preferred reporting items for systematic reviews and meta-analyses</i>) diagram for systematic review question: What are medical specialist trainees’ perspectives of the assessment messages they receive in clinical performance assessments?.....	59
Figure 6.1: Mean PADS scores across all statements by demographics.....	103
Figure 6.2: Number of items in each subscale respondents agreed with by percent.....	107
Figure 6.3: Percent of assessors who agreed with at least one item in each MUM subscale....	108
Figure 6.4: Venn diagram showing the relationship between those respondents who self- reported failing to fail or grading inflating a trainee in a clinical performance assessment.....	112
Figure 6.5: Percentage of respondents to self-report failure to fail or grade inflation of a trainee’s result in a clinical performance assessment	112
Figure 6.6: Percentage of respondents who self-reported being approached by a trainee in relation to a clinical performance assessment result	115
Figure 6.7: Linear relationship between mean PADS score and total number of MUM items agreed with	120
Figure 6.8: Quadratic relationship between mean PADS score and total number of MUM items agreed with	121
Figure 6.9: Comparison of PADS scores for respondents to self-report grade inflation on a SITA.....	122
Figure 6.10: Comparison of PADS scores for respondents to self-report failure to fail on a WBA or SITA	123
Figure 7.1: Percentage of respondent agreement with MUM avoid subscale.....	144
Figure 7.2: Percentage of respondent agreement with MUM distort (sugarcoating) subscale .	145
Figure 7.3: Percentage of respondent agreement with MUM other subscale	146
Figure 7.4: Number of items in each subscale respondents agreed with by percent.....	146
Figure 7.5: Proportion of trainees who agreed with at least one item in each MUM subscale by percent.....	147
Figure 7.6: Scatter plot showing the number of trainees who agreed with items in each MUM subscale by trainee level	148
Figure 7.7: Scatter plot showing the number of trainees who agreed with items in each MUM subscale by gender	149
Figure 7.8: Comparison of mean score for assessors with trainees over the three MUM subscales (Likert scale strongly disagree = 1, to strongly agree = 5)	154
Figure 7.9: Mean differences in scores across the three MUM subscales	155
Figure 7.10: Comparison of level of agreement MUM Distort vs 5-item Distort (with the outlier question removed)	155
Figure 7.11: Confidence intervals for differences between mean agreement with MUM Distort – full scale vs 5-item (with the outlier question removed)	156

Figure 7.12: Percentage of respondent agreement with Usefulness of feedback subscale	157
Figure 7.13: Percentage of respondent agreement with Response/Sensitivity subscale	158
Figure 7.14: Percentage of respondent agreement with Confidentiality subscale	159
Figure 7.15: Percentage of respondent agreement with Retention of feedback subscale	160
Figure 7.16: Scatter plot of number of IFOS subscale items agreed with by level of training .	161
Figure 7.17: Scatter plot of number of IFOS subscale items agreed with by gender.....	162
Figure 7.18: Percentage of respondents agreeing to statements about WBAs.....	163
Figure 7.19: Scatter plot showing correlation between trainee response on Usefulness scale and the number of MUM statements agreed with.....	171
Figure 7.20: Scatter plot showing correlation between trainee response on Response/Sensitivity scale and the number of MUM statements agreed with	172
Figure 8.1: Mean total rating compared with favourableness score.....	194
Figure 8.2: Scatter plot of mean total rating compared with favourableness score	195
Figure 8.3: Scatter plot of mean total rating compared with total word count.....	196
Figure 8.4: Scatter plot of mean total rating compared with total number of narrative units ...	196
Figure 10.1: Diagrammatic representation of the relationship of identified themes. Central or core theme is fundamental with outer level themes varying in importance between assessors and contexts.	255

List of abbreviations

ACD	Australasian College of Dermatologists
AMC	Australian Medical Council
ARC	Academic Research Committee
ARCP	Annual Review of Competence Progression
CbD	Case-based Discussions
CPD	Continuing Professional Development
CS	Clinical Supervisor
DAD	MUM behaviours of Delay, Avoidance, Distortion of a message
Derm-CEX	Dermatology Clinical Evaluation Examination
DOPS	Directly Observed Procedural Skills
DoT	Director of Training
FACD	Fellow of the Australasian College of Dermatologists
HEAG	Human Ethics Advisory Group
IFOS	Instructional Feedback Orientation Scale
ITA	In-training assessment
ITER	In-training evaluation
MCR	Multiple Consultant Report
Mini-CEX	Mini-Clinical Evaluation Examination
Mini-PAT	Mini-Pear Assessment Tool
MSF	Multi-source Feedback
MUM	keeping Mum about Undesirable Messages
OSATS	Objective Structured Assessment of Technical Skills
OSCE	Objective Structured Clinical Examination
PADS	Performance Appraisal Discomfort Scale
PBA	Procedure-based Assessments
PBL	Problem-based Learning
PIF	Performance Improvement Plan

Pro-DA	Procedural Dermatology Assessment
PS	Patient Survey
SITA	Summative in-training assessment
SLE	Supervised Learning Event
SoT	Supervisor of Training
SSP	Supplementary Supervision Program
WBA	Work-based assessment

List of third party copyright material

List of all third-party copyright material included in the thesis

Not applicable

CHAPTER 1

Introduction

Motivation for the study

Recently, on the drive to school, my youngest child asked if I would be doing anything of particular interest that day and I replied that I would be working on my thesis. She thought about this and followed up by asking why I chose the particular topic that I am working on. I explained that in my role as a supervisor of trainees, I had come to realise that there were issues with some of the assessments we use. That the ones which are set in the workplace and aim to be the most valuable type of assessment — given that they can observe the trainee's real-life work, with real patients — have some problems. That sometimes the assessors don't communicate what they really think of the trainee's performance. Sometimes they don't tell the trainee about the things that they did wrong. And sometimes they even pass them on an assessment when they didn't really think they had passed. Her response was emphatic: "But someone could get hurt!".

She went on to explain that sure, while you don't want to hurt a trainee's feelings by giving them a bad result, it would be much worse if it ended up hurting a patient. That conversation, which came in the last few months of my candidature, aptly summarises my motivation for this study.

Purpose and importance of clinical performance assessments

Assessment is a key component of any education program and has multiple purposes. Broadly, it drives learning: whether or not a topic will be assessed and how, is usually one of the first questions that students want to know and they focus their learning accordingly. Assessment can provide students with valuable feedback which can be used to both improve their performance and promote development of self-reflection skills. Robust assessments can ensure that standards are met and permit or prevent student progress, for the benefit and the protection of both the student and the public. Further, assessment results can help the educational institution to review and improve their own performance (1), again for the benefit of students and the public who trust the education and training they provide.

Assessment in medical education has changed profoundly over the past few decades; the now widespread use of clinical performance assessments being one such example.

The term “clinical performance assessment” is used in this work to refer to and describe both the newer work-based assessment (WBA) and longer-standing formats such as in-training assessment (ITA), which are viewed collectively here due to their similarities—specifically that both occur in the real-life work context. The first WBA developed — the mini-CEX (2) — involves a short period of observation of a trainee’s clinical skills. Many additional formats followed this one, with emphasis on various areas including procedural skills, clinical knowledge, and professional skills. While initially intended as formative assessments, the WBAs are used as summative tools in many areas. They may also be used to inform other performance assessments such as ITAs, which typically review a trainee’s clinical performance over a period of time, commonly some weeks or months, often in a summative capacity. While there are many differences between these assessment formats, one feature common to all is their observation of a trainee in their real work environment by a clinician assessor. Today, these clinical performance assessments commonly sit within a program of assessment, which may include the more traditional written and formal clinical exams, with each having their own benefits and challenges.

A prominent benefit of the clinical performance assessments is that they promote the observation of trainees. The realisation that trainees were often insufficiently observed throughout training (3) provided a driving force for the development and use of these formats. Clinical performance assessments allow assessors to see what the trainee “does” (4) in actual real-life clinical practice. These assessments provide opportunities to give early, frequent and meaningful feedback to trainees, with the aim of improving their practice. While other assessment formats can effectively assess a trainee’s knowledge and their application of that knowledge (for example, short answer written exams), they are less able to assess complex skills. The clinical performance assessments however, can assess these clinical skills and knowledge alongside the “broader aspects” (5) of being a doctor. When these formats comprise part of the assessment program, trainees may be provided with more feedback than they would have received otherwise. Indeed many trainees feel that feedback occurs only rarely outside WBAs, which serve to justify them approaching their assessors for feedback (6). The use of the clinical performance assessments permits the involvement of a greater number of assessors, providing trainees with a greater variety of perspectives on their performance (7). When weaved throughout a program, these assessments may have the

potential to reduce the pressure on the final exam, if trainees have had both more experience in being assessed and more performance-improving feedback prior to formal exams. When well-placed, these assessments can allow trainees who need remediation or additional support to be identified earlier and if their remediation is not possible or unsuccessful, to exit from the training program in a more timely manner and before the trainee and others have invested too heavily in an unachievable outcome. However, as assessments are now a more frequent part of both the trainee's and clinician's lives, they also introduce several challenges.

Challenges to clinical performance assessments

Many factors, including logistical and environmental ones, can have a marked impact on the assessment process and may reduce or threaten their value. The need to assess trainees in a clinic or on the wards adds to an often already stretched schedule for both the trainee and the assessor, and ongoing resentment to their introduction is reported (8). The environment in which these assessments occur is messy and often interrupted. The unstructured nature of clinical performance assessments means that “rater” cognition is a relevant consideration. Numerous biases, relating to cognitive, social and environmental factors — for example, the halo effect, data organisation and recall — impact assessors and may threaten the validity of these assessments (9). The clinician assessors conducting these assessments may not have had experience or training in assessment and motivation, and enthusiasm for assessments is of great importance to the outcomes which can be achieved (9). Some assessors may not understand or embrace the changes in assessment which have occurred. Some may be conscious of a conflict with their roles as teacher, supporter and mentor, but also as supervisor and assessor (10). Further, some may doubt their assessment may have any benefit at all. Or even that they may have the potential to cause harm. For the clinician assessor — whose background likely rests in the objective and scientific world, which favours unbiased evidence, who usually completes these assessments alone, without input or discussion with others — the subjective nature of the result may contribute to a lack of enthusiasm to deliver an assessment message. Especially when the message is negative.

Statement of problem

Clinical performance assessments rely on an assessor both forming an opinion on the trainee's level of performance and then delivering that message. This may come in a written or verbal format, given to the trainee either directly or indirectly. However, as briefly outlined, these assessments are influenced by many factors, all of which have potential to jeopardise the validity of the assessment result delivered. A substantial amount of work to date has been aimed at helping these assessments to achieve the stated goals, including development of assessor training and assessment form modifications to ensure that the message formed is as accurate as possible. But the delivery of this message is an equally important component to consider, as it is known that at times, an assessor does not deliver their assessment message (10–12). While there may be many reasons for this, in some cases it is due to an assessor's reluctance to deliver it.

When the performance observed is of a high standard, delivery of the assessment result may be easy; but when the level of performance is poor, the delivery of that message may be challenging. Outcomes including failure to give feedback, grade inflation and failure to fail may result and the consequences may be profound (12–14). The missed opportunity to give a trainee appropriate feedback on and ratings of their performance can have many effects, including the loss of opportunity for self-reflection, improvement or remediation. The trainee may be set up for failure in future assessments. Further, the training institution is denied feedback on its performance, so limiting its own development. But most importantly, when trainees are undeservedly passed or allowed to progress when they have not reached the required standard, patient safety may be compromised. Clinical performance assessments are currently an integral part of many medical specialist training programs. However, at present, the evidence suggests that they do not always achieve the aim of providing honest performance information to trainees. Among many contributing factors, the assessor's motivation to give a trainee their opinion on the performance is highly relevant to this problem.

Statement of the aim and scope of the study

In this work, the claim that reluctance to deliver negative assessment messages occurs in the clinical performance assessment setting and so jeopardises the outcome of these assessments will be interrogated. I will seek to demonstrate how this claim manifests and explain why it occurs in a medical specialist education setting in Australia. This will provide a framework for examining the phenomenon in other settings, and a platform to consider further solutions to the issue surrounding the difficulties with delivering negative assessment messages to trainees.

The work is restricted to clinical performance assessments in common usage at the time of the study; thus the delivery of assessment messages in formats such as Objective Structured Clinical Examinations (OSCEs) and written exams are not the focus, though comparisons are made to these formats where appropriate.

The setting chosen for the study is the dermatology training program in Australia and this has been selected for several reasons. As the program is Australia-wide, positioning the study in the Australasian College of Dermatologists (ACD) allows for a national viewpoint of issues to be obtained. The relatively small size of the training program makes gaining a breadth of opinion feasible in a project of this size and further, as a dermatologist, my insider status may facilitate in uncovering and understanding the relevant issues more than may be possible with another training program. In addition, my previous research (15) and my own observations, as both a trainee and as an assessor, have shown that the issue of reluctance to deliver negative assessment messages manifests here. The choice of setting does certainly not imply however that the issue of reluctance to deliver negative messages is restricted to or more severe in this setting. Similarly, it is not to imply that it is restricted to specialist or even medical training in general, though aspects of the issue may be specific to such areas.

As mentioned, there are many factors that threaten the potential outcomes of clinical performance assessments, but this work will focus on those related to the assessor's delivery of their assessment message. The work will examine the issue with focus on recent past practices to ensure that data collected is current. The overall research question for the work is: *How and why does reluctance to deliver negative assessment messages manifest in dermatology training in Australia?*

Significance of the study

Trainees deserve to receive accurate and honest information on their clinical performance. And their patients are depending on it for their own safety and outcomes. Accuracy in assessment requires attention to aspects such as rater cognition and the environment, while honesty in assessment requires consideration of factors including assessor motivation. Anecdotally, in some other training settings outside medicine, the delivery of an honest assessment message seems to be more expected and indeed demanded by those involved. For example, in elite level sports and music training learners often seem to more actively seek out corrective information on their performance to guide their improvement (16). It is also curious to observe that some clinician assessors struggle with the delivery of bad assessment news to trainees, when the delivery of bad clinical news to patients is something that all doctors are familiar with. And when the consequences of not delivering the appropriate news can be severe. The content, training and relationships involved may be of relevance to this difference between the delivery of bad news to trainees and the delivery of bad news to patients.

While it is beyond the scope of this study to compare performance assessment in medicine with those in other areas, the reasons why reluctance to deliver negative messages in the clinical performance assessment context occurs and how this may manifest will be explored. This will involve reviewing literature outside of medical education, about a phenomenon known as the MUM effect — or keeping mum about undesirable messages. This literature will inform an enhanced understanding of the problems observed, beyond those which focus on ratings and grades only, and show other ways in which reluctance to deliver a negative message can manifest. As well as developing a broader understanding of the issue, this will then permit consideration and development of other methods to address the difficulties with the delivery of negative assessment messages.

This work therefore aims to develop new understandings around the issue of reluctance to deliver negative assessment messages, and ultimately to propose methods to minimise its impact.

Overview of the study

The problem of assessor reluctance to deliver negative assessment messages in medical education will be explored from several perspectives. Clinical performance assessments have several parallels with employee performance appraisals and the business literature will be used to guide the design of part of the study. The study is situated in a pragmatic paradigm and uses mixed methods and a corresponding variety of approaches to explore the research question. This includes a systematic review of the literature on trainees' perceptions of assessment messages, a survey of dermatology trainees and assessors, a review of recorded grades compared with written comments from a sample of real-world assessments and an interview study of assessors' opinions and practices.

In the first part of the thesis, three literature reviews on various aspects relevant to assessment in medical education will be presented. Chapter 2, Background literature, will discuss the purposes of and the developments in the clinical performance assessments in the health professional education setting, and the benefits and challenges of these formats will be highlighted. A review of the history of and reasons for one issue identified with these assessments — failure to fail and grade inflation — will conclude this chapter. Chapter 3, The MUM effect, will explore a body of literature identified in the rater cognition literature as having potential implications for the issue of failure to fail and grade inflation. This body of literature on the MUM effect, will be presented and its application to the medical education setting will be discussed. The final part of the literature review chapters — Chapter 4, Trainees' perspectives of assessment messages — will present a review of the literature from the trainee's view in this assessment context. Given one of the broad aims these assessments is to provide trainees with information on their performance, a review of the trainee perspective of the assessment messages they receive is important to determine and this may well have implications for assessors on how to deliver their assessment results. These literature reviews will inform the design of studies developed to answer the research question.

Chapter 5, Research Methods will outline the theoretical perspective guiding the work and discuss the use of assessments in the study context of the ACD. The studies designed will be briefly outlined, with further details presented in each study chapter. The mixed methods approach employed will collect information from the perspectives of both assessors and trainees. Chapters 6 to 9 will present the result of the studies

undertaken. This will begin with an exploration of assessors' self-report of discomfort with assessments and their behaviours through a questionnaire study in Chapter 6. Chapter 7 will present the trainee perspectives on their assessors' behaviours and their experiences in the context of these assessments also obtained through a questionnaire study. Chapter 8 will present a review of real-world assessment data with a focus on comparing the recorded numerical ratings with written comments sections to look for differences between the valence of the message delivered by each. These chapters focus in particular on how reluctance to deliver negative messages manifests. The final data chapter, Chapter 9 will focus largely on why reluctance manifests, by presenting the results of interviews with assessors which explore their current and recent assessment practices and experiences.

In Chapter 10 'Discussion', the results from all studies will be synthesised and considered together with the literature to present answers to the research question, along with implications and recommendations for consideration and future research.

During the writing of this thesis and being mindful of several well-known aphorisms, I recognised that various depictions of my favourite animal, the elephant, could supplement the explanations and understandings provided in certain areas. As such, illustrations of elephants appear at the relevant parts of the work.

Summary

A robust assessment program should ensure that those who graduate are able to perform as effective and safe clinicians. The design and implementation of the clinical performance assessments are well-placed to provide trainees with timely information on their performance and so help their future development as clinicians. But current evidence suggests these aims are not always achieved. This study seeks to determine how and why reluctance to deliver negative assessment messages manifests in dermatology training in Australia.

CHAPTER 2

Background literature:
work-based assessment in
medical education

Introduction

This chapter discusses the assessment of medical trainees' skills and knowledge in the workplace. The purposes of assessment are broadly described, and the history of the development of the clinical performance assessments — which in this context refer to ITA and the WBA — is presented. The benefits and challenges of these assessments, as reported in the literature, are discussed with a particular focus on the issue at the core of this thesis — the delivery of the assessment message to the trainee. The chapter concludes with an introduction to the next chapter, which reviews the literature on a phenomenon relating to reluctance to deliver negative messages — the MUM effect.

Purposes of assessment

Assessment is usually a key component of any educational program and the medical education context is no different. The concept that “assessment drives learning” is well recognised (17,18) and observed through trainees' almost universal desire to know how their performance in a rotation or subject will be assessed. Further, what is assessed and how it is assessed also serves to communicate what is important to learners (19).

Assessments can achieve multiple potential functions, the focus of which can depend on the context in which they are used.

Perhaps most familiarly, assessments can function as a hurdle — to evaluate competence and ensure that required standards are met. This summative or grade focus (20) permits a certification (21) and gate-keeping role and is often employed prior to a trainee progressing (22) or graduating (1). Inherent in this is an expectation of a prediction of the future performance of those assessed (17), which in the medical education context, is fundamental for patient safety and public reassurance.

In contrast to this barrier function, assessments may have a formative or learning (20) focus as their outcome and emphasise the provision of feedback to the learner, to guide and promote their development. In practice however, such a separation of assessments into formative and summative types is often not absolute and overlap may exist (for example (23)). On an individual level, both formative and summative types of assessment can serve to motivate and give direction to learners (24). As “the means by which a trainee's learning is measured” (25) both are potential sources of information

for the trainees who undertake them, as assessments can provide feedback about their strengths and weaknesses, though the depth of feedback provided is usually greater with formative assessments (20). The identification and subsequent remediation of poor performers is important (26,27) and one goal of assessments (28).

Ideally, accurate assessment information would be used by trainees to reflect on their performance and help to promote the development of skills in self-remediation and the identification of professional development needs (1,29). Ideally, trainees would be supported with remediation programs where necessary in order to achieve this and so lead to improvement in their performance. These outcome-generating functions of assessments are important as evidence shows that both physicians and medical students have limited self-assessment abilities (30–32) meaning that external information on performance is required. Some authors report learners have a tendency to “overestimate abilities” (33,34), however, this is not always the case. As an example, one study of a group of Canadian radiology trainees reported that “generally self-assessment scores were significantly lower than multisource feedback scores” (35) and a study with internal medicine residents found that a majority underestimated their performance on an in-training exam (36). Kruger and Dunning have shown that those in the bottom quartile of performance tend to overestimate their abilities with the opposite effect seen initially for those in the top quartile (37), which highlights the importance of providing accurate external information on performance. However, it is also recognised that trainees do not always accept the feedback they are given, especially if it differs from their self-assessment. For example, some trainees think that they are not as highly performing as their feedback suggests, and suspect assessors are reluctant to deliver negative assessment information: “I don’t know if they’re not saying things. I don’t know if they’re avoiding saying things they want to say because they don’t want to hurt the resident’s feelings or what not” (38). This is a further area for consideration and will be explored in the next chapter.

As outlined in the Introduction, assessments also have important broader functions, beyond those which relate directly to the individual trainee who completes them. In the medical education context, assessment has a role in the protection of the public (1) so it is crucial that they function effectively, both in terms of accurate certification of competence and through the provision of developmental information to trainees to result

in improvements in their performance in order to achieve this. Additional functions of assessments concern the provision of information back to the educational institution itself. This is necessary for their own reflection on the effectiveness of their teaching and curriculum to permit necessary improvements (19), to inform and guide changes in the curriculum and the training of the teachers (1), as well as to satisfy accreditation needs (19) and for educational research (1). Further, assessments are valuable to help determine the impact of an educational intervention, beyond that of trainee satisfaction, by their ability to measure an increase in trainee knowledge or skills (39).

For assessments to achieve these many goals, it is necessary that the information they deliver is accurate. The features of robust assessments are well-described (17) and the importance of validity and reliability (40), and also educational impact, feasibility and acceptability (18), are recognised. In regard to impact, it is appreciated that the feedback obtained from an assessment “can only be as good as the assessment that informs it: high quality assessment is conditional for high quality feedback” (41). Features such as the timing of delivery of developmental feedback in order to result in improvement in trainee performance are also reported; with assertions it “should be immediate, specific and corrective and occur regularly” (42) and with opportunities for improvement following (43). Programmatic models of assessment permit such a longer-term view of assessment by their separation of individual assessments and ultimate outcomes or decisions (44). Thus, it is important that the assessments used in medical education contexts are designed and implemented with reference to these features. The development of assessment formats outlined in the next section shows how this has been addressed.

Assessment formats

There are many assessment formats used currently in medical education settings and these have evolved significantly over recent decades (18,40). Prior to this, assessments largely comprised the written formats and standardised clinical assessments such as Objective Structured Clinical Examinations (OSCEs) and long case examinations, though the use of ITA is longstanding. While many of these formats still have an important and ongoing role in the assessment of many trainees today, an awareness that “[t]raditional assessments have emphasised knowledge acquisition, rather than clinical ability” (45), together with the recognition that trainees were infrequently observed

during their training (1,3), were among the driving forces for the development of assessments based in the workplace (WBA). Holmboe and colleagues recognise that the “[d]irect observation of clinical skills is a critical first step in helping trainees to improve their clinical skills” (46); and limitations to standardised assessments, in particular that a trainee’s “performance in an OSCE might not be the same as performance in real life” (17) are recognised. Further, Weller identifies that while fellowship exams may be successful in assessing a trainee’s level of knowledge, they may not be able to measure a key goal — the trainee’s ability to practise successfully (45). Though stated in the context of anaesthesia training, it may be anticipated that this would hold true for other areas of medical specialty training. It makes sense for assessment to occur where training occurs, and Norcini and colleagues assert that assessment in such a real-life setting with real patients “should more easily generalize to future performance” (40). Thus in the past few decades, “the assessment of trainees in the workplace has become recognized as an essential component of evaluation”, given the “performance in the workplace is the core of medical competence” (47).

As mentioned, ITA, which report on trainee performance in the workplace over a period of time, have been used in many medical education settings for a long period. More recently, the WBA which generally look at shorter periods of trainee performance, each with a somewhat discrete focus, have been developed and widely introduced into many training settings [for example as described in Weller, Castanelli, Chen and Jolly, 2017(48)]. The first of these was the mini-CEX, which was developed by modifying and shortening the longer traditional Clinical Evaluation Exercise (CEX) and concentrates on the assessment of trainees’ clinical skills — their history taking and examination ability over a brief period of time (2). Many others have followed, including formats which focus on procedural skills. For example: Directly Observed Procedural Skills (DOPS); Procedure-Based Assessments (PBA); Objective Structured Assessment of Technical Skills (OSATS). Others have focused on clinical knowledge (Case-based Discussions (CbD) for example) or professional skills more broadly, including communication (Multi-source Feedback (MSF) and variations). Some use assessors other than the trainee’s direct supervisor, including other consultants, the trainee’s peers, nursing and administrative staff, and patients. For example, the mini-PAT [in Brown et al., 2014 (49)], and the Patient Survey (PS) [in Johnson, Booth, Crossley and Wade, 2011(50)]. In some cases, the trainee chooses their assessor(s) and

in others they are allocated by a supervisor. The number of WBAs continues to grow and table 2.1 shows a recent list of the range of WBAs developed to date from a Google scholar search.

Table 2.1: List of common workplace-based assessments in use (Google scholar 6 Aug 2019)

Name	Abbreviation
Mini-clinical evaluation exercise	Mini-CEX, mCEX and variants (e.g. Derm-CEX)
Direct Observation of Procedure	DOPS and variants (e.g. Pro-DA)
Case-based Discussion	CbD
Mini-Peer assessment tool	Mini-PAT, mPAT
Patient survey	PS
Team observation	TO
Objective Structured Assessment of Technical Skills	OSATS
Clinical Encounter Cards	CEC
Clinical Work Sampling	CWS
Blinded Patient Encounters	BPE
Multi-source Feedback	MSF, 360-degree feedback
Team Assessment of Behaviours	TAB
Patient Satisfaction Questionnaire	PSQ

ITAs share the focus of observation of a trainee in their workplace across a broad range of skills but have several key differences to the WBA. While the WBA usually focuses on trainee performance over a limited period, the ITA typically observes performance over a longer one, usually in the range of some weeks to months. Similarly, most of the WBAs are usually completed by a single assessor, while an ITA is typically completed by more than one, with the collation of assessments presented to the trainee. Both ITA and WBA formats usually comprise two components — a numerical rating one, where trainee performance is rated across various domains and assessors check a box or record a number to reflect the level of performance; and a written comments section, where

assessors can write free text comments, often guided by a prompting statement such as “Strengths” or “Suggestions for development”.

A further consistent feature across both of these formats — which as described, are collectively referred to as “clinical performance assessments” in this work — is that the observation and assessment of the trainee occurs within the scope of their everyday work environment, as opposed to a structured or controlled environment. As the content of each assessment is not planned or standardised as occurs in an OSCE, each is unique and so will vary between training sites or even between cases at one site. In addition, those conducting the assessments are usually the clinicians with whom the trainee works, who will likely vary in their individual assessment practices (40) and may have potential role conflicts with the task, though the importance of the relationship between the assessor and trainee is appreciated (51). As these formats have the advantage of being authentic to daily practice, they have the ability to assess higher order skills and constructs such as trainee attitudes (52), professionalism, counselling skills, organisation and overall competence (53), which are not effectively assessed in written or unobserved formats (40) or where performance may be rehearsed. As such the trainee can be assessed in areas beyond the memory recall or application of knowledge that are measured with formats such as written papers and OSCEs. They can measure what the trainee “does” as compared with just what they “know” (4) (figure 2.1). While these common features make the clinical performance assessments valuable and offer the potential to provide trainees with rich and holistic information on their performance, their design also introduces several problems and issues.

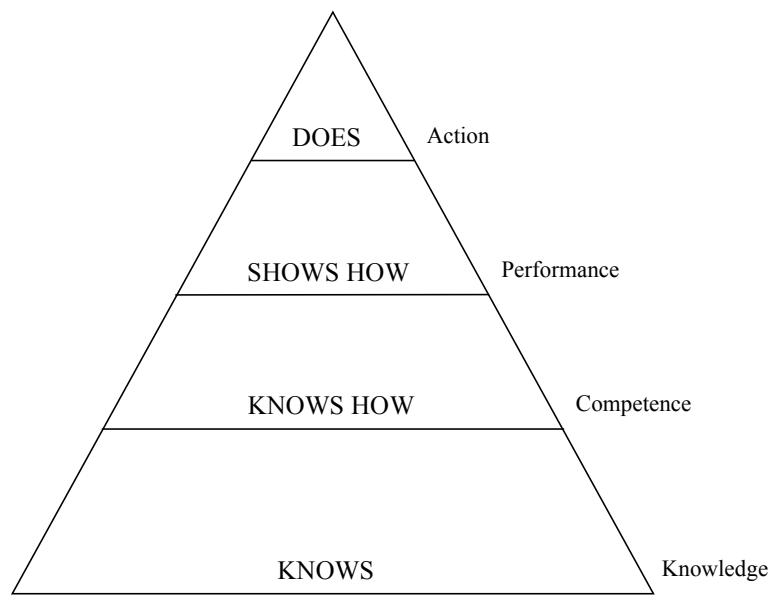


Figure 2.1: Miller's pyramid (4)

Challenges with clinical performance assessments

The clinical performance assessments occur in a messy environment. While they offer the potential benefit of assessing trainees on patients with real clinical signs, assessments may be interrupted as the clinical needs of the patient will take priority over an assessment. For example, an assessment may be interrupted if a patient becomes unstable or needs to be taken for an investigation or procedure or if the trainee being assessed is required to attend to another patient. Some trainees find WBAs are challenging to arrange or complete — to find an appropriate patient, and then an assessor and the time to complete the assessment can be difficult (8), as can completing the assessment forms themselves (54). Similarly, finding the time and a place for assessors to complete ITA information and then discuss this with a trainee can be problematic.

The outlined unpredictability of material in WBAs prevents both their standardisation as well as the training of examiners in the particular case (45). As assessors are known to vary widely in their “expertise and experience” (47) and many have had little training in workplace-based assessments (45) or in education and assessment broadly, this may affect the outcome of the assessment for the trainee. The training of assessors, while regarded as very important by many [for example (52)], is not alone sufficient to create an expert assessor and the types of training and the expected or possible outcomes that

may be achieved are seen to vary widely. For example, while Malling and colleagues report on the beneficial effects of even short training courses for assessors, with positive effects on assessors' knowledge of ITAs persisting for several months (55), others have different results. Gingerich and colleagues' synthesis of research in rater errors (56) report that some authors doubt the effects of training. For example, in their study, Newble, Hoare and Sheldrake found rater training had no effect on ratings, finding that consistent raters don't require training while inconsistent raters weren't improved by their training (57).

Others support the "training, monitoring and feedback" of assessment frameworks used by assessors (47). In their study on the nature of feedback given after a mini-CEX, Holmboe and colleagues note that most of their participants had undertaken training in feedback however, they also emphasised the need for ongoing or revision training in feedback (46). Jason (58) concurs, asserting that "being and remaining a professional educator demands far more than an occasional, brief workshop" (58), highlighting that it is a long-term endeavour. The concept of continuing professional development in their area of clinical practice is certainly familiar to doctors but may not be automatically considered in relation to their assessment role. Thus the particular assessor has implications for the assessment outcome.

The allocation of assessor, including whether the trainee chooses them, are further areas of potential issues with these formats and may impact on fairness (40). Norcini and McKinley also highlight the problem of conflict of interest with assessors assessing their own trainees in the WBA (40), which may also have application for the ITA (51).

In addition to the assessor, trainees and the assessment environment may also influence the outcome of an assessment and the assessment information received. Strategic selection of assessors by trainees can have an impact. This was demonstrated by Brown and colleagues who compared the scores obtained on an MSF when a trainee selected an assessor with those when their clinical assessor selected the assessor. They found that the trainee's "hand-picked assessors (mainly peers)" gave higher ratings in comparison to those from their clinical supervisor, with most assessors focusing on strengths rather than areas for development (49). Assessors in the study expressed sympathy for trainees' behaviour, noting "it's only human to select those who'll give favourable feedback" (49). While some trainees may have actively sought positive

assessments, there was also a realisation that doing so eventually “would be counter-productive” (49) and result in biased outcomes. Environmental factors, such as time and workload, are further factors that may influence results in the clinical performance assessments. Paget and colleagues noted that lower ratings were seen where there were longer time delays with ITA completion and from those assessors who rated more students (59). In the OSCE setting, priming assessors with highly as compared with poorly performing medical trainee performances was seen to have an effect on the assessor’s subsequent assessment behaviours, including harsher ratings of borderline performances (60). Whether the same applies to the clinical performance assessment setting is yet to be explored.

Much research and work has been undertaken in relation to the workings of assessment tools themselves, with modifications aimed to improve their use and usability and so the outcomes achieved. Alterations to the design and rating scale (45,61) have occurred as well as the ongoing development of new tools to assess different constructs. The cognitive load placed on assessors in the context of the clinical performance assessments is an important consideration, and rater cognition and influences upon raters is a well-researched area (9). Authors report that an assessment framework that “serves to reduce the cognitive load of assessors is likely to improve the accuracy of ratings” (47) and so improve outcomes achieved.

Thus the accuracy of the assessment result is a crucial element given that one of the purposes of these assessments is to provide trainees with useful information on their performance. The following section will further discuss issues which affect the outcomes of these assessments with a focus on the reporting of underperformance.

Issues with assessment outcomes – problems in delivering the outcome

To fulfil their aims, assessments need to have the ability to identify trainees who are performing poorly or below the standard required or are otherwise in need of assistance or remediation. As discussed, many assessment formats include both a numerical rating and a written comment section but it is usually the ratings component which determines the overall outcome of the assessment: the “numerical score allows rapid review and identification of a poor trainee” (25). But that, of course, is only if the rating is both

valid and reflects the views of the assessor. However, as outlined the formation of an accurate result is not always straightforward and some of the issues presented above contribute to make this a challenge. So at times, the assessor does not arrive at the “correct” result and issues relating to the assessment tool in use, assessor cognition and training can be relevant to this. However, arriving at the correct assessment result is only one part of the assessor’s duty. The result must then be delivered in the required manner — including to the trainee and the institution. At times though, the assessor may not deliver their intended message — in full or in part — and this may relate to their level of motivation to do so. In some cases, they may fail to deliver the narrative comments they wish to, referred to as failure to give feedback. In others, an assessor may knowingly elevate the grade or rating they assign, termed grade inflation, or pass a trainee whom they have determined does not reach the required standard, termed failure to fail.

The under-reporting of poor performance is widely documented in education in general. Historical reports describe the pressures faced by assessors, including implied or real threats to their physical safety (62). Reports of the practice of inflating grades and failing to fail in education in general were first most commonly noted in the mid 1960s, in the Vietnam era in the USA. Authors noted there was “little debate as to why grades were rising: Professors clearly were grading more generously” to avoid their students from being sent to serve in Vietnam (63). While the issue of grade inflation appeared to settle for a time after this period, grades were seen to begin to increase again in the 1980s leading White to argue that “[a] simple linear extrapolation of this trend predicts all “A’s” sometime in the second half of this century” (64). The reasons given for this observation were postulated to be a lowering of standards, the practice of students “gaming the system”, and the introduction of student evaluations of their teachers (63). Other authors describe that “cultural changes in the value of self-admiration [that] resulted in specific policies and behaviors (e.g. self-esteem-focused education, grade inflation)” (65) were relevant. While others recognised that improvements in teaching methods may explain the increases in grades (66), overall, most accept that the phenomena of failure to fail and grade inflation exist in education.

Several authors have in fact documented the phenomena of failure to fail and grade inflation within the medical education setting. This has been in largely in the medical

student setting (10,21,67–70) with some accounts in higher and medical specialty training settings (11,24). The clinical performance assessment formats have been involved in such reports, which are seen through both direct admissions of such practices from involved assessors and through review of assessment results. The latter includes observations which suggest the existence of failure to fail and grade inflation such as low failure rates or seemingly almost universally high ratings. For example, in their review of ITAs in an undergraduate program in the Netherlands, Daelmans and colleagues note that “borderline or fail judgements were rare” (69); and Tunbridge and colleagues in an analysis of the in-training assessments of all medical trainees in the UK over three successive years, reported that only 1.0% of medical registrars were marked as unsatisfactory and in need of a period of retraining (71). While Norcini and colleagues suggest that owing to their previous education and the selection criteria for entry into medicine, “not many trainees will be considered unsatisfactory” in the WBA (40); Bingham and colleagues also note very low rates of reports of borderline or unsatisfactory performance in their review of end of rotation evaluations for junior doctors in NSW and present other information to suggest that the “real rates of underperformance are much higher” (22). Further, Patel and colleagues (72), quote studies which suggest that the number of trainees struggling to meet the standards of their program may in fact be more like 10% of a cohort and find that in their own study, over 6% of trainees were seen to be in difficulty. Evidence for the presence of grade inflation is seen through reports of seemingly overinflated results. For example, in a review of over 1000 ITAs, Paget and colleagues reported that almost two out of every three clerks achieved ratings of “above expected level” or “outstanding” (59); and Kendal and MacRae noted they awarded “‘outstanding’ grades at twice the defined frequency” (73).

“In clinical medical education, the importance of feedback extends beyond pedagogy. The goal of clinical training is expertise in the care of patients. Without feedback, mistakes go uncorrected, good performance is not reinforced, and clinical competence is achieved empirically or not at all” (3)

There are many possible reasons for the under-reporting of poor performance or inaccuracies in assessment leading to failure to give feedback, grade inflation or failure to fail which are further discussed in the next section. However, it is important to firstly

emphasise the potential consequences of failing to identify poor performance in medical trainees. Simply put, the goals of the assessment will not be achieved and the result of this may be severe in this setting. When trainees who have not reached the required standard are passed the outcomes may be profound and serious, for both them and importantly, their patients (14). Failing to fail sends the message to the trainee that they are competent; they are given the impression that they have reached the required standard, though some may know that they have not. Failing to give feedback may leave the trainee unaware of problem areas of their performance and can mean that they generate their own feedback — for example, lack of criticism may be interpreted as meaning that there are no issues with their performance (74). Failing to fail means that poorly performing trainees potentially miss out on receiving remediation or extra support and so the opportunity to intervene and improve their performance (22,75). Further, the trainee's development of self-reflective skills and their ability to assess and so act on their continuing professional development needs (22) may be impaired. Failure to fail can lead to damage to the reputation of the profession (76); while cases with extreme consequences may be uncommon (76), even small and seemingly insignificant instances have the potential to lead to harm. Most importantly, the failure to identify and fail underperforming trainees where appropriate, leaves them vulnerable to mistakes and complications in their future practice, meaning that patients and the community who put their trust in those who care for them, may be harmed (68,77).

The reasons why the under-reporting of poorly performing trainees occurs are important to explore and rectify, but its occurrence is perhaps puzzling given evidence shows that remediation of poor performance is often successful. For example, Aram found that where documentation of underperformance occurred at the mid-term assessment, 76% of those interns identified and remediated went on to complete the term satisfactorily (26) and similarly Tunbridge and colleagues found that 64% of those who underwent repeat training had a “satisfactory outcome at their next assessment up to a year later” (71). Paice also agrees that most of those identified as needing targeted or remedial training “can be helped with extra experience and supervision” (78). A systematic review on remediation programs for medical students and medical specialist trainees, which sought to determine how and why such programs may achieve beneficial outcomes, highlighted the complexities involved in the area of remediation. Cleland and colleagues found that many of the studies included in the review lacked sufficient

detail to be able to determine the keys to a successful intervention. Many described “more of the same” or “examination coaching” interventions as opposed to “learning to learn” ones (79) which may provide longer-term benefits. Others however, advise that the costs of remediation, which include the extra time and effort invested, and perhaps diverted from patient care and other trainees, be considered carefully, particularly when the performance problem relates to professionalism issues (76). However, the failure to appropriately identify struggling trainees is seen to be a “disservice” to them (21) and denies them the opportunity to improve or where this is not possible, to exit a training program before a serious outcome occurs or excessive amounts of time, effort and money have been invested. So, when the potential negative outcomes of failing to report underperformance are so significant and the opportunity to positively influence a trainee’s performance exists, why does failure to fail underperforming trainees occur?

Reasons for the under-reporting of poorly performing medical trainees

Many authors have studied and written about the reasons for failure to fail and grade inflation in the medical education setting. As with assessment more broadly, these reasons may also relate to the assessor, the student/trainee or the environment, also alone or in combination. In 2016, a systematic review summarised the literature on assessors’ reported perspectives of the barriers and facilitators to failing trainees in health professions education, including medicine, dentistry and nursing at all levels (80). The analysis revealed the barriers to failing a failing trainee related to the themes of (i) assessor’s professional considerations, (ii) assessor’s personal considerations, (iii) trainee related considerations, (iv) unsatisfactory assessor development and assessment tools, (v) institutional culture and (vi) consideration of available remediation for the trainee. These are discussed in this section in combination with other literature where appropriate.

On the surface, some of the reasons identified for failure to fail in this review — such as assessor development and tool considerations, and practical issues including options for trainees after failing — appear relatively straightforward and intuitive and could be expected to be easily addressed. The importance of training for assessors in assessment has already been mentioned and is supported by many authors (27,47,53,67,81–84). Singh and Modi assert that assessor training is an important part of developing a

successful WBA program (85) and others have shown that their assessor workshops improved knowledge and enhanced the feedback given to trainees (52). Van der Vleuten and Verhoeven concur, emphasising the need for assessors to learn about the assessments and their importance, and be given time to practise and develop assessment skills, while receiving feedback on their own performance. They highlight that if assessments are not performed seriously, the whole process risks being seen as unimportant and so become “ritualized, with people jumping through hoops and ticking boxes without meaningful learning taking place” (86). As such, failure to fail may result. Training can serve to address the reported concerns around the inherent subjectivity in performance assessments as compared with structured assessments of knowledge (such as on written papers). Such factors may dissuade some assessors from failing a trainee, due to doubts about their own standards and the reproducibility of assessments and the interpretive nature of the task, that “there is no single ‘true’ score” (87), which may cause discomfort to assessors who more commonly deal in an objective realm. These difficulties with the scale used in the tool, identified by both assessors and trainees (88), may be addressed with a move to entrustability scales (47,89), which describe a trainee’s performance in comparison to the level at independent practice, a construct which may be easier to understand and express for many. So the barrier posed by unsatisfactory assessor development and/or evaluation tools including a lack of skills or confidence to assess (10,28,67,90) or knowledge of what or how to document concerns (11), should be able to be addressed by specific assessor training, including emphasis on the standards expected (10,91).

Lack of remediation options that are presented as a further barrier to failing to underperforming trainees (11,92) should also be able to be successfully addressed. Solid effective remediation options and outcomes which do not result in the “stigmatisation” of trainees (10) could help to avoid failure to fail and promote a change in thinking from that of “failure” as a negative concept, to one of providing support and opportunity for improvement for the trainee. This may further serve to reduce the feelings of failure reported by some assessors that a trainee’s poor result also reflects their own poor performance as teachers and supervisors (10,13,90). Trainee-related considerations that are closely related to this area could include developing options for trainees who do fail — such as the ability to exit a program with recognition or qualification of some sort, rather than with nothing. From an institutional perspective,

practical aspects that are reported to promote failure to fail, such as a lack of privacy (10) or time for quality assessments, should also not be insurmountable issues to resolve.

However, other areas highlighted in the Yepes-Rios and colleagues systematic review (80) as barriers to failing an underperforming trainee likely pose a greater challenge to resolve than those highlighted above. Inexperience in the assessor role is thought to be relevant to difficulties failing trainees (68,93) and others have also noted that “[m]ore experienced teachers reported few difficulties in providing negative feedback” (94). However, work outside of educational settings, in the similar employee performance appraisal setting, suggests that junior assessors may have some advantages over more senior ones. For example, Whitla and Tirrell showed that “the level of raters closest to the ratees were best able to rate them” and that this group was also the only one “able to discriminate between sections of the rating instrument” (95). Looking more broadly at the appointment of assessors, Noel emphasises the need for “careful selection and orientation” of those who teach and work with the trainees (96); and Bing-You and colleagues recognise that not all clinicians will make good assessors (97). While Norcini suggests we should “identify a core group of faculty whose only educational job is assessment and formative feedback” (53); many medical education settings, including in Australia, rely on a volunteer assessor pool, meaning there may be therefore little choice of assessor available.

Assessor professional concerns around legal repercussions (28,92), or the anticipation of an appeal (11) or challenge to their assessment by the trainee (10,13,21,67,68,90) or of being over-ruled by a higher authority (28) are more challenging barriers to failing trainees; but could also be targeted through a combination of robust assessor training and support by the institution and colleagues (10,11), though are difficult to eliminate completely. The contribution of role conflict to the issue, both from teacher/evaluator and evaluator/service provider aspects (10,28) could be minimised by assigning independent assessors to the assessor role. This may also serve to reduce the well-reported impact of personal relationships on assessment results (28,68,91,92,98), address the lack of time available to assess (10,67,90), and issues of the extra time and work involved in failing a trainee (10,28) that are noted as professional considerations for the occurrence of failure to fail. However, while Cavalcanti and Detsky argue that a

trainee's coach is best able to observe their development over time, which an external or unknown assessor would not have knowledge of, they still caution of the role conflict between that of supervisor and assessor (24). Further, the significant financial commitment this would involve may not be possible in many settings. Whether assessors are anonymous or known to the trainees whom they assess can impact assessments, and Klein and Babineau have noted the potential for differences to exist between anonymous and direct evaluations and so potential implications for the assessor (98).

Aside from these reasons highlighted in the systematic review (80), further reasons for the failure to fail trainees include the potential consequences or outcomes of an assessment, for both the trainee and the assessor. Discussing the experience of midwifery students with an observational diary in The Netherlands, Govaerts and colleagues report that the purpose of the evaluation can be a “decisive factor in the way evaluation of performance is documented” and lead to failure to fail on occasions (43). As described earlier, failure to fail or grade inflation have at times, been used consciously and deliberately to avoid a negative consequence for a trainee, such as conscription in the Vietnam era or more recently to prevent the loss of a college sporting scholarship (63). In the medical education setting, some have inflated reports to help a trainee secure a favourable future training position (67,99). The desire not to “penalize” a student since everyone else grades high” further motivates some assessors to engage in such behaviours (68). Some settings have made local modifications to separate learning and evaluation, such as by removing the satisfactory/unsatisfactory rating component from WBAs (100) or by developing further WBAs, for example Supervised Learning Events (SLEs) whose focus is formative (72); in part to address this.

In ITA formats, some assessors note that lack of feeding forward to future assessors in combination with short assessment periods and limited exposure to trainees may make assessment challenging (16), and perhaps encourage them to err on the side of caution and not fail a trainee about whom they have concerns. Ensuring that assessors are aware of their responsibilities and obligations in assessment, and involving the previous assessors, may reduce this as a reason for failure to fail, though the contra-argument of

the “potentially negative consequences for residents” of feeding forward the results from their previous ITA (16) is an important aspect to consider.

Authors have recognised further aspects relevant to the assessor in the issue of failure to fail. Van der Vleuten and Verhoeven recognise that “the value of instruments depends more on the users than on the instruments themselves” (86) and assert that for an assessment program to be successful, attention to its implementation is of vital importance. The attitudes of assessors to the assessments are particularly important to consider, in order to foster commitment to and engagement with them. Persisting resentment around the manner in which new formats were introduced is recognised (8) and would need to be addressed in order to achieve the best outcomes. In their study of WBAs, Fokkema and colleagues report trainees and assessors view them “as both a stimulus and a hindrance to teaching and learning” (101), and the fundamental importance of assessor engagement in achieving a useful assessment outcome is reported (102).

Therefore while many reasons are identified for the issues of failure to fail and grade inflation, not all appear insurmountable, although some require significant financial and higher-level institutional interventions. While it is recognised that many modifications have already been made to date, including improvements in the design of assessment forms or the training of assessors (80), these are yet to eliminate the problem of failure to fail and grade inflation. Thus it appears likely that there are other explanations for some cases — those perhaps related to the fallibility of assessors that “cannot easily be overcome through either training or more detailed assessment frameworks” (103) are important. Sim comments on the “natural” desire of assessors to see a junior colleague succeed and states that it “is well recognised that doctors find it very difficult to report the serious shortcomings of a colleague” (104) and others seem to agree. In some of the first studies on the issue of failure to fail in the medical education setting, Dudek and colleagues reported that “faculty do not lack the ability to use the forms as they currently exist, rather they lack the willingness” (11); and similarly Cleland and colleagues pointed to the “‘human’ gap in the assessment process” (10). Paice explains that assessors may be reluctant to “face a difficult conversation with the trainee, or [have] a sense of guilt at not having done a better job of training” (78). Norcini recognises that widespread satisfactory ratings in WBAs more likely reflect assessor

reluctance to award failing grades than universal satisfactory trainee performance (40). Thus, more fundamental issues appear relevant to the issues of failure to fail and grade inflation, with assessor reluctance to deliver negative information to a trainee seeming to be a key part. The next section focuses on those reasons for failure to fail and grade inflation which appear to stem from an underlying reluctance of an assessor to deliver a negative message.

Reluctance to deliver a negative assessment message

“There is a general reluctance for faculty to give lower grades for a variety of reasons” (21).

Failure to fail and grade inflation sometimes occur due to more complex underlying reasons. These are separate to those, perhaps more somewhat straightforward barriers discussed in the previous section, and offer explanations for the persistence of these phenomena, despite the efforts made to date to resolve them.

It may be expected that difficulties with the delivery of negative assessment information may be restricted to summative assessments, but authors have noted the under-reporting of performance issues in general formative evaluations also including Problem-based Learning (PBL) cases (105). As Ende notes, there is some level of evaluation included in all feedback information, this may explain why some assessors find it difficult to give feedback to trainees, even when not summative (3). As such, the consideration is given to an innate reluctance of the assessor to deliver negative messages, which broadens the view of the problem, beyond that of one which focuses on grades and ratings (and thus failure to fail and grade inflation) in a summative setting to include the failure to give comments on performance more broadly.

This assessor reluctance to deliver negative messages has been reported by several authors (10,11,21,28,68,70) and seen to relate to many factors. Complex human interactions including interpersonal relationships, fear of confrontation and outcomes, and generational factors have relevance to this. Cohen and colleagues study on the delay in recognising struggling medical students that they observed was occurring in their department, concluded that “it is not the insensitivity of the instrument in detecting problems that is the issue, but rather the reluctance of faculty to give negative evaluations in the face of such problems” (28). The prior interpersonal relationship between assessor and trainee can lead to a tendency to be “more likely to cut them slack” (106) or upgrade their result if they really liked them (12), though this tendency does not affect all assessors (12).

The impact of personal feelings on assessment results is well-recognised. Outside of medicine, in the similar employee performance appraisal setting, authors have

questioned whether it is even reasonable to “expect raters to ignore or set aside their feelings when participating in performance appraisal” (107), and this arguably applies to the clinical performance assessment setting. Others have noted that while a hierarchical relationship exists between a trainee and the assessor, it may not always operate in the direction expected. For example, where assessors are close to trainees in terms of academic level, junior assessors may feel “their personal relationships with [trainees] impeded negative judgements, as did their feelings that sometimes they were barely more proficient than the [trainees]” (69).

Other authors have discussed that the “Millennial generation”, “raised with an emphasis on being special” may not accept negative feedback (34) and so react with emotion which assessors wish to avoid. Assessors may be therefore concerned that raising issues with a trainee who lacks insight may be met with “denial, defensiveness and counterclaim” (78), leading the assessor to be reluctant to deliver their message and rather “hope the problem will resolve itself, especially if the training rotation is likely to move the trainee to another unit before too long” (78). Borges and colleagues, in a study comparing Millennial with Generation X medical students at their university, noted differences do exist in the personality traits of those in different generations, and questioned whether Millennial students, who have often grown up in very structured environments, may therefore need to be treated differently in many areas of education, including those of evaluation and feedback (108).

Difficulties delivering assessment information, especially when negative, are not only confined to real-life situations. The task is difficult for some even in experimental settings. For example, a difference between an assessor’s observation and the judgement they delivered, has been observed in experimental situations, even where there were no negative consequences for either assessor or trainee. In a study, the participants reported that their “public judgement (feedback in writing or verbal feedback to the trainee) may differ from their private judgement (judgement resulting from assessment of performance according to private performance theories and standards)” (41). A concerning deficiency in behavioural specificity in the feedback and that narrative comments appeared “to be less useful for learning and professional development than verbal feedback” (41). This comparison of the assessor’s written comments with their verbal ones, though in a simulated environment, is of great interest

and echoes Colletti's work, where a study showed that assessors awarded higher grades to medical students when they were required to give the assessment results directly to them compared with giving results in writing only (70).

Reluctance to deliver negative information has been considered by some to be a part of the "culture" of medicine. Cavalcanti and Detsky have observed the somewhat pathologic situation in medicine where "strengths are highlighted and praised, but weaknesses may be hidden" (24); and others suggest that reluctance to deliver bad assessment information, may be consistent with a medical culture of reluctance to reveal "errors and shortcomings" (109). Some authors have contrasted this to other settings such as sport where the coach is expected to identify relative weaknesses in order that they may be addressed and so lead to improvement (24), and others have also noted differences to other training cultures (sport and music) with respect to the delivery of performance information (16).

When there is not an immediate consequence for not delivering negative performance information or results, many may find it preferable to engage in failure to fail and grade inflation. Long's study participants acknowledged that assessment outcomes depend on the levels of honesty and strength within the assessor, with one noting that "it is far too easy to just pass the trainee and let someone else deal with the problem" (25). While others note that delaying the delivery of such news may make the situation worse, so making it harder to subsequently "challenge behaviour that has been tacitly accepted for some time", and potentially leading to "accusations of discrimination or bullying" (78); when the balance is tipped in favour of not telling, failure to fail and grade inflation may easily result.

The setting for the clinical performance assessments promotes some of the difficulties seen with them. To contrast, with formal written and oral exams, much time is spent writing and preparing the questions and marking guides, and this usually occurs in collaboration with other assessors. Candidates are usually anonymous to essay markers and examiners may calibrate results and do not have to give feedback immediately, or in fact at all, directly to the candidate. OSCEs are scored with reference to a rigid checklist, and long cases and other formats often involve two examiners who can discuss performance afterwards. However, in clinical performance assessments, clinician supervisors, in the setting of a normal day, are expected to assess, rate and give

comments and grades to trainees. The cases are non-standardised, the trainee usually works with the assessor, and the result is often required to be delivered immediately and/or without the opportunity to think about the performance or discuss it with a colleague. On paper this seems an onerous task for assessors.

The elephant in the room

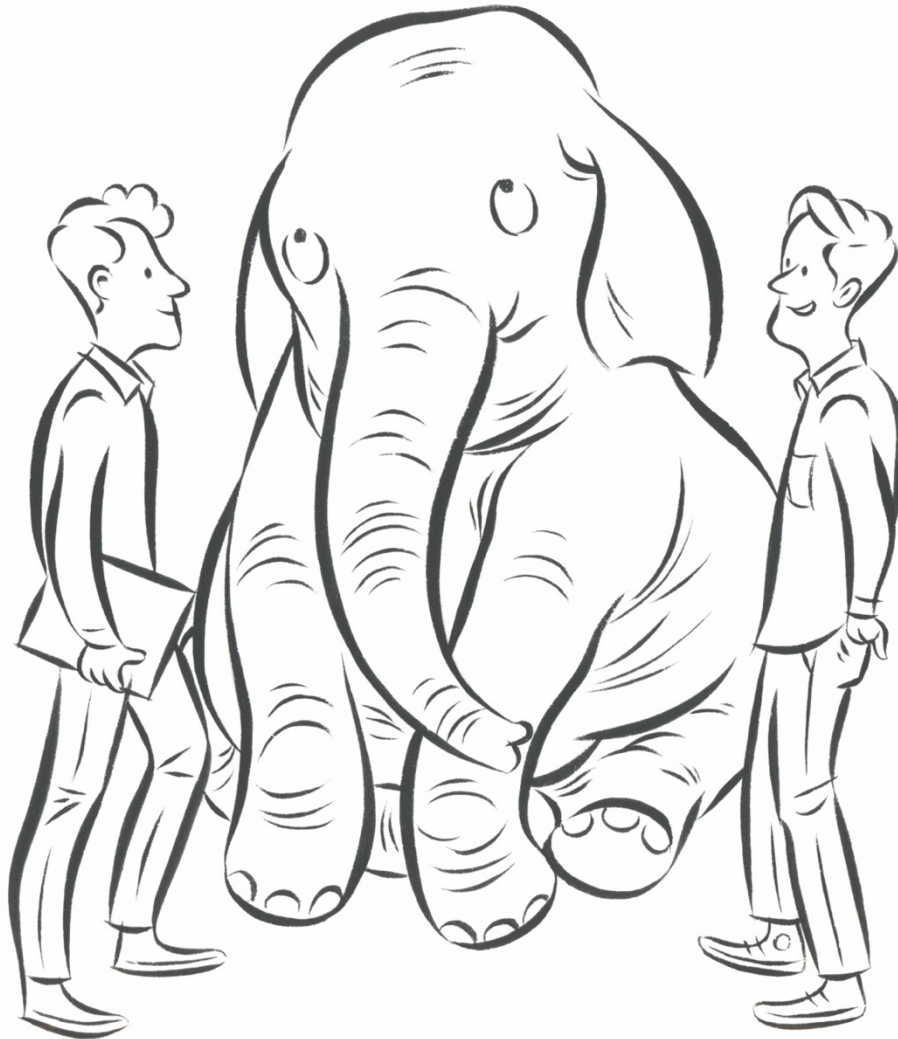


Illustration 2.1: The Elephant in the room (illustration by Jo Martin)

The clinical performance assessments aim to give trainees feedback on their performance in order to help their development. Many training programs currently place great emphasis on the results they provide and at times they fulfil progress requirements or contribute to progress decisions. However, there is an elephant in the room. As described by the Cambridge dictionary, this term describes a situation where

“there is an obvious problem or difficult situation that people do not want to talk about” (110), at least publicly. That is, for various reasons at present, the clinical performance assessments do not universally achieve the aims they set out to as at times, assessors do not wish to deliver their assessment message. While the generation of an accurate and valid assessment message is important, this does not matter if it is never communicated to anyone. Many works from the literature have highlighted that there are issues with the assessment results delivered and there is much anecdotal evidence to support this. My own experiences as a supervisor of trainees reveals many examples of where assessors displayed reluctance to deliver assessment information in full, and comments from many other supervisors in many other settings echo these ones. Therefore, in this work and as the previous literature has demonstrated, I assert that critical problems exist with the clinical performance assessments as they currently operate, namely that they currently do not universally deliver the assessment message completely, especially when it is negative. This has the potential for significant negative outcomes for all stakeholders — including the trainee, the educational institution and most importantly, the patient. This work will seek to explore how and why reluctance to deliver information on trainee underperformance manifests. This will commence with two further literature reviews which frame the empirical studies — the first to explore the literature in regard to reluctance to deliver negative messages and the second to explore the trainees’ views of the assessment information that they receive in the context of the clinical performance assessments.

Summary

This chapter has summarised the purposes of the clinical performance assessments, described formats in current use and highlighted some of the benefits and issues with their use. The issue of failure to give feedback, grade inflation and failure to fail have been discussed, with a focus on where these occur due to assessor unwillingness to deliver negative assessment information to a trainee. The next chapter discusses the phenomenon of the MUM effect and considers how it may be relevant to medical education and in particular, the issues around the delivery of negative assessment messages.

CHAPTER 3

The MUM effect and its relevance to assessment in medical education

Introduction

This chapter presents a review of the literature on the phenomenon of reluctance to deliver negative information broadly, and considers the relevance this may have for the issues of failure to give feedback, grade inflation and failure to fail in medical education.

The chapter discusses the MUM effect — the tendency to keep mum about undesirable messages — which was introduced to the medical education literature by Williams, Klamen and McGaghie in 2003 (9). A summary of the original MUM experiments which show the manifestations of and reasons for the effect is provided. Newer studies are also included which supplement and extend the original work. While the MUM effect has been discussed in many settings, including business and information technology, it has not been widely explored in medical education to date. The chapter therefore includes a discussion of the possible implications of the MUM effect in medical education, including that it may offer an explanation for the observation that attempts to solve the issues of failure to fail and grade inflation have not met with complete success (80). The chapter concludes with the introduction to the final literature review, which presents the trainees' perspectives on the assessment messages they receive in the context of the clinical performance assessments.

This chapter is the original work which underlies a narrative literature review published in 2019 (111). This paper was written independently of this thesis and the commentary within it reflected the collected views of all authors. For this reason and together with the requirements of the University of Melbourne's thesis with publication policy, where narrative reviews cannot be included in theses, the published article has therefore not been included.

Background

Chapter 2 has presented literature which shows that many clinical supervisors feel reluctant to deliver negative performance results (10,11), and that this reluctance can affect their behaviour. At the most extreme, this may result in failure to fail and the passing of a trainee who has not reached the required standard. At other times, they may give a trainee a higher rating than the one they feel is deserved. As discussed, failure to fail is a serious issue in medical education with equally potentially serious

outcomes. To summarise, when trainees are passed undeservedly, the aims of an assessment are not met and in turn, this can ultimately affect the wellbeing of the patients who trust in their doctor, in addition to the effects this may have for the trainee and the training organisation. But as presented, the problem is larger than the focus on grades that the use of the terms “failure to fail” and “grade inflation” may imply. Indeed, anecdotally, in the real world, examples where assessors fail to give negative comments, written or verbal, on a trainee’s level of performance are arguably more common than failing to give negative grades. While the systematic review on failure to fail (80) thoroughly summarised the literature on the issue across all health professions education, it also acknowledged the persistence of the problem despite the many efforts made to address it to date. The review highlighted how process or logistical issues — in addition to an assessor’s willingness, as opposed to their ability to report poor performance can contribute to this issue — echoes work from outside of medical education in the performance appraisal literature (112). As such, a broader search of the literature including aspects of the rater cognition, human resources and organisational psychology literature was undertaken in order to gain a different perspective on the issues.

In 2003, in the rater cognition literature, Williams, Klamen and McGaghie introduced the concept of “the MUM effect” in their article on cognitive, social and environmental sources of bias in clinical performance ratings (9). As defined in Chapter 1, Introduction, this effect describes the tendency of people to keep mum about undesirable messages (113), and was originally reported in the psychology and sociology literature in the 1970s. In simple terms, “the ‘MUM effect’ describes this reluctance to give negative feedback” (114). Since the initial work, this phenomenon has been extensively researched, but not within medical education. The research findings into the seemingly obvious claim — that people don’t like to tell others bad news — presents a valuable way to reconsider the issue of failure to fail and the difficulties assessors may face in delivering negative feedback to medical trainees in clinical performance assessments. Further, it offers an explanation as to why efforts which have not directly addressed this issue of reluctance, have not completely eliminated failure to fail.

This chapter summarises the literature on the MUM effect with the aim of considering how it may influence the delivery of negative assessment messages more broadly in medical education — beyond that of focusing on grades to consider comments on performance also — that is, the whole assessment message. While, as mentioned, none of the studies are set in a medical education setting, there are still potential lessons to learn from them. As presented, Cleland and colleagues note, the human element of the assessment process seems to have not developed in assessment as much as tool development and standard setting approaches (10). It is proposed that examination of the extensive work on the MUM effect may lead to new ways in which to consider and manage the issue of the delivery of negative assessment messages in clinical performance assessments.

Rosen and Tesser, the original investigators into the MUM effect, studied the claim that people are more reluctant to communicate bad news as compared with good (113). The first MUM study, published in 1970, tested the proposition that “the reluctance to transmit information is directly dependent on the inferred desirability of the message for the potential recipient” (113). In the study, subjects “overheard” an urgent and important message and their views on whether they would pass it on to the intended recipient or not were explored (113). It was shown that when the news was bad, subjects took both a longer time to deliver their message and did not deliver it as completely compared with when the news was good. This depended on their view of the “pleasantness” of the news for the recipient and the urgency of the message (113). As a result of seeing the differences in news transmission depending on its valence, Rosen and Tesser subsequently modified the commonly used phrase to become “no news is bad news” (115).

Since these original works, many studies have explored the MUM effect as a phenomenon. These include both in experimental [for example (116,117)] and in real-life environments. Settings for the latter include information technology (118) and other industries (119), psychology (120) and education (121,122). While some argue that the MUM effect is pervasive (123–125), or at least widespread (122), it is reported that “individuals may vary in the extent they engage in these behaviours” (126) and may keep MUM in different ways. The next part of this chapter outlines the manifestations of and reasons for the MUM effect and then proposes to show how these may apply to

the medical education context, specifically to highlight how knowledge of the MUM effect may shed new light on the issue of failure to fail. Methods of working with the MUM effect and where further research is required in order to understand the phenomenon in the medical education context are also highlighted.

Manifestations of the MUM effect

The literature reports three main ways the MUM effect may manifest:

- Delaying giving the message
- Avoiding giving the message
- Distorting or “sugarcoating” the message

The acronym DAD has been coined to both summarise these behaviours and as it fortuitously complements the MUM terminology. The DAD behaviours are discussed in this section.

Delay in the delivery of bad news was noted as a manifestation of the MUM effect from the early experiments (113) and studies (127–129). In more recent experiments, Benedict and Levine supported these works, showing a delay in the performance ratings given to lower performers by their supervisors (130); and Dibble and Levine similarly reported a delay in the sharing of negative information in comparison with good (116,124). However, delay in the delivery of bad news is not a consistent finding (131) and further, may not always represent a harmful MUM behaviour. For example, a delay in giving results can help signpost that they are negative, so a limited amount may in fact enhance communication in some instances.

Keeping MUM can also mean avoiding delivering a negative message or a part thereof (113). In addition to the original studies by Rosen and Tesser (113,129), the avoidance of delivering negative messages has been discussed more recently in real-life settings (121,132,133). In an interview study with school principals as an example, more than half of the participants admitted avoiding dealing with the issue of underperforming teachers. One principal even begged students to behave in order that he could avoid having to deal with their teacher’s failings (121). Marler and colleagues also showed that avoiding the delivery of negative messages occurs in real-life work settings, and that personality and the prevailing norms can modify the expression of this (132).

Distorting the unpleasant message is the third common way of keeping MUM. While again initially seen in laboratory studies (131,134), later works support these findings. Distortion is described in various ways. For example, being indirect was a technique used by subordinates to manage the negative message of denial of an unethical request in one experimental study by Ploeger and colleagues (135). Some business settings use the term “sugarcoating” to describe distortion of the message (132,136). As with delay, not all distortion is seen to be harmful and to a point, some distortion of a message may enhance communication. However, when in excess, distortion, like delay can become harmful to communication (136).

As discussed, while these manifestations of the MUM effect are reported from settings outside of medical education, the pervasiveness of MUM makes it reasonable to predict they would also occur within it. The next section of this chapter describes the reasons which underlie the expression of the MUM effect. Scenarios from the health professions education literature are given to highlight how they may be viewed as manifestations of the MUM effect.

Why does the MUM effect occur?

From their original work, Tesser and Rosen proposed that the MUM effect occurred due to three main reasons: the “communicator’s self-concern”, the “communicator’s concern with the recipient” and the “communicator’s concern with norms” (125).

Communicator’s self-concern

This explanation predicts that MUM behaviours result from a teller’s concern for themselves, specifically in relation to guilt they may feel, concerns about being thought of negatively by others or effects on their own mood should they tell the bad news.

Early studies (113,137) suggested that tellers may feel guilty when delivering bad news and so may engage in MUM behaviours in an attempt to escape this. Few subsequent studies have considered the impact of guilt and instead primarily focused on other factors; for example, the effect of certainty of outcome, and either closeness of relationship (138) or similarity of outcome for participants (139). Guilt may well be a relevant motivating factor for MUM behaviours in medical education as many assessors report being mindful of the lack of time they have devoted to supervising and teaching

trainees (10,16,140). This may be compounded by their knowledge of the time, effort and money (10,92) trainees have invested in their training, and so lead to feelings of guilt and to MUMing.

Assessors may also be concerned that they will be thought of poorly if they deliver negative assessment messages and the phrase “don’t kill the messenger” (132) supports this. Authors have demonstrated that tellers do think about how their delivery of news will be regarded by others (141) and use strategies to minimise negative evaluation of the teller, including the deliberate use of delay (142,143). The health professions literature similarly reports that assessors are concerned about how trainees view them. Yepes notes that assessors’ professional considerations, including their concerns that “failing a student would harm their professional standing” (80) pose as a barrier to failing a trainee. However such concerns are not universal among assessors (12), which supports the assertion that reasons for and the extent of MUM behaviours may vary between individuals.

Avoiding the required negative change in the teller’s mood when delivering bad news is given as a further reason for MUM behaviours (125). The breaking bad news literature both in clinical settings (144–149) and outside of the doctor-patient situation (121,143) documents the negative effect that breaking bad news may have on tellers. This too may apply to the assessment setting and promote an assessor keeping mum.

Communicators concern for others

MUM behaviours can arise from the teller’s concern for the recipient and these are explained through the inferred recipient’s desire to hear the news and the emotionality hypothesis (125).

Most people may assume that others will want to be given good news and conversely that they will not want to receive bad news, and this may therefore impact on the teller’s behaviour (125). In an experiment, Conlee and colleagues tested how knowledge of a recipient’s desire to receive news, even if it was bad, affected the teller’s behaviour. They found that news was told more if the teller was aware that the recipient wanted to hear the news, but the MUM effect was not eliminated entirely (150). When a desire to hear news is not communicated, tellers presumably come to their own conclusions, which may differ from the potential recipients’. Assessors and trainees often hold

different views about other aspects of assessment (151), so this may be relevant for their desire to hear also.

Tesser and Rosen assert that “Communicators prefer not to put the recipient in a negative affective state” (125) and that the “emotional impact a particular message would have on the recipient” may affect them (125). A teller’s decision of whether or not to tell a negative message can therefore depend on their understanding of how a recipient may respond to it. So many factors may influence tellers in their decision whether to tell or not: the teller may simply want to avoid making another unhappy; they may fear being around someone who has been upset by bad news, especially if they don’t know how to help or console them; or they may be concerned about being judged harshly by an emotional recipient (125). An eagerness to avoid upsetting trainees is documented by some clinical assessors in the health professions education setting (12,80) and the teller’s personal experiences with negative feedback may also colour their future behaviour in relation to giving feedback (3,126).

The communicator’s concern with norms

The third explanation for the MUM effect as described by Tesser and Rosen relates to the prevailing societal rules or norms. The first was that norms proscribe giving bad news, and the second and favoured explanation, was that norms for giving bad news are ambiguous in comparison to giving good news (125). Some authors have demonstrated how the delivery of negative messages is impacted by the organisation’s climate (152,153). Others however, note that additional factors, such as the teller’s personality are of greater relevance to whether MUM behaviours manifest or not (132).

What can influence the way MUM manifests?

In addition to the underlying reasons for the MUM effect already discussed, several factors are reported to modify the behaviours seen. It is important to understand these influences upon the delivery of a negative message in order to help predict and recognise when the MUM effect may be likely to occur, so that ultimately methods to minimise harmful effects may be developed.

The following section of this chapter summarises these factors, again giving examples from the health professions education literature which may illustrate how the MUM effect may manifest in the setting of the delivery of negative assessment messages.

The clear assignment of responsibility to communicate a negative message can impact a teller's behaviour. In an experiment, Tesser, Rosen and Batchelor (154) found that the subjects who were made responsible for the communication of news also felt more obligation to do so, in comparison to others who were either asked not to deliver or were allowed to come to their own decision, and this is supported with more recent work (153). In the performance assessment context, clarifying an assessor's responsibility to pass on negative assessment messages when appropriate may increase their willingness to do so. This is important as some supervisors do not feel comfortable in an assessment role (155) and some feel conflict between their teaching and assessing roles (24). Others simply "do not view failing students as part of their role" (140), so assigning responsibilities to give news may potentially result in less MUM behaviours.

Similarly, a teller's own satisfaction in their work environment may impact on their willingness to deliver negative messages. For example, perceptions of job security (133) and job satisfaction (119) can influence their likelihood of keeping MUM. These findings are likely to have direct application to the assessment setting, where those who are engaged and committed to their role as assessor may be expected to show lower levels of MUM behaviours compared with those who are reluctant or resentful of the role.

"Uncertainty about the conclusion" (153) or outcome of a message may also affect the likelihood of a teller passing it on. Keil and colleagues argued this in the situation of workers reporting to management when information systems projects had become "runaway", and parallels also may be drawn to the assessment setting. If an assessor lacks confidence in their own judgement of a trainee's performance or the potential consequences that their assessment may lead to, they may be more likely to keep mum and not pass on their concerns. Additionally, the likelihood of consequences arising from a negative message may affect the likelihood of its transmission. Weenig has shown that where consequences are indefinite, bad news is delivered more readily than where the consequences are definite as the former is considered to be "more urgent and more helpful" (138). Furthermore, in a performance appraisal setting, Smith and

colleagues have shown that assessors' levels of discomfort with an appraisal increase in line with the importance of the appraisal (156); and Cox and colleagues demonstrated an association between discomfort levels and MUM behaviours (157). These findings have important implications for the performance assessment setting and suggest that the MUM effect may vary with the possible outcomes of the assessment — that is for summative clinical performance assessments as compared with formative ones.

Other factors which have been reported to modify the expression of the MUM effect relate directly to the individuals involved, and include the individual's preferences around the delivery of messages and more personal attributes such as age and gender. The relationship between those involved is also reported to impact on the telling of news.

In the performance assessment setting, some assessors express a desire to be anonymous, and some report being concerned and influenced by a lack of anonymity (158). While the literature shows that the MUM effect may still occur in situations where the teller is anonymous (159), some works show that the transmission of bad news may be enhanced when the teller is not visible to the recipient (142). Rosen and Tesser had previously and unexpectedly found greater bad news transmission in a non-anonymous compared with an anonymous situation, however, this may be explained as their study participants did anticipate having contact with each other in the future (127). O'Neal, Levine and Frank similarly noted the importance of expected future contact and predicted that withholding the transmission of bad news would not occur if the consequence of doing so would be to show the potential teller as "cowardly or irresponsible" (159). This would seem to have significant application to the clinical performance assessment situation where trainees and assessors often have an ongoing relationship.

The medium used for the delivery of news is important in communication and individual variation in preferences is seen. Sussman and Sproull found computer-mediated methods resulted in less distortion of negative feedback compared with telephone or face-to-face methods (160). Further, a meta-analysis by Ye and colleagues concluded that less direct methods yielded less positive responses when compared with those seen with telephone interviews (161). In contrast, Dibble showed that individuals vary in their preferences, finding that compared with face-to-face delivery, email did not alter self-reported levels of reluctance to communicate (123). However, this was a hypothetical situation involving the delivery of news between friends so may not

extrapolate to the assessment situation. Studies in education settings support the claim that the mode of delivery can affect the results delivered. Both in experimental (162) and as presented in Chapter 2, in real-life settings (70); the delivery of feedback face-to-face compared with indirectly, was seen to be associated with upward distortion of results.

Other personal attributes of tellers have been studied in relation to their delivery of news. While Tesser and colleagues found that the study subjects' mood did not significantly affect their transmission of news (128), subjects in the bad mood condition were seen to show more spontaneous communication compared with those in the good mood one. Marler and colleagues showed that those who are more aware of social cues are more likely to display the MUM behaviours of avoidance and sugarcoating compared with low self-monitors and extraverts (132). And while contrary to their prediction, Waung and Highhouse found that those with lower empathy levels gave higher performance ratings as compared with those with high empathy, when they expected to be giving their feedback directly to the person. They explained this as showing that those with low empathy were motivated by a greater concern for themselves than the recipient (162).

Other features may affect a teller's likelihood of sharing negative messages, though in some areas, further research is required to determine the impact. The teller's gender may be of relevance, but the literature is not consistent in the direction of the effect (124,130,132,154). In regards to the recipient of the news, Tesser and colleagues showed that their subjects felt a greater responsibility and obligation to deliver news to a female recipient as compared with a male (154). Marler and colleagues showed that older workers were less likely to engage in avoidance MUM behaviours (132) and this is consistent with some work from the medical education literature, where more junior assessors were seen to give significantly higher ratings on daily performance evaluations to trainees than more senior ones (93). However, Noel and colleagues found that variables including years of practice and experience as an assessor did not affect accuracy of assessments (163). Outside of medicine, others (for example Smith et al., 2000), show that additional or a combination of factors may be relevant.

The nature and presence of the relationship between the teller and the recipient may affect the likelihood of news transmission, and this is highly relevant to the medical

education situation. Some early work suggested that the non-delivery of bad news related to the relationship between the teller and the recipient (125); and Weenig and colleagues assert that the experienced moral responsibility between friends may explain the increased in the delivery of bad news (138). However, in a more recent experimental study, Dibble could not demonstrate modification in the delay in delivery of news due to relationships (116) though the “seriousness” of the news may explain the difference in results. (Dibble’s study involved delivery of a “social skills” test result, while Weenig’s involved a simulation of giving news of work promotion or dismissal.)

The precise nature of the relationship is also relevant to its impact on giving news. Sussman and Sproull suggest that if a teller likes the recipient, he or she may fear damaging the relationship by delivering negative messages (160). Smith, Harrington and Houghton reported that the level of performance appraisal discomfort reported by evaluators did not necessarily reduce with the increasing length of a relationship (156), and the relationship between discomfort and MUM behaviours has been mentioned (157). Hierarchy can further impact behaviours, and Ploeger, Kelly and Bisel have shown that subordinates were more likely to keep mum to a senior (135). The additional complexity of delivering bad news in an educational setting is noted as the feedback from a teacher to a student needs to address both these relationship and educational aspects (164,165).

Finally, while some note the MUM effect could be considered a “universal” tendency, (166), anecdotally, cultures vary in their norms around the giving of news. This literature review has included works from many settings including Asia (118,122,167), Israel (121), and Europe (159) including the Netherlands (138,139) and anecdotal descriptions from the UK (168). In some cultures, MUMing has been reported as contrary to expectations. Yariv notes the MUM effect was seen in Israel, despite the “direct style of discourse within Israelis society” (121). In contrast, Ramingwong and Sitthichok, who note that Asian culture teaches against opposing seniors, demonstrated that cultural influences on the delivery of bad news in an IT setting were not very strong and “could even be easily dampened by a strong organisational culture” (118). The culture of medicine is an important consideration in the delivery of negative assessment messages. In comparison to other settings such as elite sport and music, medicine is likely to promote MUM behaviours as criticism — especially when documented

formally — is largely “not done” (38); and authors have noted the difficulties giving constructive criticism to high achieving individuals, “we’re not great at getting it; we ... somehow take it personally” (16).

What does the MUM effect mean for medical education?

Failure to fail and grade inflation are well-documented in medical education settings as well as in education more broadly. The MUM effect literature offers a new approach to understanding such issues in more depth and to hopefully then lead on to new methods to minimise their impact.

The concept of the MUM effect helps to broaden the focus of the issue with reluctance to deliver negative messages beyond that which emphasise grades and ratings, to a wider view that considers reluctance to deliver all types of negative information, including numerical ratings, and written and verbal comments. The MUM effect can offer an explanation for the observation that attempts to address failure to fail and grade inflation to date have not met with universal success. Dudek and colleagues argued that “It would appear that faculty do not lack the ability to use the forms ... rather they lack the willingness” (11); and the MUM effect explains this by recognising the fundamental reluctance of people to deliver negative messages. It asserts that the **delivery** of the message, in addition to its formulation, is important in assessment and explains why modifications such as form development, scale modifications, and assessor training; and attention to cognitive aspects have not and will not eliminate all cases of failure to fail and grade inflation. The terms “failure to fail” and “grade inflation” themselves may focus on the importance of getting the numbers right, given this is the part of most assessment forms which gives the outcome of an assessment. While this is of course necessary, an understanding of the MUM effect in turn increases our understanding of the complexity of the issue of why assessors often feel reluctant to deliver negative assessment messages of all types, and emphasises the importance of both the generation of the assessment message and its delivery.

The MUM effect explains the seemingly obvious claim that bad news can be hard to deliver, and predicts why and how assessors may behave in the situation of needing to deliver negative assessment messages. Williams and colleagues assert the effect seems to have “clear implications for supervisors’ communications of poor performance to

practitioners” (9) and many of the examples highlighted in the previous sections of this chapter show how issues in the delivery of negative assessment information can be viewed as manifestations of this pervasive effect.

The MUM effect can predict the ways that assessors may behave when faced with having to deliver negative assessment messages — being through avoidance or delay in delivery or distortion of news. Promoting an understanding and awareness of these behaviours would help to identify when the MUM effect is occurring. As well as facilitating assessor self-identification of engagement in MUMing, increasing awareness and the meanings behind these behaviours would allow others to identify this also. For example, when the delivery of an assessment result is delayed, trainees and the training organisation may take this as a flag for possible MUM behaviours occurring and take action to investigate further.

Research of the MUM effect in the medical education setting is required in order to determine which assessors are particularly likely to engage in MUM behaviours and under which circumstances. For example, assessor demographics such as age or experience may prove to be of relevance, and this may guide the appointment and training of assessors. Similarly, identifying assessors’ preferences in relation to mode of delivery of feedback and whether these are able to be modified with training or by other factors is important. Cox, Marler, Simmering and Totten have developed scales to measure feedback medium preferences (157) which may be of value in identifying individual assessor preferences in the delivery of assessment messages to trainees. This in turn may guide their selection and training for involvement in particular assessment formats, and Cox proposes that specific training of assessors based on their individual personality factors and tendencies to MUM may be of benefit in decreasing the detrimental MUM effects. Research into variations in MUM behaviours within the components of assessment formats is also an important area for consideration. Some studies have noted a lack of alignment between the numerical rating and narrative written comment parts of an assessment (28,73) may occur, and while one explanation for this may be that the two components are distinct and hence not be expected to align, a further is that assessors keep mum in different ways. For example, some assessors may MUM in the ratings area of an assessment, but deliver their message more completely in the written area or through verbal feedback (12). Development of ways to

identify this numerical MUMing may be an important consideration for future research, and ensure that the entire assessment message delivered to the trainee is consistent. Consideration of how the trainee as the recipient of the news may affect the delivery of the assessment message is also an important element to consider. Trainees vary in their opinion and acceptance of feedback (169) and their reactions to negative feedback may in turn affect assessor behaviour (94), as predicted by the MUM effect.

An ultimate aim for research into the MUM effect would be to minimise potential harm arising from such behaviours. As discussed, it is recognised that a small amount of behaviours such as delay (123) or sugarcoating (136) may in fact promote effective communication. For example, those who deliver bad news frequently, may be specifically taught to signpost upcoming bad news with preparations prior to delivering it (170). Similarly, Govaerts and colleagues' recognition that "accurate low ratings will tend to turn into self-fulfilling prophecies by demotivating students who will subsequently perform at substandard levels" (171) may support the potential benefit of some amount of sugarcoating. However, excessive MUM behaviours may in fact promote failure to fail, should they lead too far down a path whereby telling the truth becomes ultimately too difficult. Promoting awareness of MUM behaviours would ideally lead to the development of robust measures to support the assessor so that not MUMing (or not excessive MUMing) becomes the preferred behaviour. Knowledge of the MUM effect may therefore be used to help guide development of solutions to failure to fail. Additionally, motivation to support non-MUM behaviours (139) or reward those who raise issues (122) may prove valuable. Educating trainees to detect MUM behaviours and to take a proactive approach to defend their learning may be additional areas to explore. An example of this could be in the situation of where numerical ratings and written narrative comments are seen to not align, in order to determine if the MUM effect is the cause for this.

Strengthening the clinical performance assessment setting and ensuring that trainees receive the assessor's true judgement of their performance would have multiple beneficial outcomes. Assessors would feel more confident to deliver their true assessment. Trainees would receive the assessment they have earned in a more timely manner and so allow remediation to occur where required. Ensuring that robust assessment results are followed with equally robust support for the trainee is crucial to

the process to avoid issues such as trainee demoralisation, increased stress in the workplace and accusations of bullying, which are all key areas of current conversation and could all potentially lead to increased MUM behaviours. The ultimate goal is to ensure that patient safety is not jeopardised through failure to fail.

In summary, the MUM effect offers one explanation for the observation that efforts to eliminate failure to fail and grade inflation to date — such as by training of assessors and modification to assessment formats — have not been entirely successful as yet. In addition, it shows that the difficulties faced with negative message delivery are not unique to our setting but are part of the human condition. As such searching for solutions likely require investigation beyond the medical education context. The MUM effect provides a useful framework to use in future research in order to explore the issue of failure to fail and grade inflation in medical education.

Conclusion

In years gone by, many doctors believed the practice of shielding a seriously ill patient from their diagnosis or prognosis was a benevolent one (145). Similarly, many assessors believe they are acting compassionately when they withhold information that a trainee's performance is below the required standard, despite the evident risk to the trainee's eventual progression and of course to their patients' safety and outcomes. In years to come, we may reflect on this practice as being as inappropriate as the clinical one.

The MUM effect tells us that no news is not necessarily good news (115). It can impact on the delivery of undesirable messages by delaying, avoiding or distorting the news (DAD behaviours). As a robust phenomenon, it occurs across many settings, with many influencing factors. While in some circumstances, measured amounts of MUM behaviours may facilitate effective communication, detrimental effects result when this goes too far. These must be countered, in order that assessment messages can be delivered as intended by the assessor.

Several issues with the current use of the clinical performance assessments have been explored in these first two literature review chapters. As the clinical performance assessments were designed to benefit trainees, the next part of the project considers their perspectives on these formats, focusing on their views on the usefulness of the assessment message provided to them through a systematic review of the literature.

CHAPTER 4

Trainees' perspectives of assessment messages: a narrative systematic review

Introduction

The intended purposes of the clinical performance assessments have been outlined in Chapter 2. This chapter now presents the trainees' perspectives of the assessment messages that they receive in the context of these assessments to see how well these match the aims.

This review is important in the context of the thesis, as the findings will contribute to a deeper understanding of issues with clinical performance assessments from the perspective of the trainees who undertake them. Views on problems with the assessments have been presented thus far largely from the perspective of assessors and the educational institutions, but in order to consider how best to address the problem identified for this study overall, it is important to consider the trainee view also. Thus the question for this systematic review focuses broadly on perspectives of the assessment messages received by trainees and not only on problems, to permit a wide range of views to be collected. This in turn permits an enhanced understanding of problems with the delivery of negative assessment messages which informs the design of the study in Chapter 7 on trainees' perspectives.

This chapter was published in *Medical Education*, 2019 (172). It collates the published English language literature to answer the research question: What are medical specialist trainees' perspectives of the assessment messages they receive in clinical performance assessments? In keeping with thesis with publication requirements, University of Melbourne (<https://gradresearch.unimelb.edu.au/preparing-my-thesis/thesis-with-publication>), "the 'author accepted' version, that is the final draft as accepted by the publisher, including any changes based on referees' suggestions but before it has undergone copy-editing, typesetting and proofing" is included here.

In order to conform to this thesis's formatting, table and figure numbers have been altered from the version of the article submitted for publication.

Abstract

Purpose

To conduct a narrative systematic literature review on medical specialist trainees' perspectives of the assessment messages they receive in the context of clinical performance assessments. The aim of the study was to determine if the trainees value the information received through the formats which are designed to promote their development, and if not, the reasons for this.

Methods

The authors searched the ERIC, Embase, Ovid Medline and PsycINFO databases for articles published up to 16 June 2018 which presented original data on trainees' perspectives of the assessment messages they receive in the context of work-based assessments (WBA) and in-training assessments (ITA) used within their training programme. After screening by title by one author, all authors screened the total of 938 abstracts and 139 full-text articles were assessed. Descriptions of quantitative data and thematic analysis of qualitative data were used to present the opinions of the trainees.

Results

33 articles met inclusion criteria. Twenty-six articles (79%) described trainees' perspectives in the context of WBA with the remaining describing ITA formats. Wide-ranging opinions were reported and the analysis categorised these into three themes: trainees value developmental assessment messages; trainees become disengaged when assessment messages aren't developmental; trainees' views depend on the environment, the assessor and themselves. Some trainees reported the assessment messages were valuable and provide input on their performance to guide their development; but many disagreed. In particular, the trainee's own level of engagement with the assessments influenced perspectives on the messages received.

Conclusions

Trainees do not universally perceive that clinical performance assessments provide them with the valuable developmental input on their performance as they were designed. Factors related to the trainee, the assessor and the environment influence their perspectives.

Keywords

Evaluation/assessment of clinical performance, personal characteristics/attitudes, postgraduate training.

Introduction

Clinical performance assessments – defined here as workplace-based assessments (WBA), in-training assessments (ITA) or in-training evaluations (ITER) – are designed to provide medical trainees with holistic information about the quality of their practice. They recognise that “[t]he highest level of evidence of attainment of a skill is the demonstration of the behaviour in routine daily practice, i.e. performance in vivo.” (173). The workplace setting also allows the assessment of additional attributes (174) and provides trainees with the opportunity to receive qualitative and quantitative judgements on their performance from multiple assessors during the course of a placement. Whether primarily formative or summative in design, the information provided is intended to promote improvement. However, trainees and their assessors sometimes hold very different views regarding the performance information given and received in general (151). It is important to understand trainees’ views on the assessment messages they receive in the workplace in order to know if they are achieving the aim of providing guiding information from the trainee’s perspective. If these assessment messages are not helpful, then reasons for that need to be explored.

We use the term “assessment message” to emphasise consideration of quantitative information, in addition to written or verbal comments, as the former is sometimes not included in feedback research. We also use the term to deliberately distinguish the information provision that is of interest here from more recent versions of feedback, which have broadened beyond the “teacher-specific behavioural principles” of information delivery (e.g. timely, specific) (34) to include the learner’s barriers and

reactions to feedback and the relationship between the teacher and learner (34). Indeed feedback can be seen as a matter of co-construction between teacher and learner (175). Our focus here is on the initial part of the feedback process – the message – comprising both ratings and comments that are delivered to the trainee.

To date, scholarly papers have evaluated many aspects of performance assessments, including their validity or reliability (176), their educational impact and feasibility (177,178) and evidence of their impact on doctors' future work (174). Many of these have studied this from the perspective of the assessor, however, none have determined trainees' perspectives on the information received from the two types of clinical performance assessment together. This is important as perspectives may be relevant to both formats. Bing-You, Hayes, Varaklis, Trowbridge, Kemp and McKelvy recently conducted a scoping review of what is broadly known about feedback in health professions education. Though high-level evidence was lacking, repeated and immediate feedback was seen to be beneficial for learning and the importance of additional aspects including emotions, credibility of the source and demographics was highlighted (179). Massie and Ali (180) performed a critical narrative literature review on medical trainee and assessors' views of WBAs in training, finding widespread negativity towards them, particularly around their perceived purpose and associated logistics, such as time available to perform them and training for those involved. Poor quality feedback to trainees was a common outcome reported.

The goal of our study was to extend these works and bridge the gap between assessment and feedback literature by focusing specifically on medical trainees' perspectives of the assessment messages that they receive in the context of clinical performance assessments. In order for these assessments to achieve their aims, trainees need to value the assessment information they receive. We therefore sought to collate published reports on trainee's perceptions of their worth. The question for this systematic review is "What are medical specialist trainees' perspectives of the assessment messages they receive in clinical performance assessments?"

Methods

Definitions and explanations

“Medical specialist trainees” refers to qualified doctors in a medical specialty training programme: for example, those who are training to become surgeons, physicians or general practitioners. In North American, these doctors are known as residents, while in the UK, Australia and New Zealand, they are termed registrars. For simplicity, we refer to them as “trainees” from here.

“Perspective”, refers to “a way of regarding situations, facts etc, and judging their relative importance” (181). We use this term as it was commonly used in the research question or aims of included articles; it also encompasses a broad range of ideas by including the concepts of attitudes, outlooks and views.

An “assessment message” describes the information provided on the assessment form (ratings, comments) and any verbal information delivered to the trainee by the assessor about their performance.

“Clinical performance assessments” refers to assessments of trainees conducted in authentic settings, by an assessing clinician with the delivery of an assessment message to the trainee. They occur within usual clinical practice, rather than a formal examination or simulation setting, and the assessment content is ‘natural’ and not predetermined and so varies between learners. Clinical performance assessments therefore do not include formats such as Objective Structured Clinical Examinations (OSCE) or similar.

This review was performed with an interpretive stance. That is, we sought to interpret others’ interpretations by inductively grouping results from studies into themes of similar ideas. Ethical approval for this literature review was not required.

Search strategy and information sources

The review followed the PRISMA (182) guidelines for systematic reviews.

The databases used for the search were Ovid Medline, Embase, PsycINFO, and ERIC which were selected in consultation with library staff as being the most likely to yield

references relevant to the research question. In the search design phase, databases were explored for appropriate subject headings relating to the question. The list of participant and interest search terms (English language only) were combined by ‘and’ to generate the search yield (table 4.1).

Table 4.1: Search terms used in database searches

Participant		Interest
specialist trainee* specialist registrar* intern SpR not (surface plasmon resonance or spastic or secondary) registrar* interns learner education, professional/ trainee learners education, continuing/ trainees medical staff, hospital/ schools, medical/ medical trainee internship and residency/ internship health profession* student* adj3 (medical or medicine) residency resident residents exp education, medical/ health personnel/	AND	clinical rotation assessment* end of clinical rotation assessment* clinical performance report* DOPS tool patients' assessment questionnaire* t-mex workplace-based assessment* mini-peer assessment tool in training report* WPBA procedur* based assessment* eportfolio* work based assessment* direct observation of procedural skills in training assessment* multi source feedback in training evaluation* mini-clinical evaluation mini-CEX multisource feedback case based discussion* clinical assessment tool workplace-based assessment* WBA peer assessment DOPS MSF clinical supervision assessment form* performance assessment* CBD not (central business district) evaluation not quality

MeSH terms are denoted by ‘/’

Database searches were undertaken on 26 October 2016, with an update performed on 16 June 2018 (no limit was placed on the start date for the search), yielding 9248 articles. Citations were managed in Endnote X7 (Thomas Reuters, New York, New York). Duplicated papers (n = 3663) and non-English papers were removed (n = 260). Exclusion of articles by title (n = 4391) was conducted by CS. The remaining articles (n = 938) were imported into Covidence (Melbourne, Australia). A double abstract review was undertaken (CS for all and ST, NC and MB approximately one third of the articles each) with any disagreements resolved by discussion and reference to the inclusion and exclusion criteria (table 4.2). Four additional articles were obtained through hand searching of the reference lists of the 32 articles identified for data extraction and one was included for data extraction.

Table 4.2: Inclusion and exclusion criteria for this systematic review

	Inclusion criteria	Exclusion criteria
Participant	Medical specialist trainee (vocational / specialist / postgraduate).	Medical student. Consultant doctor (i.e. Fellow), doctor not in training programme to become specialist. Other health professional groups. School students, US College students.
Interest	Trainee perspectives, views and opinions on assessment messages received from clinical performance-based assessment formats in actual use for assessment e.g. WBA, ITA. Satisfaction only insufficient. Minimum of approx. half of focus of paper is about trainees.	Studies concerning other assessments (e.g. assignments, written exams, standardised oral exams including OSCEs, short and long cases). Assessor opinions on clinical performance-based assessment formats. Trainee opinion is less than half of the focus of the study. Pilot assessment.
Trial design	Data is reported.	Literature reviews, commentaries, letters to editors, non-peer-reviewed works. Non-English.
Publications	Peer-reviewed journals.	Other literature.

Full-text review for inclusion or exclusion was performed by CS with all other authors double reviewing approximately 34% (n = 47/139) of the articles. Double review was performed where there was uncertainty as to an article's inclusion status. Thirty-three articles met inclusion criteria and data extraction and quality assessment was conducted by CS. Double review of quality assessment on 14/33 (42%) articles was performed by MB, ST and NC. Consensus on disagreements was achieved through discussion. See figure 4.1.

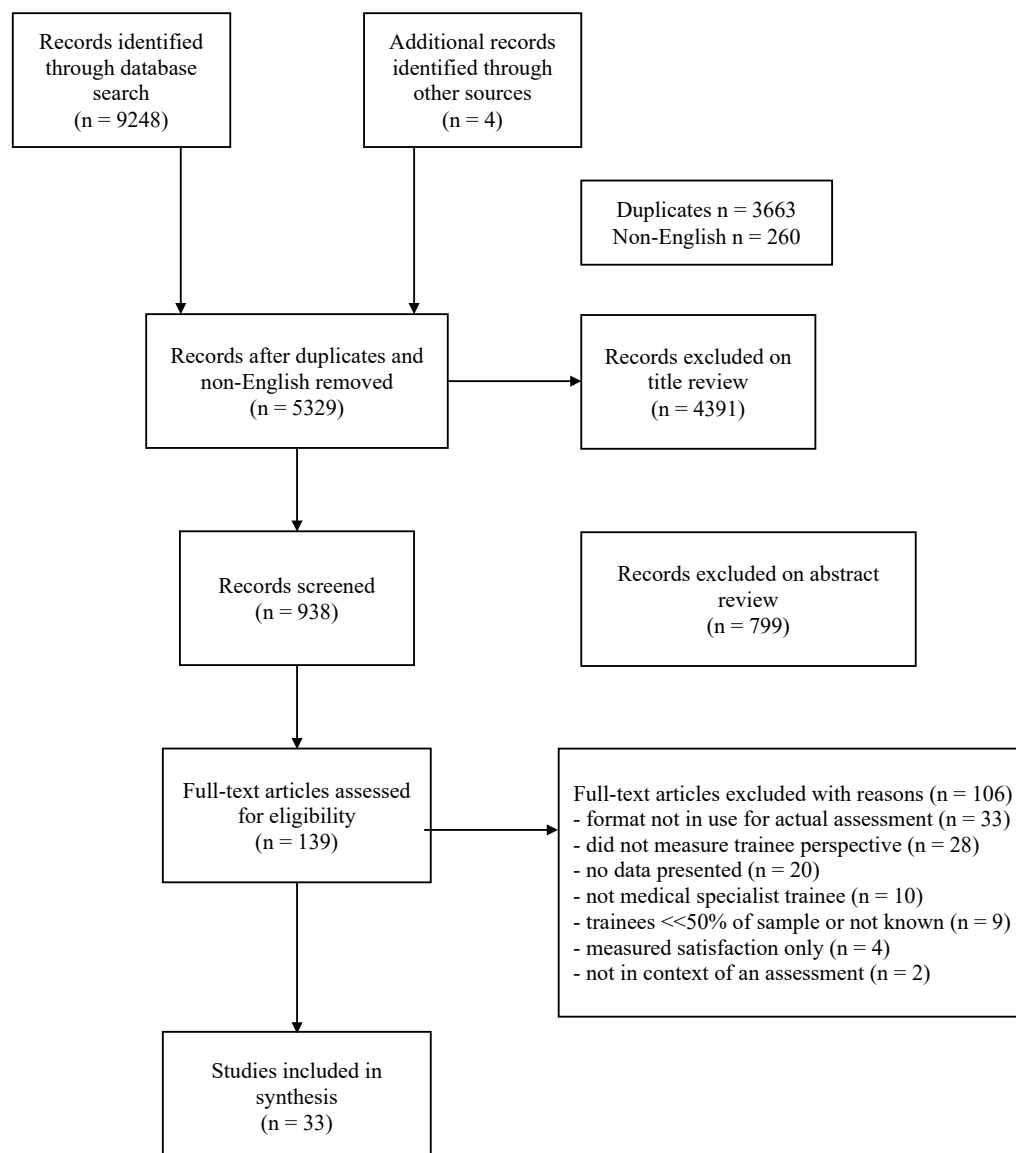


Figure 4.1: PRISMA (*preferred reporting items for systematic reviews and meta-analyses*) diagram for systematic review question: What are medical specialist trainees' perspectives of the assessment messages they receive in clinical performance assessments?

Inclusion and exclusion criteria

Articles were included if they presented qualitative and/or quantitative data describing trainee reports of perspectives on the assessment messages received in a clinical performance setting in current usage in their training program. This included logistical aspects, such as the time spent getting feedback and opinions on its quantity, quality and usefulness. Articles that reported pilot assessments were not included as perspectives on feedback in these likely differ from those in actual use. Articles that noted trainees' satisfaction with or general perspectives on the assessment process only were not included.

In addition to ITA, the WBAs in included articles are:

- mini-CEX
- Direct Observation of Procedural Skills (DOPS)
- Case based Discussions (CbD)
- Multi-source Feedback tools (MSF)
- Procedure Based assessments (PBA)
- Objective assessment of Surgical and Technical Skills (OSATS)
- Multiple Consultant Report (MCR)

Many of the studies included used the term “feedback” to describe what we have defined as the “assessment message” and we have not modified their expression. The reader is alerted to this and asked to be mindful that “feedback” in this context relates to the “assessment message”.

Data extraction

Data extraction from included articles' results section was performed by CS. Each article was read carefully and data were entered into a spreadsheet in Microsoft Excel for Mac Version 15.30 (Microsoft). Data extracted included the author details, year and country of publication, assessment type, research aim, data collection method, number and type of participants, and the qualitative and quantitative data on trainees' perspectives of assessment messages. Supporting quotes were recorded. Quality

assessment was performed using the MERSQI tool (183) for quantitative data and the Walsh and Downe tool (184) for qualitative data.

Data analysis and synthesis

Data analysis was performed through an interpretive paradigm and thematic analysis was the method employed for the synthesis. The data extracted from the results of included studies were organised into open data-derived codes and all instances of codes were collated together; in this way, all studies contributed data to the thematic analysis (185). Similar codes were clustered together into themes and CS determined three categories that ran across the themes. The qualitative analysis from the articles with the highest quality assessment (6,8,102,109,186–189) informed this process. In this way, the tension between the analytic work done by authors and our own analytic work was balanced (190). Additional themes were inductively coded and included in an overall theme as required. MB and CS met to iteratively discuss codes, categories and themes. MB re-read the original data of six of the top papers to review the final categories. Quantitative data were included in this analysis and presented in a descriptive form with the theme to which they relate.

While we did not use the analyses from studies with lower results on the two quality assessment tools used (level determination described in Appendix A – Quality level determination), we did include their reported data. This decision was made post hoc. While we determined that lower quality articles had a valuable contribution to make, they had fewer interpretive approaches to the data and therefore contributed less to our own interpretive processes.

Results

Study Characteristics

The 29 articles included in this systematic review are summarised in Appendix B – Articles included.

Most studies were from the UK ($n = 23$, 70%), with others from Canada ($n = 3$, 9%), Ireland ($n = 2$, 6%), the Netherlands ($n = 2$, 6%) and one (3%) each from Argentina, Australia and New Zealand and India. Twenty-six (79%) articles covered WBA (mini-

CEX n = 4, 12%, CbD n = 4, 12%, MSF n = 2, 6%, PBA n = 3, 9%, OSATS n = 1, 3%, MCR n = 1, 3%, mixed n = 11, 33%) and seven (21%) focused on ITA. Fourteen (42%) of the studies were primarily quantitative in design, 13 (39%) were primarily qualitative and six (18%) were mixed methods. Where studies included qualitative and quantitative data, a quality assessment for each part was conducted. Eight studies achieved ten or more 'yes' scores overall on the Walsh and Downe scale while 11 scored less (184). Seventeen studies scored 7 or more on the MERSQI scale (183), with 3 scoring less than 7.

Thematic analysis

There were three main categories derived from thematic analysis which addressed the research question. These are:

- **trainees value developmental assessment messages;**
- **trainees become disengaged when assessment messages aren't developmental; and**
- **trainees' views depend on the environment, the assessor and themselves.**

These are outlined in turn.

Trainees value developmental assessment messages

Many trainees highly valued the assessment messages they obtained from the clinical performance assessments. They reported that the assessment messages they received were fair (191) and useful (191–194), and promoted deeper understandings and learning (23,195):

"if it is done well you can get an awful lot of learning opportunities out of it and ...it gives you an awful lot of feedback" (109).

They described the assessment messages as beneficial by being real-time and objective (196), though some recognised they could be subjective (191). They noted the assessment messages indicated and documented their performance (5,23,197) so increasing awareness of their strengths and weaknesses (196):

"The feedback was useful, either for confirming what I had done well and for suggesting other things I could have done" (23).

Trainees often ascribed different values to the individual components of the assessment messages, with qualitative information valued above quantitative:

"I think the most important thing about all of these things is the feedback. It's not the mark you get at the end of the day" (8)

For example, one study on the MSF reported that 89% of trainees agreed the comments were useful, compared with 38% agreeing the numeric report was useful (7). Trainees particularly valued the verbal delivery of performance information (198); in Sabey and Harris, 94% of trainees escribed their verbal feedback as useful compared with 74% for written comments (189). This was especially when in-depth conversations occurred (188) and areas of weakness were identified (187,195,196,199):

"... You look for assessors that you know are knowledgeable and where you get something out of it, a good dialogue or really learn something. Not just marks on a sheet of paper ..." (5).

Others highlighted that verbal feedback was more effective in achieving outcomes (such as by addressing specific learning points or self-directed learning (194)).

Written narrative information, with detailed examples, was highly valued, especially when it explained ratings (102) or was coupled with verbal input as to how to work to achieve competencies (23). These high values placed on written comments and verbal discussions with assessors were noted for both the WBA (8,187,191,197,199) and ITA formats (188):

"you can actually figure out why you didn't meet [expectations] or why you were doing well so you can keep doing it" (102).

Some trainees wished to avoid quantitative scores altogether (7) or at least to have a better understanding of scales (188), consistency in their use (189) or a form redesign to focus on developmental feedback (200).

Evidence of an effect on learning was limited and was generally based on self-report alone. For example, between 20–29% of trainees felt they improved practice because of feedback (198,201), with those who reported receiving good quality feedback being more likely to agree (198). Some felt that the assessment messages, often coupled with dialogue, helped their development of reflective skills (191,200,202) and self-assessment skills (188).

Trainees highlighted additional aspects of the assessment messages they valued such as their breadth including coverage of nontechnical skills (187) or giving a more holistic picture of performance (7,8,191):

"What I think is more useful is when you get feedback about how you're performing as a trainee in general" (191)

Some acknowledged the ability of the message to provide encouragement (23) or reassurance (199) and to convey the feeling of being appreciated (189), and some noted that feedback being affirming was a potential outcome in itself (6). Some trainees reported the assessment messages led to an increased confidence with aspects of their work (196):

"It's nice that they have approved that you can do such and such. You get confidence - I know I have been able to do this correctly" (5).

Others highlighted the limitations of positive feedback:

"positive feedback is nice but it's not always going to be productive for you" (109).

In addition, some acknowledged the potential difficulties for assessors with giving negative feedback :

"people always say they want negative feedback, but people don't take it as well... puts the evaluator in a very awkward position" (102).

Trainees become disengaged when assessment messages are not developmental

By contrast, many trainees reported highly unfavourable views of their assessment messages, often noting “widespread dissatisfaction” (102) with the process.

Between 33–41% of trainees reported they rarely or never received verbal feedback immediately after a WBA (193,201), with 16% receiving no feedback in one study (203):

"But the majority of consultants will just sort of go "Oh God, you want one of these again". Fill it in really quickly, give you no feedback whatsoever and it is just a tick box" (109).

Wrongly or rightly, this left some trainees to infer from their feedback:

"...if nobody has criticised you or had a go at you then you know you are alright" (109)

When trainees did receive feedback, not all reported it to be useful or helpful (199,204):

"...it's a box ticking exercise. I only do it because I have to. I don't do it because I think it benefits me in any way." (109)

Trainees reported variability in rating (186,187) and that assessors vary in their level of leniency (187).

Many trainees noted frustration at a lack of explanation of written or verbal comments (199):

"They don't mention anything in the notes or just one line and that doesn't show you a great deal as in which areas there are that you should improve and that sort of leaves you in a state of why did they mark you poor?" (54).

and poor quality comments were also noted (102,194):

"It's all just platitudes" and "They'll write something a bit vague and a bit meaningless" (189)

Some trainees doubted the ability of the assessments to reflect their progress (205) or to detect struggling trainees (8,186,187) both in the ITA, where 39% of urology trainees in one UK study believed they would not identify a failing trainee (206) and WBA formats, where only 45% of GP trainees in the UK agreed that WBAs could detect struggling trainees (189).

"Generally they are a waste of time for a good doctor. They won't be useful to differentiate good from bad doctors" (186).

Some trainees noted that assessors could concentrate on unrepresentative episodes (102) or be deficit-focused (6), noting that feedback delivered poorly can be “demotivating” (23) and if too negative, counterproductive (54).

These perspectives lead to a view of such assessments as “form-filling” (198) or “tick box” initiatives (6,109,186,202,204,207):

“Now it is just so people can tick the boxes. The inspiration and pride in our profession has gone” (186).

Trainees’ views depend on the environment, their assessors and themselves

The reviewed articles highlighted factors that influence the trainee’s perspectives on the assessment messages they have received. These included the environment, the assessors and themselves.

Environment

The time available for feedback influenced trainee perspectives. Many noted a lack of dedicated time (194) and the consequent potential for poor quality feedback with “limited educational value” (191). They reported that feedback usually took less than five minutes (for the CbD (193), PBA (203) and WBA in general (201)) but in some studies, only the minority (14%) felt time for feedback was insufficient (199), while other studies reported this varied with the trainees’ level. For example, the majority (67%) of senior trainees in Tsagakataki and Choudhary’s study felt that feedback time with the OSATS was insufficient (204).

Perspectives also depended on timeliness of receipt of assessment messages. While some trainees noted completion of the form and delivery of feedback was mostly timely (198), others disagreed, reporting delays in receiving assessment messages (188,193,194,201) or difficulties with getting the forms completed in full (186) or at all (54):

"The main obstacle or issue I find is from higher levels and not actually from me getting them achieved...finding consultant input to sit down and go through them with to sign them off was a difficulty...it's so difficult to get them..." (8)

The setting was important, with trainees reporting the negative impact of interruptions and inability to speak freely (191) on feedback occurring in a public place.

Assessor

Many trainees felt that the “style” of trainer impacted heavily on the assessment message provided (202). The background of the assessor was relevant: for example, GP assessors were seen to be more skilled than hospital based ones (189). The seniority of the assessor was contentious; some reported that senior registrars provided better feedback than consultants (207), while others reported that consultant feedback was more likely to be constructive (205) and honest:

“[r]egistrar feedback tends to always be positive, but consultants will tell you how it is” (202)

Assessor enthusiasm in the process was very influential, with almost 80% of trainees in one study reporting that “the enthusiasm of their trainer for PBAs determined the extent of benefit to be gained from using them” (198). This engagement was highlighted in many studies (102,188,191,202) and trainees reported that feedback from an engaged assessor is more valuable :

"If you choose someone who wants to teach and can teach, it makes a big difference. If they just fill the paperwork in then you're not going to get anything from it" (191).

Unfortunately, large numbers of trainees felt that assessors were not engaged in assessments (186,189):

"[t]he trainers aren't engaged...the majority...they'll just help you in order to get by" (109)

The assessor's skills in giving feedback were reported as very important for feedback to result in learning and improvement (23). Some trainees reported concerns about assessor motivation and how personal preferences or biases (197,205), victimisation (199) and retribution (191) could affect the assessment message.

Many trainees raised concerns about the level of honesty in their assessment messages. This stemmed from the influence of interpersonal relationships (54,189) and from poor assessment practices such as retrospective completion (6,186,187) or completion of parts of the form related to unwitnessed tasks (188,189):

"... they just go through and fill them out anyway" (188)

Half the participants in one study reported they had been told to complete their *own* forms (8). Some trainees doubted the honesty of messages delivered in face-to-face settings such as mini-CEX and CbD (189) and were aware of assessor reluctance to document deficiencies (109):

"What you want from an assessment is someone to actually tell you what your fault is" (194).

Themselves

Perspectives on the assessment messages were further influenced by trainees' own characteristics and behaviours. Training level influenced views of feedback by tool. For example, in one study, 52% of senior trainees felt CbDs were more useful at junior levels (202); while the MSF was reported to be useful in all years of training (7). Some senior trainees highlighted their desire for different types of feedback:

"I am wanting to be assessed more globally. A sort of overall impression of 'am I almost ready for consultant practice or not' " (8).

Many studies showed that trainees influenced their assessment messages strategically. Those who didn't just want "marks on a sheet of paper" (5) chose assessors they knew were serious about assessments (5), and whom they anticipated would give them quality feedback (187) and take time to go through the assessment with them immediately afterwards (198). Others focused on ratings only (189) or "gamed" the system (8,187), seeking out assessors known to be lenient in their assessments (109):

"If I know a person wouldn't rate me high, even for a teaching thing, I would rather find someone else. In terms of real development, that's not how it should be" (200).

Others avoided assessors they felt gave negative feedback especially when the assessment carried implications such as supporting progress decisions (109):

"Like I don't get certain consultants to do things particularly because arguably they're more critical on paper and then in an ARCP environment it then looks bad on me" (109).

Some strategically chose cases (8,109,187,189,191), with between 2–11% of trainees agreeing they often /always avoided a "complex or difficult cases owing to the fear of receiving negative feedback" (193,201):

"I actively pick a procedure that I know I did well in and ask someone to rate me on it" (109)

Some acknowledged the consequences of this on the assessment message:

"I rarely bring forward patients that I did badly with and so rarely get negative feedback. So it's useful but there's not a lot of constructive criticism to go on most of the time" (191).

Discussion

Summary of findings

This review collates the perspectives of medical trainees on the assessment messages they receive through clinical performance assessments. While these assessments have a well-established position in the program of assessments undertaken in many speciality training programs and are designed to provide trainees with valuable input on their performance, this review shows that trainees do not universally agree that the assessment messages they receive achieve this aim. Instead, it shows that trainees' perspectives can vary widely even within the same study (186,191) and depend on multiple factors.

Their own and their assessor's engagement in the assessment process were seen to be major drivers on perspectives on the assessment messages received. The assessors' enthusiasm and engagement was fundamentally important (102,188,191,198,202) and predicted the quality of assessment messages. Unfortunately, many trainees cited examples of disengaged behaviour by their assessors (109,188,189,193,201), and so had poor opinions of the resultant assessment messages. However, in contrast, where assessors were engaged, trainees often valued their messages highly (5).

The assessor's engagement often predicted the trainee's engagement, and further influenced perspectives on the messages received. Three studies (54,109,187) highlighted the polarisation of trainees into two groups – those who valued formative feedback and sought quality information to promote their learning and development (learning goal oriented); and those who were results-focused, primarily interested in receiving positive comments and high marks (performance goal oriented). The former group clearly expressed their desire for more information – for explanation of their assessor's views on their performance (102) and their frustrations when this does not occur (54). They recognised a lack of honesty (189), calling for more truth in the assessment messages provided (194,202), though they acknowledged the difficulties of doing this (102). Reluctance to deliver negative messages is well studied in the psychology literature (113). The performance-orientated group focused on their goal of just getting the assessment ticked off (109) rather than deriving any educational benefit. This aligns with work from Dijksterhuis, Schuwirth, Braat, Teunissen and Scheel who

distinguished performance-orientated from mastery-orientated behaviours in assessors and trainees in general formative assessment situations (208).

These different trainee behaviours may not stem solely from innately different trainee types. Some trainees may be fundamentally performance goal orientated, while the circumstances of the assessments may alter the orientation of others. Ongoing resentment to the introduction of these assessments has been noted (8), and challenging environmental circumstances exacerbate this. Lack of time and an appropriate location for feedback (191) promote a cursory approach to them, as may imposing quotas for completion (8); most assessors and trainees do not think that increasing numbers of assessments leads to improvement in clinical skills (209). These factors coupled with assessor disengagement in the assessments would likely challenge the most enthusiastic learning goal orientated trainee to become more performance orientated.

Implications for practice

Identification of the problems with assessment messages from clinical performance assessments is the first step is to ensuring they function as intended.

The lack of engagement, in part related to ongoing resentment at their being “dropped on trainees and [assessors]” (8), may be reduced by reorientation to the assessments, with the close involvement of trainees and assessors. This should include their input in defining the purpose of the assessments in each training context and so minimise a tick-box view of them (6,109,186,202,204,207). Addressing trainee desires for written and verbal comments on their performance and increasing emphasis on constructive comments is important (97), and others have also reported that feedback is often lacking (210–212). Comments could be provided alone, as introduced in some settings (100). Training assessors in assessment and feedback is clearly necessary for production of quality assessment messages, but similarly, training of trainees in how to incorporate assessment messages into practice may be required. That is, while many of the notions about co-construction through feedback (175) are evident in the data (particularly through third theme Trainees’ views depend on the environment, the assessors and

themselves), many assessors simply write a message for trainees and both assessors and trainees should understand how to maximise the value of this performance information.

This study highlights that environmental aspects of these assessments are equally important to address and neglecting to do so will limit their potential. The training organisation (97) must provide more support, the number of assessments to be completed must be balanced and justified (for example (209)) and dedicated time to conduct quality assessments provided. Some have implemented ways to address this, for example with dedicated WBA clinics (213), but time and service pressures remain a significant challenge for most and other options need to be developed.

In addition to optimising environmental aspects, recognition of the difficulties of delivering negative assessment messages is needed. The pervasive human tendency of reluctance to provide negative information (113) needs to be acknowledged, and emphasising the educational aspects of these tools may encourage assessors to deliver a clear assessment message, even when it is ‘negative’. Some authors have recognised how the culture of medicine and “reluctance to reveal errors and shortcomings” (109) may also impact on the delivery of assessment information and this is an important area for further research from an educational perspective.

The use of clinical performance assessments in a training program is only one step in the process of providing trainees with assessment messages that contribute meaningfully to their development. This review combines the assessment and feedback literatures and has shown that without commitment to and engagement in the assessments by the trainee and their assessor and the provision of environmental support, resulting assessment messages will be of little value. Attending to the formative purpose of the assessments, by not only providing timely and specific but also quality information based on performance, that leads on to a feedback exchange, is key to their benefits. Without this, the message will be disregarded at best and invalid at worst, with potential ongoing harmful effects, as trainees who are unaware of their true level of performance may not improve and subsequently fail a future examination or — of more concern — cause harm to a patient. Promoting such engagement and attention

to the intended outcomes of these assessments is essential if they are to function as designed and help trainees to develop into the best doctors they can be.

Limitations

This review has several limitations. The inclusion criteria limited the search to studies published in English, meaning relevant articles in other languages were missed. This was done for logistical and financial reasons. The focus on medical specialist trainees excluded others, including junior doctors not in training programs and medical students. This was done as the consequences of assessments are different for each, and therefore perspectives on them may differ. Similarly, although four databases were searched, it is possible that relevant articles only listed in other databases were missed. Exclusion of articles by title and by only one researcher had the potential to overlook relevant articles. Similarly, data extraction by a single researcher is a further limitation and may have missed relevant information. Most of the articles included were from the United Kingdom and this “UK-centricity” of works on WBA in particular has been noted in previous studies (180).

Analysing and presenting qualitative and quantitative data together is challenging and different authors may have emphasised different areas. As highlighted, eight articles formed the framework for the analysis and while these are of higher quality, this may have unduly influenced the review.

As noted in several of the included studies, surveys of people’s perspectives tend to attract responses from those with polarised views and the data subsequently included in this study may be overly representative of such groups. However, opinions at all extremes were noted. Social desirability bias encourages socially acceptable answers from participants and is of particular relevance for data extracted from focus group or interview settings.

Conclusion

This systematic review collates the perspectives of medical trainees in speciality training programs about the assessment messages they receive in clinical performance assessment contexts. We believe this is the first collation of such opinions and they are crucial to consider in relation to further developments of these important assessment formats. Some trainees report their assessment messages are valuable, but some report they are not. This appears to stem from both their and their assessors' engagement, and importantly, environmental factors. Addressing aspects of all three will be needed for these assessments to realise their potential benefits.

Implications of this study

This literature review has important implications for the thesis overall. Firstly, by documenting what is reported in the literature about medical specialist trainees' perspectives of the assessment messages they receive in the context of the clinical performance assessments, the review provides a point of comparison for the views of the trainees in the current study setting. This is important as the vast majority of studies were set outside of Australia, so results may differ to this context. Secondly, as the literature review has revealed that some trainees are aware that at times, their assessors are not completely honest in their delivery of assessment messages and may withhold assessment judgements, this review helps to frame the design of studies for this work. Factors revealed in the MUM literature which may modify reluctance to deliver messages, such as trainee desire for information (150), in combination with other influences on assessor behaviour, such as trainee reaction to feedback (94), are used to guide the development of the studies in the next sections. Thus this literature reviews together with the MUM effect review inform how to explore trainees' views about reluctance to deliver assessment messages in the study context by viewing them as MUM behaviours.

Summary

This chapter has presented a summary and interpretation of the published literature on medical specialist trainees' perspectives of the assessment messages that they receive in the clinical performance assessment context. This shows that opinions are wide-ranging but that trainees do not universally value the messages that they receive. They recognised the contribution of the environment, their assessor and themselves to the outcome achieved, and at times questioned the honesty of the messages they received.

The preceding three chapters have summarised the current assessment climate in medical education and highlighted some of the difficulties that are occurring at present. There is strong evidence that the clinical performance assessments are currently not always achieving their aims of providing helpful and accurate performance information to trainees. Examples of poor practice around the use of these assessments have been presented, and possible explanations and understandings for the difficulties involved in delivering negative assessment information are seen. As such, these literature reviews have provided the background for the design of studies to answer the research question of "How and why does reluctance to deliver negative assessment messages manifest in dermatology training in Australia?"

CHAPTER 5

Research Methods

Introduction

This chapter presents the methods used in the project. It discusses the theoretical position taken for the work, and introduces the methods which were used to answer the research question, and presents the overall methods of analysis. The methods are revisited and discussed in further detail in the individual study chapters to which they relate. This chapter also discusses the ethical considerations relevant to the project.

Theoretical framework

The theoretical framework or paradigm, which “sets down the intent, motivation and expectations for the research” (214) for this project, is the pragmatic one.

The research question, which focuses on the “how” and “why” of the problem with the delivery of negative assessment messages, has driven the research methods. An amount of measurable or quantifiable information was required for the initial part of the study and as such, there are elements of the project which are situated within a post-positivist paradigm. This is consistent with my background in medicine, and medical and scientific research and understandings are largely said to exist in a positivist paradigm. Similarly, most of those working in medicine, including the assessors and trainees to which this study refer, largely understand or conceptualise knowledge with reference to this paradigm. This positivist paradigm, which is also seen to have dominated research in social and educational topics in the first half of the twentieth century (215) and which some authors assert “is still the dominant framework of many medical educators, researchers and practitioners” (216), asserts that reality exists irrespective of observers. As such and in order to document the existence of grade inflation and failure to fail in the study context, methods such as questionnaires and a review of real-world submitted assessment forms were appropriate methods to employ to answer this aspect of the research question.

However this work overall is not positioned entirely within a positivistic paradigm, nor with the ontological assumption of realism and the existence of a solitary truth (217) or answer to the research question. Nor does the work propose that “reality is static and fixed” (216) as this may vary across time and context. As the creation of knowledge from the perspective of those involved within a social and cultural setting — here being medical specialist training — is key in this project; relativism — with the recognition of

“multiple constructed realities” (217) — is an appropriate stance. In keeping with this, there was no attempt to “remove subjective influences on knowledge production” (217) in this study and in fact, it is believed that they contributed to the knowledge created, by virtue of the shared understanding that the candidate held with the participants, both as a previous trainee and now as a Fellow of the ACD.

As such, the methods of data collection for the study include both ones traditionally viewed of as quantitative and ones viewed of as qualitative, as these were deemed the most appropriate to answer the research question. While some argue that both methods can exist within a positivistic paradigm (217) and more recent work extends the association of mixed methods and pragmatism beyond that of practicality (218), this work does not believe that knowledge will be uncovered, but rather created through the conduct of the research project.

Further, given that many of the stakeholders in this project — the clinicians and trainees — likely embrace largely positivistic views of knowledge, it is important that the work be both accessible to and accepted by them. As presented, quantitative methods alone were deemed insufficient to answer the research question, which required an exploration of underlying reasons and experiences. Hence pragmatism, which is “not committed to any one system of philosophy or reality” (214), but is a new paradigm that does not rely on “metaphysical assumptions about ontology and epistemology” (218) is most suitable for the project. That is not to assert that this research is “theory-free” (219), but focused on the knowledge formation “from taking action and experiencing the outcomes” (218).

Methods of data collection

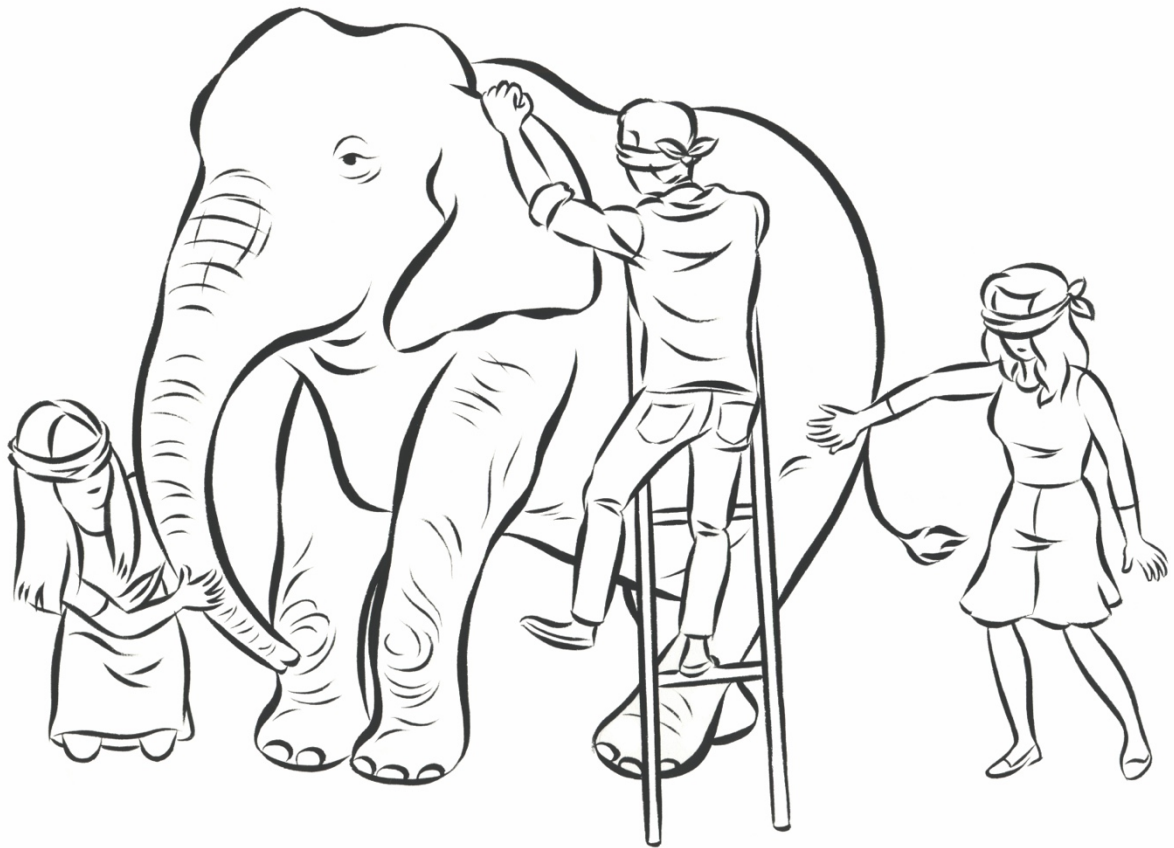


Illustration 5.1: The Elephant and the blind people (illustration by Jo Martin)

As is depicted by the illustration, the story of the elephant and the blind people (with each one imagining a different animal), has relevance for the current study. This Indian fable teaches the importance of considering multiple perspectives and experiences, and that focusing on one piece of information to the exclusion of others will limit the understanding that can ultimately result. As such, and following the literature reviews, and with the aim of exploring the issue from multiple different perspectives, four studies were designed in order to answer the research question. Their results will be synthesised together in Chapter 10, Discussion, and aim to provide a sum that is greater than the individual parts.

In keeping with the pragmatic research paradigm, the studies for the project included both quantitative and qualitative methods, which aimed to consider the question from multiple perspectives and sources, and to provide practical and meaningful answers to

and suggestions for future directions for the issue of reluctance to deliver negative assessment messages (217,220). The studies and associated data collection methods were:

- A questionnaire measuring assessor experiences of discomfort in various aspects of their assessment role and their self-reports of reluctance to deliver negative assessment messages
- A questionnaire measuring trainees' impressions of their assessors' behaviours in relation to delivering negative assessment messages of clinical performance assessments and their views of the clinical performance assessments
- A review of written comments and numerical ratings in sample of real-world assessments in order to compare the message delivered through both parts of the assessment
- Interviews with assessors to further explore their experience with clinical performance assessments and the delivery of negative assessment messages.

These are outlined in brief below, with further detail provided in the relevant individual chapters. The questionnaires were distributed first and the audit of submitted assessments followed this. The interviews were conducted after the data had been collected for these first three studies.

Questionnaire – assessors

A three-part questionnaire was designed for assessors. The first part collected demographic information. The second part used a modified scale, the Performance Appraisal Discomfort Scale (221), to explore the impact of discomfort on assessment behaviours, as this was one factor identified in the literature as potentially relevant to MUM behaviours. The third part used a modified scale to measure assessor self-report of MUM behaviours (132). Additional questions which asked directly for self-reported history of failure to fail and grade inflation and of being approached by trainees in relation to assessment results were also asked. The questionnaire was delivered electronically through the SurveyMonkey platform and was distributed and open for completion between August and September 2016. The results were analysed statistically. The full questionnaire is presented in Appendix C – Questionnaire – assessors.

Questionnaire – trainees

The questionnaire distributed to trainees also comprised three parts. The first part also collected demographic information. The second part used a modified scale, the Instructional Feedback Orientation Scale (169), to explore trainees' views of various aspects of the clinical performance assessments, thereby extending the systematic review chapter. The third part used the same questions as distributed in the assessor questionnaire to explore trainees' views of MUM behaviours by their assessors (132). As with assessors, the questionnaire was delivered electronically in the SurveyMonkey platform and was distributed and open for completion from September to October 2016. The results were analysed statistically. The full questionnaire is presented in Appendix D – Questionnaire – trainees.

Comparison of comments to ratings in the Summative In-Training Assessment (SITA)

While the first two studies used a questionnaire format to collect self-reports of behaviours, this study focused on a review of real-world data using actual previously submitted examples of assessments. In particular, it aimed to compare the assessment messages conveyed by the numerical ratings component with that of the written comments recorded in the SITA. It also aimed to explore the features of the comments, including the word counts and the number of themes/topics addressed, and suggestions for development provided. The SITA and SITA summary forms are presented in Appendix E.

Interviews

Interviews with assessors formed the final study in the project. These were conducted in order to extend the largely quantitative data collected with the first three methods and explore the results in greater depth. The interview format was chosen over focus groups as the logistical issues of arranging a group that several assessors could attend were felt to be too great. In addition, it was felt that participants may feel more comfortable discussing issues in a one-to-one setting. The interview schedule is presented in Appendix F – Interview schedule.

Setting for the study

The study was positioned within the Australasian College of Dermatologists (ACD). This college is the national organisation for the training of medical practitioners to become dermatologists, and is the only organisation in Australia that is accredited by the Australian Medical Council to train and assess dermatologists. The ACD comprises over 500 Fellows who practise in all states and territories of Australia, and at the time of the study, there were 114 trainees in the ACD training program. These trainees have undergone a competitive selection process and must successfully undertake the components of the program run by the ACD in order to be eligible to apply for Fellowship. These components include:

- Completion of the required training time in an accredited training position each year (being 46 weeks in each full-time year)
- Attendance at required workshops and conferences
- Completion of Clinical Sciences Online Competency Modules
- Completion of portfolio requirements, including WBA and research requirements
- Attainment of a satisfactory SITA for each period of training (usually in June and December each year)
- Attainment of a pass in the Fellowship Examination.

Training sites, where trainees undertake rotations of usually 6- or 12-months duration, are located in public hospitals and some private settings and all are accredited by the ACD. Trainees are supervised and assessed in these rotations by Fellows of the College as follows. Each state has at least one Director of Training (DoT) who is responsible for the overall training program in that state, including the allocation of trainees to individual training positions. Each training site has at least one Supervisor of Training (SoT), who is responsible for day-to-day training, supervision and assessment of trainees. The SoT meets with trainees at the start of a rotation to plan and set goals for the period and also meets with them in the SITA meetings. Clinical Supervisors (CS) at the sites also supervise and may assess trainees in a WBA. The trainee is responsible for arranging this. WBAs are completed by all trainees in each year of training and include the Case-based Discussion (CbD), Dermatology Clinical Evaluation Exercise (Derm-CEX, based on the mini-CEX) and Procedural Dermatology Assessments (Pro-

DA, based on the DOPS). Trainees must successfully complete a minimum number of WBAs in each year of their training. In addition, in order to progress, trainees must successfully complete the SITAs required of their training position. At the direction of the SoT, two CS anonymously and individually complete SITA forms on trainees and these are collated with that completed by the SoT and then presented to the trainee in a face-to-face SITA meeting. The completed summary form is submitted to the ACD as a record (Appendix E). The Fellowship exams are held annually and trainees in their fourth year of training or above who have completed all portfolio requirements are eligible to sit the examinations (222). See box 5.1 for a summary of terms (222).

General definitions

Assessor

Refers to those who perform an assessment, for example senior clinicians, consultants, faculty members. Can also be peers in some circumstances.

Trainee

Doctor in a training program.

ACD related role definitions

Directors of Training (DoT)

Fellow of the ACD who is responsible for overseeing the organisation and implementation of the Dermatology Training Program in their State Faculty.

Supervisors of Training (SoT)

Fellow appointed by the health institution where they are attached who oversee the Training Program and supervise and assist with day-to-day teaching, mentoring and supervision of trainees in their department.

Clinical Supervisors (CS)

Fellows who are selected by SoTs and provide on the job teaching, mentoring and feedback to trainees.

Clinical performance assessments in use

Summative In-Training Assessment (SITA)

Assessment completed by the clinical supervisors of a trainee designed to support the trainee's development and provide an assessment of their performance at the end of a rotation.

A minimum of two are completed per year and the trainee and SoT meet to discuss their performance. The SoT determines whether the assessment is "satisfactory" or "unsatisfactory".

Work-based Assessments (WBA)

Competence-based assessments conducted in the clinical setting which reflect everyday performance and the curriculum. These include:

- Dermatology Clinical Evaluation Exercises (Derm-CEX)
- Procedural Dermatology Assessments (Pro-DA)
- Case-based Discussions (CbD)

Box 5.1: Definition of terms

As discussed, this study focuses on the clinical performance assessments, which in this context comprise the WBAs of the Derm-CEX, the Pro-DA and the CbD and the SITA (see Appendix E for examples of each assessment form). All assessments have two

components: a numerical rating scale and an area for written comments. WBAs are completed with a single assessor, and SITAs are usually completed by three assessors (being two CSs and the trainee's SoT). In order to be deemed competent in the WBA, the trainee must be given the rating of "4" on each criteria specified on the individual form. For their performance to be recorded as satisfactory on the SITA, a trainee must have completed the required number of WBAs commensurate with their year of training and not receive "2" or below from two or more CSs on two or more of the 25 criteria. At present, a total of four CbDs, three Derm-CEXs and depending on the year of training, either four (year 1 or 2) or three (year 3 or 4) Pro-DAs are required to be successfully completed each year. Trainees in first year can be marked unsatisfactory in their SITA if they fail to complete the necessary online training modules (Pharmacology, Clinical Sciences and Evidence Based Medicine). A trainee may make as many attempts that are necessary to become competent in a WBA. The Training Program Handbook notes that "being unsatisfactory is not an indication of poor performance but the opportunity to fine tune a procedural/clinical skill" (222). There are no consequences for having to attempt an assessment more than once, and only the record that competency has been achieved is recorded in the trainee's online portfolio. A trainee who is deemed unsatisfactory on a SITA may be required to enter a Supplementary Supervision Program (SSP), which is prepared by their DoT and SoT, and may involve an individually tailored additional supervision and assessment program (Performance Improvement Plan, PIF) and the timeframe over which this is to occur. Failure to complete the requirements of the SSP can result in a trainee being exited from the training program. A summary of the SITA process is presented in Appendix G.

Participants

Participants for the studies were assessors (Fellows) and trainees of the ACD. Participation was open to all Fellows and trainees who satisfied the criteria for participation, being recency of participation in assessments as outlined in specific chapters. There was no limit placed on the number of participants in any study and all those who contacted the candidate and wished to participate in an interview were able to.

Limitations of the methods

All methods of data collection have limitations and the specifics are discussed in the individual results chapters, Chapters 6 to 9, however, some general comments that are relevant to all are presented here.

Firstly, it is recognised that the setting for the study places limitations on the transferability of the results and outcomes of the study. The cultures within medical specialist training colleges and educational institutions broadly are likely to vary, meaning that the generalisation or applicability of the results obtained here may be limited. While the ACD was chosen in part for the study due to its national program, even within it, it is likely that local cultures and ways of doing things apply. As such, the relevance of the results of this work to other settings may need to be considered on an individual basis.

The methods of data collection place further limitations on the results. While a mixed methods approach allows a more complete understanding of a problem than either method alone could produce (223) and so ideally present more nuanced proposals for solutions and the guidance for future research, each method has restrictions. For example, questionnaires are a convenient method of data collection, with the ability to reach many, however they are subject to various limitations, including not getting the “full story” (220). Errors including nonresponse error, where the characteristics of responders and non-responders differ (224), are relevant and some authors argue that those with more extreme experiences are more likely to participate (225). Further, while interviews can collect rich data and permit an expansion in the depth of information collected (226), it is recognised that they are more challenging to conduct and analyse, and biases affect both the interviewer as well as the participants (220).

The choice of sampling technique is relevant and influences the results obtained. Purposive rather than convenience sampling may have increased the number of participants, for example if directors of training or SoTs from each state had been directly invited to participate. The offering of incentives for participation in either the interviews or the questionnaires may also have been considered as a method to increase participation, but this was decided against for several reasons. These included a lack of

available funding but more importantly, the recognition that the use of incentives may influence the respondents who participate and so the results obtained.

Ethical considerations

Approval to conduct the project was requested and obtained from the Department of Medical Education, Human Ethics Advisory Group (HEAG), University of Melbourne, ethics ID 1545651 dated 3 Aug 2016. The ethics application involved outlining the aims and objectives of the project, the key questions and the research design, including the methods to be used. Written permission to set the study in the ACD was obtained from the CEO of the ACD, Mr Tim Wills.

During the course of the project, two amendments to the ethics application were submitted and approved. The first was for an extension to the advertisement period for the Assessor questionnaire aiming to increase the response rate, and the second was to notify of the change of address of one of the supervisors. The project was conducted under the supervision of the three supervisors and regular face-to-face meetings (approximately monthly) were held with them to report on the progress of the project. There were no breaches to the approved project, and annual reports and documentation were submitted in accordance with the ethics protocol.

Informed consent

Informed consent was obtained by all participants prior to participation. For the assessor and trainee questionnaires, the plain language statement appeared at the start of the questionnaire, and informed consent was implied by submission of the questionnaire data. For the assessor interviews, the plain language statement and consent forms were completed in hard copy prior to or around the time of the interview being conducted. Consent Forms were kept in a separate file from that of the data obtained.

Recruitment

Invitations to participate in all parts of the study were made via email sent by the Honorary Secretary, ACD. As such, contact email addresses were not required for nor obtained by the candidate.

Risk concern

The two main risks of the project concerned identification of participants and disclosure of behaviours. These were managed as follows.

Possible identification of participants:

In comparison to other colleges, the ACD is a relatively small one, with small numbers of assessors and trainees. While any Fellow is eligible to conduct a work-based assessment with a trainee, for a variety of reasons, not all do. This means that of the 506 Fellows (at the time of the study), not all were eligible to participate in the study, which reduced the possible number of participants. The number of Fellows eligible to participate was estimated to be 135 as outlined in Chapter 6 in the Results section. All 114 trainees, however, were able to participate in the part relevant to them as all had completed at least one assessment (which are usually performed in June and December each year) when the questionnaire was distributed.

To address the risks of possible identification of participants, several methods were used. Firstly, questionnaire data was collected electronically (SurveyMonkey was set to “anonymous” distribution) and stored in a password protected file in accordance with the ethics requirements. Interview transcriptions were professionally transcribed and stored in password protected files (demographic data was separated from interview data) again in accordance with HEAG requirements. For both questionnaires and interviews, minimal demographic data was collected, and participants were permitted to not provide any of this information if they wished. Any identifiable material or information that was submitted was made anonymous by the candidate and interview participants were referred to by a code. The transcribed and deidentified interviews were sent to interview participants prior to the analysis of data. As will be discussed, some participants made adjustments to the transcripts and these were the versions used for analysis. Interview participants were also shown the first draft of the interview results and none requested any changes. In regard to the study involving the review of previously submitted assessments, the ACD made the forms anonymous prior to releasing them to the candidate.

Disclosure of behaviours

As the project asked both assessors and trainees to reveal instances where their behaviour may have deviated from what would be considered to be ideal, the potential for this to then precipitate feelings of regret or discomfort in the participants was considered. This was felt to be unlikely to lead to significant issues, however the contact details of the lead supervisor for the project and the candidate were given at the end of the questionnaires and in written forms to interview participants if they wished to discuss issues arising from the study further. No participant made use of this, though it was also offered verbally to interview participants who expressed high levels of emotion during the study.

There were no potential risks to non-participants, and neither participation nor non-participation in the studies were known to anyone other than the candidate. Similarly, there were no direct benefits for individuals from participating in research, though it was anticipated that involvement or knowledge of the project may prompt assessors to reflect on their assessment practices. This occurred with one interview participant, who stated that their assessment practices had changed due to awareness of this project. As discussed, participants were not offered any incentives for participation in any part of the study. The overall and global benefits expected for trainees, the educational institution and ultimately the public by improving the ability of assessors to provide their most honest feedback on performance to trainees was felt to greatly outweigh any risks of the small sample size, such as the risk of participant identification or the risk of regret of previous behaviours by participants. Participants were made aware that results of the project would be published in peer-reviewed journals and presented at relevant education and dermatology conferences, as well as being presented in the PhD thesis of the candidate. The data files and hard copies of materials will be stored for a period of five years after the last publication or public release of the research in accordance with the HEAG approval.

Funding

There was no external funding for the project, however the candidate was the recipient of a Research Training Program Scholarship (formerly known as Australian Postgraduate Award) for its duration.

Conflicts of interest

The candidate is a Fellow of the ACD and was previously involved in the supervision and assessment of Victorian trainees at four training centres. From the beginning of the research project (February 2015), she did not participate in any assessment activities in order that no trainee should feel compelled to participate, nor be concerned about their assessment should they not participate. The candidate was also a member of the ACD Academic Research Committee (ARC), which oversees the research projects that trainees are required to undertake during the four years of their training period.

Members of the ARC perform several tasks including to assess the suitability of the trainee's project plan. It was decided by the then chair of the committee, Associate Professor Sam Zagarella (who is a member of the Advisory Panel), that the project plans would be submitted in an anonymous format to the committee, who as a whole will judge the submission. As such, the candidate had a contributory role only to decisions on trainees' research project plans. In addition, the candidate was excused from participating in any decisions which related to recognition of prior learning application in relation to these trainee research projects during the study period.

From January 2018 to July 2019 the candidate was appointed to the role of Trainee Welfare and Engagement Officer for Victorian trainees. This role included providing pastoral care and support to trainees, and planning of their educational activities and work logistics. Data collection for both the assessor and trainee questionnaires had been completed by this time.

Analysis of data

Analysis of data proceeded according to the type of data. Broadly, quantitative data — that being most of the data generated by the questionnaires and the review of the SITA — was analysed statistically using IBM SPSS Statistics Version 25. This included descriptive statistics, including determination of means, confidence intervals and *p-values*, and comparisons of responses and means between groups and proportions (percentages). Statistical analyses were undertaken using Excel and IBM SPSS Statistics Version 25 and with the guidance of Dr Sue Finch, Statistical Consulting Centre, University of Melbourne and Dr Kate Reid, Department of Medical Education, University of Melbourne.

Qualitative analyses of interviews proceeded according to the “systematic approach for identifying, analysing and reporting patterns – themes – across a dataset” — thematic analysis (217). As a process which requires considerable review, reflection, discussion and frequent return to transcriptions, this process involved all supervisors and several meetings, both in person and via email over an extended period in order to develop the themes from the data.

Relationship of studies to each other

In order to answer the research question, data from all studies as well as the information obtained from the literature reviews needed to be synthesised and considered as a whole. The systematic review literature and the results of the questionnaires and interviews gave insight into how and why reluctance to deliver negative messages manifests. The review of submitted assessments also contributed to this aspect, though could not definitely contribute answers, rather offer suggestions. The MUM effect literature review provided a lens by which to consider the interview data in answering why reluctance manifests, but in addition, care was taken to ensure that themes other than those relating to the MUM effect were able to be identified. This included the extension of some; for example, those around discomfort and the inclusion of others, such as the underlying views on failure. As such, and in line with a mixed methods approach and design, all data collected contributed to the discussion, implications, consideration of future directions and the overall conclusions of the study.

Summary

This chapter has summarised the relevant considerations and the background to the choice of methods employed to answer the research question of how and why reluctance to deliver negative assessment messages manifests in dermatology training in Australia. In keeping with the pragmatic theoretical framework, a mixture of methods was selected and the findings from each study will be presented separately in the results chapters which follow this chapter. The overall findings and the individual studies contributed to answer the research question will be presented in Chapter 10, Discussion.

CHAPTER 6

To provide an honest rating of a poorly performing trainee is difficult:

Assessor self-reports of discomfort with clinical performance assessments and how this impacts upon assessment behaviours

Introduction

This chapter seeks to explore assessors' experiences with clinical performance assessments, with a particular focus on discomfort and self-report of MUM behaviours. Using a questionnaire format, assessors were asked to self-report their levels of discomfort with various aspects of the assessment process and their assessment practices. The relationships between discomfort and assessment practices were then explored in the study context. The chapter concludes with an introduction to the next chapter, which seeks to explore trainees' views on their assessors' behaviours around the delivery of assessment messages and their perspectives of the assessments.

Background

As discussed, clinicians are now frequently called upon to judge trainee performance in the context of the clinical performance assessments, which form a key part of the program of assessments in many medical training programs, including the ACD. However, many factors including rater training (52), cognitive and other biases (9) and the assessment tool design itself (227) are known to potentially affect the assessment result which is recorded and delivered. As presented in Chapter 2, failure to fail and grade inflation are well-documented issues in medical education, though, their exact prevalence is not always precisely defined. For example, Bingham and Crampton comment that their documented rates of unsatisfactory performance were likely to be an underestimate, with “real rates of underperformance [being] much higher” (22). Other studies have reported figures — for example, national surveys in the USA of clerkship directors in internal medicine report that 18% of directors admitted that failure to fail of medical students had occurred in their department in the 2004 report (68), and the figure was 38% in the 2009 report (67). A recent study in nursing in Australia reported that almost one quarter of participants had “sometimes or often given the benefit of the doubt” to a student (228). The actual figures may in fact be greater. The recent systematic review on failure to fail in medical education discussed in Chapter 2 highlighted barriers to the delivery of a negative assessment message, including assessor concerns about the impact that failing may have on a trainee (80). Though referring to a different (PBL) setting with medical students, Qadan and colleagues concur arguing that concerns about the impact of the assessment result, both for themselves and the student, combined with assessor awareness of the culture of

feedback are reasons why assessors “are not providing honest feedback” (105). They assert this could result in a difference between the assessor’s private thoughts on the trainee’s performance and the public judgement they deliver (105), and others have also recognised this gap (9,229). Exploring the impact that such concerns have on assessor behaviours in other settings is therefore important in order to work to reduce their prevalence.

Difficulties with the delivery of performance information are not unique to educational settings, and other areas which face similar problems may offer relevant insights. For example, the business environment also faces difficulties with delivering performance appraisal information, especially in the context of negative results, and have undertaken studies in this area (157). Performance appraisals share some common features with the clinical performance assessments, given both require an assessor to form an opinion on another’s performance and then deliver this judgement. As such, work from the performance appraisal literature may offer new insights and approaches to the problems of delivering negative assessment messages in the medical education context (112). To illustrate, Bernardin and Beatty (221) recognised the importance of rater discomfort on the appraisal process and its outcomes. They published the Performance Appraisal Discomfort Scale (PADS) in 1984 to measure rater discomfort in their context and subsequently, Cox and colleagues explored the impact of discomfort on rater behaviour. Using an experimental study, they looked specifically at the relationship between rater self-reported level of discomfort and subsequent reluctance to deliver a negative assessment message in business settings — the MUM effect — finding that the more discomfort participants felt (which included justifying ratings), “the more likely she or he engaged in avoidance when giving feedback” (157). While the setting for their study was in a different context with a different assessment format, these studies have potential relevance to the medical education setting, due to the common situation of needing to share negative information with another. Discomfort with the process is previously unexplored in medical education and may impact upon the delivery of negative assessment messages.

As argued, failure to fail may represent a manifestation of the MUM effect when it arises from assessor reluctance to deliver a negative assessment message.

Understanding the effect helps to account for the observation that the issue of failure to

fail persists despite efforts to address it (80) and also allows us to appreciate other, more subtle behaviours which may occur due to reluctance, including delay in (113) or avoidance of the delivery of messages (230) or distortion of the content (132).

The aim of this study is therefore to explore assessors' experiences of discomfort in clinical performance assessments, such as WBA and ITA, and whether this has an impact on their behaviour. To date, there are no formal studies of the MUM effect concept published in medical education literature, however, its widespread presence and documentation in many other settings predicts it is likely to occur here also. This study will therefore extend previous work by measuring assessor discomfort and exploring the relationship between discomfort and assessment behaviours in a medical specialist training setting in Australia. If discomfort does affect assessor behaviour, this is important to determine as this may lead to new opportunities by which to reduce reluctance to deliver negative assessment messages. When MUMing denies trainees accurate assessment messages, the potential to harm their professional development occurs. And ultimately, patient care may be compromised.

Research questions

The research questions for this study are:

What are Australian dermatology assessors' experiences of discomfort in relation to their assessor role?

How does the MUM effect manifest in clinical performance assessment messages in Australian dermatology training?

What is the relationship between self-reported level of discomfort and self-report of MUM behaviours in clinical performance assessment for assessors of Australian dermatology trainees?

Methods

A questionnaire was developed and distributed in order to collect data on Australian dermatology assessors' reported experiences of giving clinical performance assessment messages to trainees in the study setting (the questionnaire is presented in full in Appendix C). All Fellows of the ACD who self-reported participation in the clinical

performance assessment of trainees in the preceding 12 months prior to distribution of the survey were invited to participate in the survey.

Development of the questionnaire

The questionnaire comprised three sections:

1. “Information about you”, which included demographic questions
2. “Experiences with performance appraisal”, which included questions about discomfort in the assessment process
3. “Making and delivering your assessment”, which included questions about MUM and other behaviours.

Sections 2 and 3 included questions from previously published tools, and general questions relating to assessors’ previous experiences with giving assessment results to trainees. The questions in section 2 were based on the Performance Appraisal Discomfort Scale (PADS) by Bernardin and Beatty (221). The PADS was used to measure assessor discomfort with various aspects of the assessment process. This 20-item scale (table 6.1) was designed to measure discomfort felt by appraisers in performance appraisals in business settings, but given their similarity to clinical performance assessments, it was deemed appropriate for use here also. The original scale was modified to suit the local use. For example, the terms “employee” and “subordinate” from the original scale were replaced with “trainee”; the term “interview” in the original version was replaced with “feedback session”; and “removal from the training program” was used instead of the term “discharge” from employment. As with the original scale, a five-point Likert scale was used, with “1” being labelled as “No discomfort” and “5” as “High discomfort”. Originally, results were suggested to be used descriptively to assess self-efficacy in tasks; but others, for example Cox and colleagues (157) have also presented mean scores for select items. In this study, means and ranges are presented for each item in order to explore the breadth of assessor experiences — to explore which tasks are most problematic and how discomfort levels vary between individuals. Results are compared across assessor demographics. While this departs from the previous published structure, this was performed to look for differences between groups which may inform interventions.

The questions in section 3 were based on a self-report scale of MUM behaviours by Marler, McKee, Cox, Simmering and Allen (132). This group of authors has used other versions of MUM scales to explore rater behaviour in business performance appraisal settings in relation to variables such as trust, job satisfaction and feedback medium preference (119,157). The scale development process included expert content validation and authors presented mean scores for each subscale. In order to obtain results from recent practice, the question logic function in SurveyMonkey was employed to limit the display of items in this scale only to those assessors who reported that they had delivered a negative assessment message to a trainee within the preceding 12 months of distribution of the questionnaire. The MUM scale consisted of three subscales, each focusing on different behaviours (table 6.2). These are the Avoidance subscale (10 items), the Distort (sugarcoating) subscale (6 items) and an Other subscale (3 items). The first two subscales were similarly modified from the original scale for local circumstances. For example, “I asked others to give any negative information” was replaced with “I asked another consultant to give the trainee any negative information”. The Other subscale included questions about other ways in which assessors may behave when they have negative information to give, derived from the authors’ own questions and other sources (231), and were not part of the original scale. Marler et al.’s original scale also included questions about the organisation’s norms in relation to giving negative messages. As this was not the focus of the current work, these questions were not included in this study. As with the original work (132), mean scores for each subscale are presented. To permit deeper understanding of results, additional analyses — the number of items participants agreed with and those agreeing with at least one item — are also presented.

In addition to the two scales, respondents were asked directly to self-report certain behaviours. Firstly, they were asked whether they had engaged in failure to fail and/or grade inflation in the specific assessment formats used in the ACD. Those who had were asked to nominate their reasons for doing so (with options presented being: due to the impact that delivering an unsatisfactory result would have on themselves; the impact it would have had on the trainee; disagreement with the outcome that would result due to failure; or “other”) and any other behaviours accompanying this were asked for. Secondly, respondents were asked if they had received a direct approach from a trainee, requesting them to give a certain rating or change a previously assigned rating on a

clinical performance assessment. A final free-text section asked assessors for any comments they would like to make about performance assessments.

The questionnaire was developed in conjunction with feedback from the project supervisors and was piloted prior to its distribution to a group of 17 non-dermatologist doctors, who were involved in clinical performance assessment of medical specialist trainees in other colleges, including general practice, physician, surgical, paediatric, anaesthetics, obstetrics and gynaecology, ophthalmology, intensive care, and radiology trainees. Minor modifications to the survey were made based on review of their comments (see Appendix C for the full questionnaire).

Distribution

An invitation to participate in the questionnaire was distributed by the Honorary Secretary, ACD between 12 August and 16 September 2016 to Fellows via the email address registered with the College. This was distributed to all Fellows and invited those with recent experience (being within the previous 12 months) in conducting clinical performance assessments to participate in the study. An electronic platform (SurveyMonkey) was used to allow ease of distribution, completion and collection of data. Results were collected anonymously.

Data analysis

The quantitative survey responses were imported from SurveyMonkey into IBM SPSS Statistics Version 25 for statistical analysis. The quantitative data were analysed descriptively to show respondent demographics and the range and mean of variables: being their responses to the PADS scale, MUM scale and other questions. Comparisons of these responses between groups (by seniority as clinician and assessor, gender and SoT status) were made using independent samples t-tests and inferences for proportions (percentages). Free-text responses underwent thematic analysis and are presented in the themes to which they relate.

Results

Demographics

Responses were received from 61 of 135 eligible assessors giving a 45.2% response rate. This number of eligible assessors is an estimate of the number of assessors from the ACD as complete figures for the number of clinical supervisors are not available for all states. The numbers estimated are as follows: NSW 67, Queensland 9, South Australia 15, Victoria 25, Western Australia 19.

Respondents practise in all states and territories with the exception of Tasmania and the Northern Territory. The majority were from NSW (36%, $n = 20$), followed by Victoria (20%, $n = 11$) and Queensland and Western Australia equal (15%, $n = 8$ each). Less than 10% were from South Australia (9%, $n = 5$) and the ACT (5%, $n = 3$). Most respondents were male (58%, $n = 32$) and the majority were not a supervisor of training (SoT) (73%, $n = 40$). Most respondents reported that they complete relatively small numbers of assessments each year, with the majority completing less than five of each (Appendix H, table H.1).

Forty-four percent ($n = 24$) of respondents report having practised for between 16 and 25 years and 31% ($n = 17$) of respondents report having been involved in conducting performance assessments on trainees for less than five years. For the purposes of analysis, years practising as a dermatologist and years assessing have been condensed to two groups and described as junior (less than five years) and senior (more than five years). These groupings were made as some reports in the literature showed conflicting results about the effect of seniority as an assessor on assessment ability (93,163) and the relationship of discomfort to this was of interest.

Performance Appraisal Discomfort Scale (PADS) results

Respondents reported a wide range of levels of discomfort with situations involving giving assessment messages to trainees. Table 6.1 shows that for all items, the range of discomfort reported by assessors varied from “1” to “4” or “5” (where “1” = no discomfort and “5” = high discomfort). The mean PADS score across all items for all respondents was 2.45 ($n = 49$).

Situations reported to be associated with the lowest levels of discomfort were: conducting a feedback session with a trainee who is performing well (mean rating = 1.21, n = 44 (83%) respondents reported no discomfort); letting a trainee talk during a feedback session (mean rating = 1.28, n = 38 (76%) respondents reported no discomfort); and letting a trainee give his or her point of view regarding a problem with performance (mean rating = 1.52, n = 34 (65%) respondents reported no discomfort).

The highest levels of discomfort were reported with: situations of having a trainee accuse an assessor of playing favourites in relation to their assessment activities (mean rating = 3.37, n = 14 (29%) reported high discomfort); supporting a recommendation to remove a trainee from the program (mean rating = 3.54, n = 17 (34%) reported high discomfort); and warning a trainee that the assessor would support their removal from the training program unless performance improved (mean rating = 3.71, n = 14 (29%) reported high discomfort).

Table 6.1: Mean assessor self-reported level of discomfort and range of ratings by individual feedback giving situation across all respondents.
Reported in ascending order of discomfort (1 = no discomfort, through to 5 = high discomfort), Performance Appraisal Discomfort Scale (PADS) (221).

Description of situation	Mean rating	Range
Conducting a feedback session with a trainee who is performing well	1.21	1–4
Letting a trainee talk during a feedback session	1.28	1–4
Letting a trainee give his or her point of view regarding a problem with performance	1.52	1–4
Giving a satisfactory rating to a trainee who has done a satisfactory (but not exceptional) job	1.52	1–4
Encouraging a trainee to evaluate his or her own performance	1.57	1–4
Telling a trainee that he or she must stop taking long breaks	2.10	1–4
Asking a trainee if he or she has any comments about your ratings of his or her performance	2.13	1–4
Telling a trainee that he or she must stop coming to work late	2.15	1–4
Talking to a trainee about his or her performance on the job	2.29	1–5
Telling a trainee that his or her performance can be improved	2.38	1–5
Telling a male trainee that his performance needs to improve	2.51	1–5
Telling a female trainee that her performance must improve	2.63	1–5
A trainee's challenging you to justify your evaluation in the middle of a feedback session	2.84	1–5
Telling a trainee that his or her work is only satisfactory, when you know that he or she expects an above satisfactory rating	2.85	1–5
Telling a trainee who has problems in dealing with others that he or she should do something about it	2.94	1–5
Responding to a trainee who is upset over your rating of his or her performance	3.08	1–5
Conducting a feedback session with a poorly performing trainee	3.30	1–5
A trainee's accusing you of playing favourites in the rating of trainees	3.37	1–5
Supporting a recommendation that a trainee be removed from the training program	3.54	1–5
Warning a poorly performing trainee that unless performance improves, you will support a recommendation for his or her removal from the training program	3.71	1–5

Independent samples t-tests showed that the mean PADS score did not show statistically significant variation with number of years reported practising as a dermatologist (collapsed to either junior or senior), nor number of years as assessor (collapsed to either junior or senior) though there was a slight tendency for junior clinicians and assessors to have higher PADS scores. Similarly, mean PADS scores did not show statistically significant difference with assessor gender (female mean PADS = 2.41, male mean PADS = 2.48, $p = 0.74$), nor with SoT status (SoT mean PADS = 2.29, non-SoT mean PADS = 2.52, $p = 0.27$), though there was a tendency for SoTs to have lower PADS scores (figure 6.1 and Appendix H, table H.2). Considering the questions regarding discomfort levels when telling a male or female trainee that their performance needs to improve, again, there were no statistically significant differences in the mean level of discomfort reported by assessors based on their gender (Appendix H, table H.3).

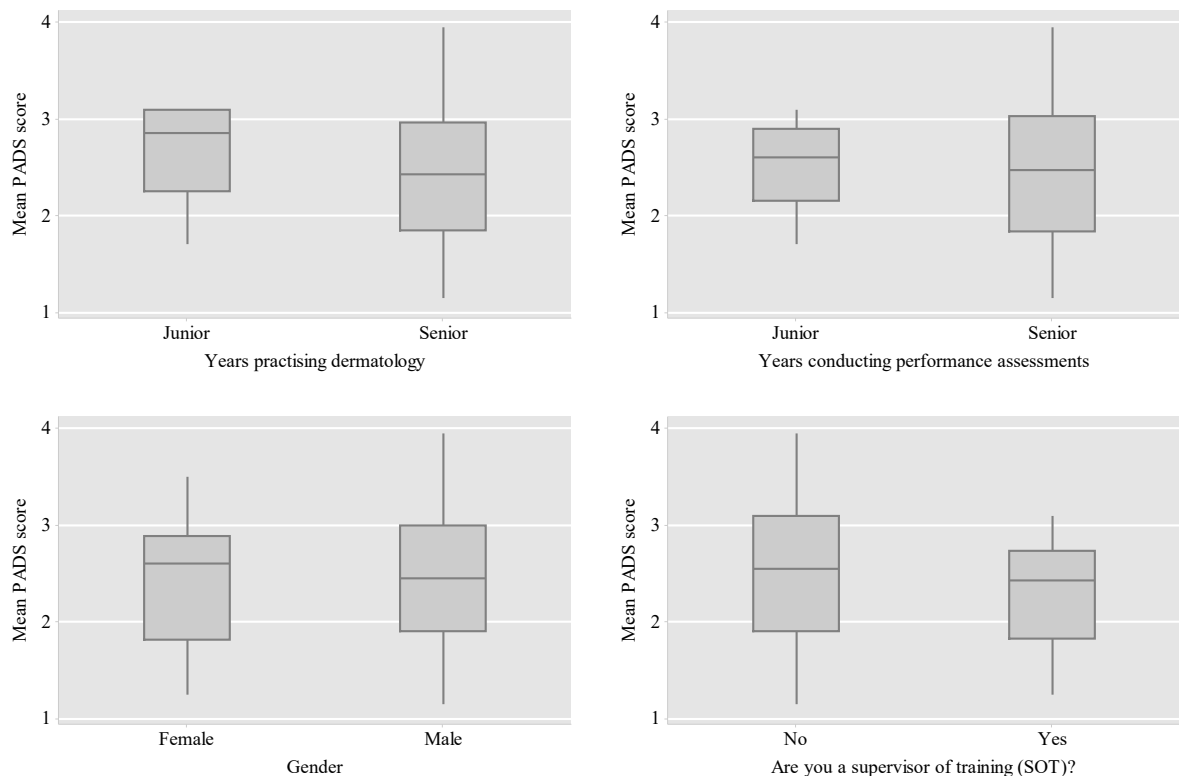


Figure 6.1: Mean PADS scores across all statements by demographics

Considering the three statements with the highest individual PADS mean ratings overall, SoT status was significant for one statement only (SoTs reported significantly lower PADS for the statement “Warning a poorly performing trainee that unless performance improves, you will support a recommendation for his or her removal from the training program”, mean PADS SoT = 3.14, mean PADS non-SoT = 3.94, $p = 0.027$). There were tendencies for the more senior clinicians and assessors to report lower PADS scores (seen in two of the three questions), but these were not significant (Appendix H, tables H.4–H.6).

MUM behaviours and self-report of failure to fail and grade inflation

MUM behaviours subscale

Forty-six percent of respondents ($n = 23$) reported having had to share some negative information in relation to a WBA or a SITA (such as a low rating or negative feedback) with a trainee within the preceding 12 months, and only this group were then asked to self-report MUM behaviours. The mean PADS scores for those having reported sharing negative information was not statistically significantly different from those who had not, (2.28 vs 2.60 $p = 0.095$) (Appendix H, table H.2). Similarly, the respondents SoT status was not statistically significantly related either (proportion SoTs having shared negative information = 57.1%, proportion of non-SoTs having shared negative information = 41.7%, difference in percentage = 15.5%, 95% CI – 13.9%, 41.4%, $p = 0.251$), meaning SoTs reported they were not statistically significantly more likely to have shared negative assessment information than non-SoTs.

Considering the subscales, the mean MUM score (Likert scale 1–5) for the Avoid subscale was 1.78 ($n = 23$). The mean score for the Distort (sugarcoat) subscale was 2.61 ($n = 21$) and the score was 1.92 ($n = 21$) for the Other subscale. At an individual item level, respondents in this subgroup reported they did not engage in most MUM behaviours, with the exception of one example of distortion (sugarcoating), where all either agreed or were neutral that they “tried to make the trainee feel better by emphasising positive things”. For all other behaviours, respondents self-reported low levels of behaviours, with agreement with sugarcoating behaviours (range 4.5–81.8%) being reported more often than avoidance ones (range 0–21.7%). There were 4 items in

the Avoid subscale and 1 in the Other subscale which no respondent agreed with. The number and percentage of respondents agreeing to each item in the subscales is shown in table 6.2.

Table 6.2: Assessor self-report of MUM behaviours in relation to giving negative assessment messages to trainees

Possible behaviours in relation to giving negative assessment messages to trainees	MUM subscale	Agree		Disagree or neutral		Total number of respondents
		%	n	%	n	n
I tried to make the trainee feel better by emphasizing positive things	Distort (sugarcoat)	81.8	18	18.2	4	22
I put the negative feedback between two positive comments	Distort (sugarcoat)	27.3	6	72.7	16	22
I did not point out all of the weaknesses	Avoid	21.7	5	78.3	18	23
I tried to make it sound nicer than it was	Distort (sugarcoat)	19.0	4	81.0	17	21
I adopted a very formal or clinical manner when I gave the feedback so the trainee would be less likely to question me	Other	19.0	4	81.0	17	21
I told them only part of the problem	Avoid	17.4	4	82.6	19	23
I left out some of the negative information	Avoid	17.4	4	82.6	19	23
I tried to make the feedback more positive than it really was	Distort (sugarcoat)	13.6	3	86.4	19	22
I changed the feedback or rating so it was not so negative	Distort (sugarcoat)	9.1	2	90.9	20	22
I changed the feedback during the feedback session due to the trainee's response to it	Other	4.8	1	95.2	20	21
I changed the feedback or rating to make it more positive	Distort (sugarcoat)	4.5	1	95.5	21	22
I didn't go into much detail about the negatives	Avoid	4.3	1	95.7	22	23
I asked another consultant to give the trainee any negative information	Avoid	4.3	1	95.7	22	23
I left giving the feedback for as long as I could	Avoid	4.3	1	95.7	22	23
I avoided the trainee at work, so I would not have to talk to him/her	Avoid	0	0	100	23	23
I avoided it altogether	Avoid	0	0	100	23	23
I steered clear of circumstances in which I'll have to give negative information	Avoid	0	0	100	23	23
I found ways to get out of sharing negative information	Avoid	0	0	100	23	23
I gave the feedback as quickly as I could and before the trainee could react to it	Other	0	0	100	21	21

Figure 6.2 and table 6.3 show the number of items respondents agreed with in each subscale agreed with by percent. For the MUM Avoid and Other subscales, this was most frequently zero items and for the Distort (sugarcoat) subscale, this was most frequently one item (“I tried to make the trainee feel better by emphasising positive things”).

Total number of MUM items agreed with

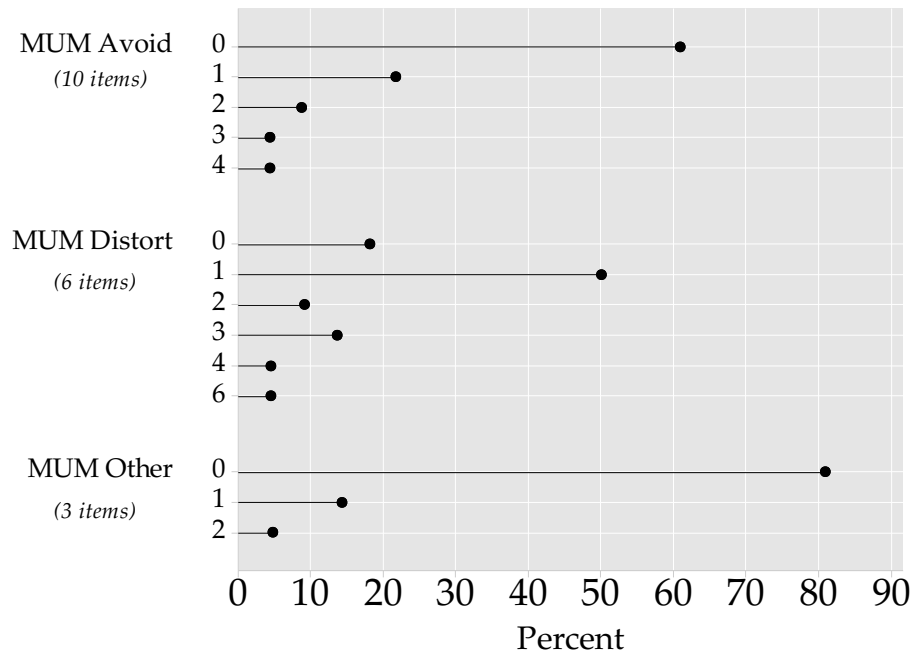


Figure 6.2: Number of items in each subscale respondents agreed with by percent

Table 6.3: Number of items in each subscale respondents agreed with by number and percent

Number of items in subscale respondents agreed with	MUM Avoid (10 items)		MUM Distort (6 items)		MUM Other (3 items)	
	n	%	n	%	n	%
0	14	60.9	4	18.2	17	81.0
1	5	21.7	11	50.0	3	14.3
2	2	8.7	2	9.1	1	4.8
3	1	4.3	3	13.6	n/a	n/a
4	1	4.3	1	4.5	n/a	n/a
5	0	0	0	0	n/a	n/a
6	0	0	1	4.5	n/a	n/a
Total	23		22		21	

Figure 6.3 shows the proportion of respondents who agreed with at least one item in each subscale by percentage. A total of 39.1% ($n = 9$) respondents agreed to at least one item in the MUM Avoid subscale, 81.8% ($n = 18$) agreed to at least one item in the MUM Distort (sugarcoat) subscale, and 19.0% ($n = 4$) agreed to at least one item in the MUM Other subscale.

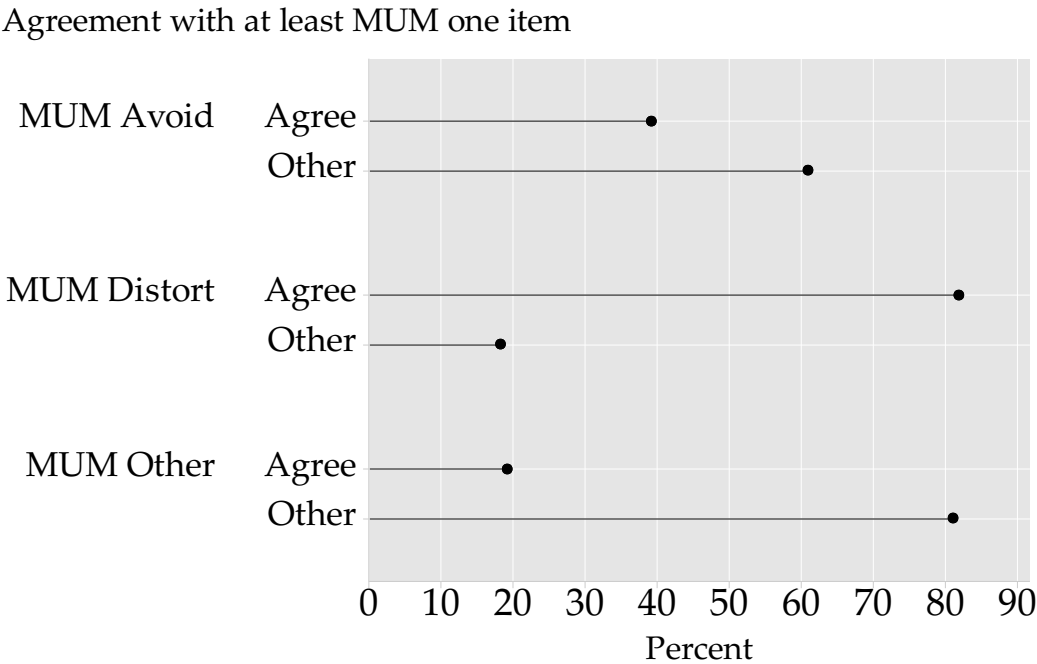


Figure 6.3: Percent of assessors who agreed with at least one item in each MUM subscale

Respondents were asked to describe ways that they manage their discomfort around delivering negative assessment messages to trainees. Several free-text responses were recorded, including offering support and encouragement, asking for the trainee's opinion on the feedback, and giving the message as soon as possible and without delay. Some described providing suggestions for development to the trainee, having a second supervisor present, and offering a follow-up discussion a week later. Two assessors emphasised the need to be clear in discussions, and to state to the trainee that the assessment and performance criteria need to be addressed.

Self-reports of MUM behaviours were compared with years as a dermatologist, experience as an assessor, gender and SoT status. Junior consultants generally reported non-significant tendencies to engage in more MUM avoidance and distortion behaviours than senior consultants (Appendix H, table H.7); and more junior compared with senior consultants agreed having engaged in at least one or more MUM behaviours, but this is not significant (Appendix H, table H.8). There were no significant differences between MUM behaviours for junior and senior assessors, either in the mean number of MUM behaviours they reported engaging in (Appendix H, table H.9) or the percentage who agreed with engaging in at least one behaviour (Appendix H, table H.10). Male assessors tended to engage in more MUM behaviours than female assessors, but the results were not significant (Appendix H, tables H11–H12).

Those who were not SoTs tended to engage in more MUM behaviours than SoTs, but the results were only significant for MUM distortion behaviours, where SoTs reported in engaging in a mean of .75 of these behaviours, while non-SoTs reported in engaging in a mean of 2 (95% CI $-2.214, -.324$, $p = 0.021$) (Appendix H, tables 13–14).

Self-report of failure to fail and grade inflation

All respondents were asked whether they had marked a trainee as competent on a WBA or SITA when they did not feel that they were (i.e. failed to fail), or had given a higher rating on a SITA than they felt was justified (i.e. grade inflated) in the preceding 12 months. Those who reported they had were asked to give their main reason for doing this and if they did anything else in addition to doing so.

Two assessors (of total $n = 48$, 4%) indicated that they have marked a trainee as competent (i.e. failed to fail) in a WBA when they did not think they were (both were on the Pro-DA format), three were unsure and 40 (of total $n = 48$, 83%) had not (figures 6.4, 6.5). By way of explanation, those who had failed to fail a trainee did so equally due to the impact this would have had on themselves and the impact it would have had on the trainee (table 6.4). While one assessor indicated they gave accurate verbal feedback to the trainee, the other did not (table 6.5). Of the respondents who indicated that they were not sure if they had failed to fail on a WBA, one subsequently reported that they had also grade inflated on a SITA and one reported that they had also failed to fail on a SITA. The third reported they were not sure if they had grade inflated on a SITA.

Four assessors (total $n = 48$, 8%) indicated that they had marked a trainee as satisfactory (that is, not given a score of 2 or less on two or more areas) on a SITA when they did not think they were (i.e. failed to fail), while two (total $n = 48$, 4%) were not sure and 40 (total $n = 48$, 83%) had not (figures 6.4, 6.5). The reasons given for avoiding to fail the trainee included the impact that doing so would have had on the trainee (2 assessors), the impact that it would have had on themselves (1 assessor), and a combination of factors including the timing of the SITA in relation to upcoming exams and the fact that in the respondent's view, the trainee had not been given appropriate negative ratings on previous assessments, the implication being it would be difficult to do so now (table 6.4). Two assessors indicated that they did nothing in addition to assigning the rating, while one gave the trainee accurate written feedback and one spoke to the trainee's SoT (table 6.5). The respondents who admitted to failure to fail in the SITA were different to those who admitted to failure to fail in the WBA. Therefore, this means that in this questionnaire, a total of 6 out of 48 assessors (12.5%) answering this question admitted to having failed to fail a trainee in a clinical performance

assessment in the preceding 12 months (figures 6.4, 6.5). In addition, one respondent who said they had failed to fail on a SITA responded that they were not sure if they had failed to fail a trainee on a WBA, and two responded that they had not completed a WBA in the time period in question.

Eight assessors (total n = 45, 18%) indicated that they have given higher ratings than they felt were justified in the SITA (i.e. grade inflated) in the previous 12 months, while 2 (total n = 45, 4%) were not sure and 35 (total n = 45, 78%) had not. Of these eight respondents, one admitted to failing to fail a trainee on a WBA, one was not sure if they had failed to fail on a WBA, two admitted to failure to fail of trainees on a SITA, and two were not sure if they had failed to fail trainees on a SITA. Two of the eight respondents reported that they had not failed to fail on either a WBA or a SITA.

In summary, three respondents who admitted to grade inflation on the SITA also admitted to failure to fail on either a WBA or SITA, and three others were not sure if they had failed to fail on these assessments (figures 6.4, 6.5). Assessors indicated that the reasons for this included the impact doing so would have had on themselves (4 assessors), the impact it would have had on the trainee (3 assessors), due to disagreement with the process outcomes (1 assessor), and their opinion that the process itself was inadequate (2 assessors) (table 6.4). In most cases, the assessors indicated that despite grade inflating, they gave accurate verbal feedback to the trainee instead (5 assessors). Three assessors did nothing in addition to assigning the higher rating, one gave accurate written feedback and one questioned the trainee about their commitment to dermatology (table 6.5).

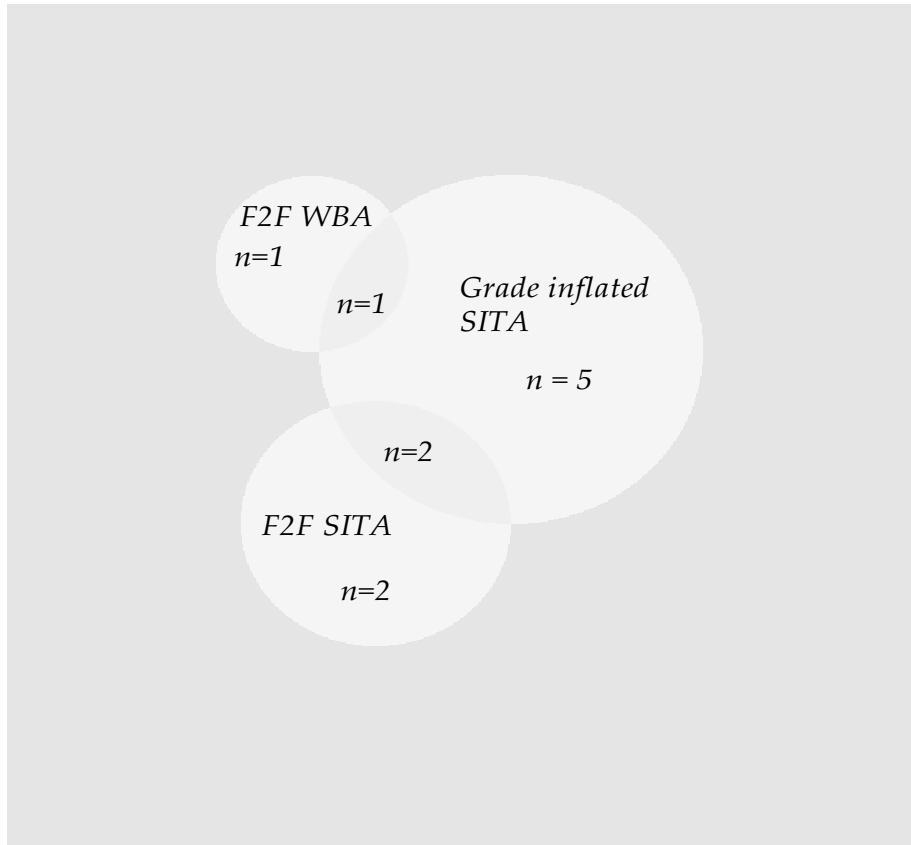


Figure 6.4: Venn diagram showing the relationship between those respondents who self-reported failing to fail or grading inflating a trainee in a clinical performance assessment

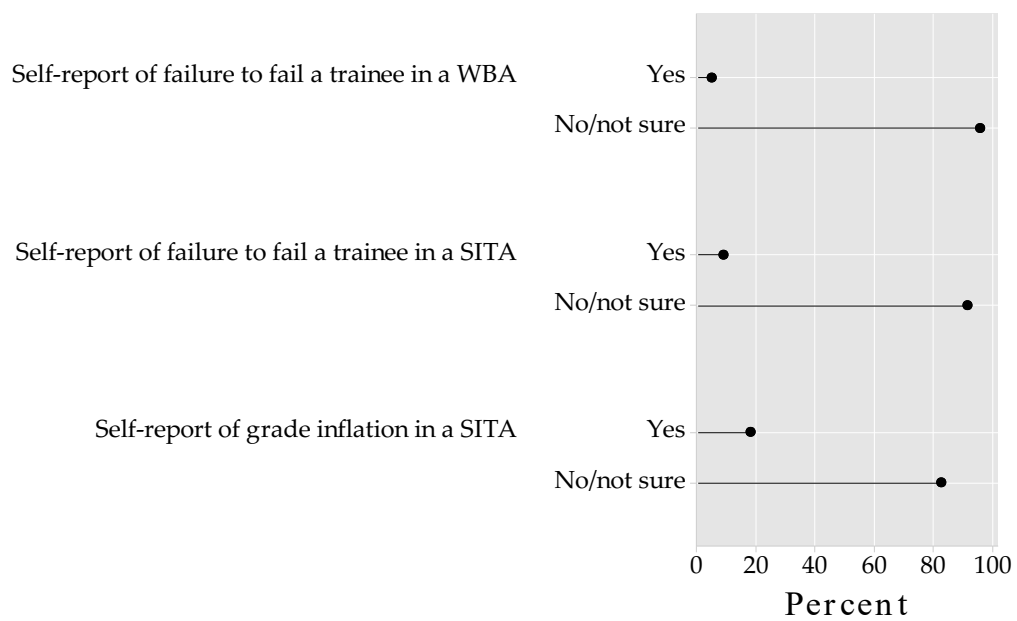


Figure 6.5: Percentage of respondents to self-report failure to fail or grade inflation of a trainee's result in a clinical performance assessment

The main reasons respondents listed in explanation for their decision to grade inflate or fail to fail a trainee are shown in table 6.4.

Table 6.4: Reasons given by assessors for not recording their true assessment judgement on the assessment form

Reason	Assessors who failed to fail on WBA or SITA	Assessors who grade inflated SITA results
Due to the impact that giving my real rating would have had on me — e.g. I'd feel guilty, the trainee may think negatively of me, I'd end up in a negative mood, I wanted to avoid any conflict etc.	2	4
Due to the impact that giving my real rating would have had on the trainee — e.g. the result would have upset the trainee, the result would have affected the trainee's confidence, the trainees prefer to hear good results etc.	3	3
I disagreed with the outcome that would have results — e.g. trainee progress being interrupted, trainee being put on performance improvement plan, getting a black mark against his or her name etc.	0	1
Other	1 ¹	2 ^{2, 3}

¹ – previous assessors had failed to give negative ratings to the trainee despite verbally reporting this and now too close to exams: “leading up to exams the SITA would have been adverse to this doctors whose previous assessors had failed to give negative ratings even when they personally said to me that the trainee was underperforming”

² – inadequate tool

³ – related to exposure and lack of correlation between poor clinical skills but good technical skills
Note answers add to greater than respondent number as one respondent stated they did both the first two reasons in the text box “Other”.

Additional actions that assessors undertook when they decided to grade inflate or fail to fail a trainee are shown in table 6.5.

Table 6.5: Additional actions by assessors who did not record their assessment judgement on the assessment form

Other action	Assessors who failed to fail on WBA or SITA	Assessors who grade inflated SITA results
Nil other	3	3
Gave verbal feedback to the trainee	1	5
Gave written feedback to the trainee	1	1
Other	1 ⁴	1 ⁵

⁴ – spoke to their supervisor of training

⁵ – questioned the trainee about their desire to practice dermatology

Note answers add to greater than respondent number as one respondent gave two reasons (verbal and written feedback) in “Other”.

Self-report of failure to fail on WBA or a SITA and grade inflation on a SITA were compared with seniority as a dermatologist, seniority as an assessor, gender and SoT status. No significant differences were found between self-report of either failure to fail or grade inflation and years practising as a dermatologist (Appendix H, table H.15). While there was no significant difference in rates of failure to fail or grade inflation compared with years as an assessor, those who had been assessing for more than five years showed slight tendencies to engage in both behaviours compared with those who have been assessing for less than five years (Appendix H, table H.16).

Male respondents were non-significantly more likely to admit to failure to fail than female assessors (Appendix H, table H.17). Although the difference in proportions of those failing to fail was not significantly different based on SoT status, it is interesting to note that no respondent who admitted to failure to fail was an SoT (Appendix H, table H.18).

Direct approaches from trainees in relation to assessor ratings

Small numbers of respondents agreed with the statements relating to being approached by a trainee either pre-empting a result or asking them to change a previously assigned one (figure 6.6, table 6.6).

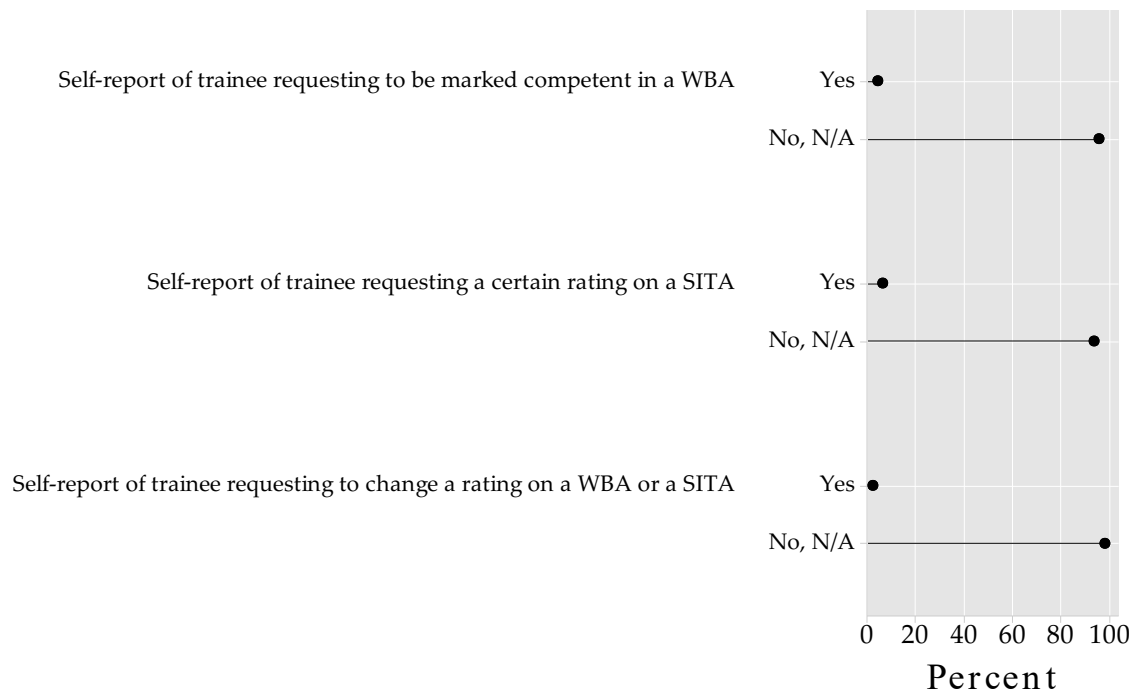


Figure 6.6: Percentage of respondents who self-reported being approached by a trainee in relation to a clinical performance assessment result

Table 6.6: Number and percentage of respondents who self-reported being approached by a trainee in relation to a clinical performance assessment result

Description of approach by trainee (total n = 48)	n	%
Self-report of trainee requesting to be marked competent in WBA	2	4.2
Self-report of trainee requesting a certain rating on a SITA	3	6.3
Self-report of a trainee requesting a change to a rating on a WBA or a SITA	1	2.1

As the numbers were small, the total number of assessors who reported being approached in any of the above scenarios were considered together and compared with seniority as a dermatologist, seniority as an assessor, gender and SoT status. Note that there were five assessors involved in direct requests, two requests were in relation to a

WBA result, three were in relation to a SITA rating and one was to change a previously assigned rating. One assessor was approached for both a WBA and a SITA rating (table 6.6).

Male assessors were statistically significantly more likely to report having been approached by a trainee with a direct request about their assessment, and in fact no females reported being approached. Similarly, no SoTs self-reported being approached, though the difference to non-SoTs was not significant. Other differences were not significant, though the more junior consultants and the more senior assessors report being approached more often. None of the respondents who reported being approached directly by a trainee in relation to the assessment results admitted to failure to fail also (all stated that they had either not, or were not sure if they had, failed to fail a trainee), while one quarter of those who admitted to being approached also admitted to grade inflation (table 6.7). This means that most of those who did grade inflate do not report being directly approached by trainees in relation to their assessment results. In summary, this suggests that the assessor's decision to fail to fail or grade inflate is not related to any direct approach or pressure from trainees.

Table 6.7: Percentage of respondents self-reporting being approached by a trainee in relation to a clinical performance assessment result by demographic and other features

Percentage agreeing to being directly approached by a trainee in relation to a clinical performance assessment separated by demographics		Difference in percentages (first column – second)		
Junior dermatologist	Senior dermatologist	Estimate	CI	P-value
14.3% (n = 7)	9.8% (n = 41)	4.5%	–.128, .420	.562
Junior assessor	Senior assessor			
6.3% (n = 16)	12.5% (n = 32)	–6.3%	–.226, .171	.454
Female	Male			
0% (n = 21)	18.5% (n = 27)	–18.5%	–.367, .001	.047
SoT	Not an SoT			
0% (n = 14)	14.7% (n = 34)	–14.7%	–.301, .084	.163
Had to share negative information	Not had to share negative information			
14.3% (n = 21)	7.4% (n = 27)	6.9%	–.116, .279	.379
Failed to fail	Not failed to fail/not sure			
0% (n = 6)	12.2% (n = 41)	–12.2%	–.255, .274	.489
Grade inflated	Not grade inflated			
25% (n = 8)	8.1% (n = 37)	16.9	–.053, .514	.211

The PADS scores were not significantly different for those assessors who reported being approached directly about their ratings by trainees (Appendix H, table H.19), nor was there a difference in the number of MUM behaviours self-reported between the two groups (Appendix H, table H.20). However, there was a significant association seen between those who self-reported experiencing direct requests in relation to clinical performance assessments and self-reported at least one MUM behaviour, where all (100%) of those who reported being approached in relation to an assessment result, also admitted to engaging in at least one MUM avoidance behaviour, compared with only 27.8% of those who had not been approached (95% CI .115, .875, $p = 0.042$) (Appendix H, table H.21).

Comparison of Performance Appraisal Discomfort Scale (PADS) scores and self-report of MUM behaviours (including failure to fail and grade inflation)

Demographics

As previously presented, PADS scores were compared with assessor demographics to explore for differences between groups. There were no significant differences between PADS score and assessor gender, years practising as a dermatologist, years as an assessor or SoT status; however there were non-significant tendencies to lower PADS scores with more experience as a clinician and assessor and with being an SoT (Appendix H, table H.2).

PADS scores and MUM items

In order to explore whether discomfort with assessment was associated with MUM behaviours, mean PADS scores were compared with responses to the MUM subscales.

Means for each subscale are compared between those who self-reported a history of failure to fail and those who did not. The mean MUM distort for those with a history of failure to fail was significantly greater than those who did not, though it is important to appreciate the small numbers of respondents (table 6.8).

Table 6.8: Comparison of mean MUM scores between those with and without self-report of failure to fail

	Self-report of failure to fail	No self-report of failure to fail	Estimate	95% CI	P-value
MUM avoid	2.00 (n = 1)	1.76 (n = 20)	.24	-.87, 1.35	.656
MUM distort	4.00 (n = 1)	2.55 (n = 19)	1.45	.18, 2.71	.027
MUM other	3.33 (n = 1)	1.85 (n = 20)	1.48	.38, 2.58	.011
MUM total mean	2.84 (n = 1)	2.03 (n = 19)	.81	-.20, 1.81	.108

There were no significant differences found between the mean PADS scores of those who self-reported at least one MUM behaviour and those who did not (table 6.9).

Table 6.9: Respondent mean PADS score compared with agreement with at least one MUM statement by subscale items

MUM	Mean PADS score		Mean difference (Agree – Disagree/neutral)		
	Agree with at least one MUM statement	Disagree/neutral with MUM statements	Estimate	95% CI	P-value
At least one MUM avoid statement	2.219 (n = 9)	2.322 (n = 14)	.103	–.489, .630	.700
At least one MUM Distort	2.235 (n = 18)	2.275 (n = 4)	.040	–.510, .679	.904
At least one MUM Other	2.725 (n = 4)	2.178 (n = 17)	–.547	–1.156, .338	.222

The association between the mean PADS score and the total number of MUM items agreed with was explored. This showed a non-statistically significant association, with a correlation of 0.309 accounting for approximately only 10% of the variation (table 6.10).

Table 6.10: Pearson correlation of MUM behaviours self-reported and PADS mean.

		Total MUM behaviours	PADS mean
Total MUM behaviours	Pearson correlation	1	.309
	Sig. (2-tailed)		.173
	N	21	21

Linear and quadratic regression analyses are presented for the mean PADS score predicted from the total number of MUM items participants self-reported agreement with. Neither yielded statistically significant associations ($p = 0.173$ for linear,

$p = 0.296$ for quadratic) (figures 6.7 and 6.8). Hence the linear association between the PADS score and self-report of MUM behaviours is weak indicating that discomfort is not a large explanation for MUM behaviours (figure 6.7).

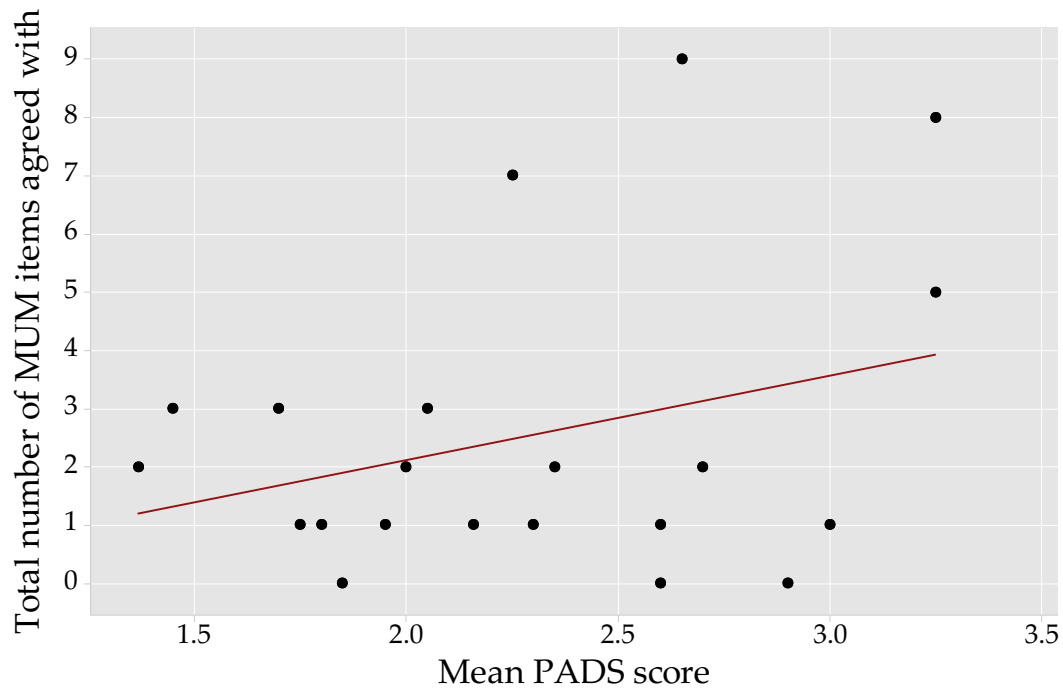


Figure 6.7: Linear relationship between mean PADS score and total number of MUM items agreed with

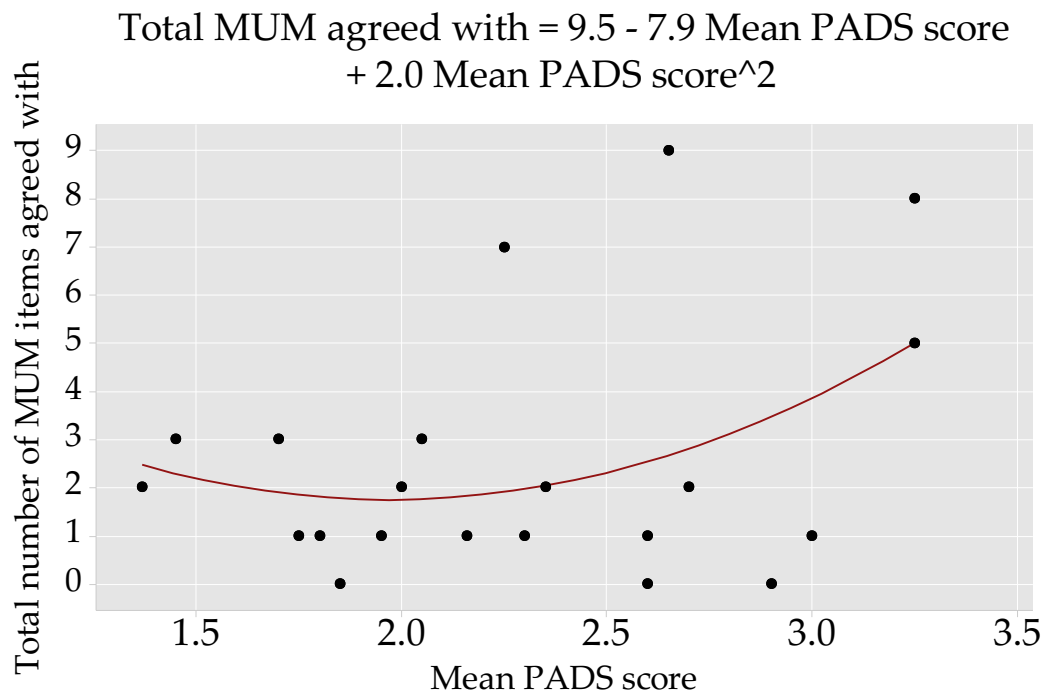


Figure 6.8: Quadratic relationship between mean PADS score and total number of MUM items agreed with

When the median split of the PADS score was considered (median = 2.25) for the 23 participants who answered the questions about MUM behaviours, the number of total MUM behaviours agreed with and the standard deviations are seen to differ between the two groups. Table 6.11 shows that those assessors with PADS scores below the median self-report a mean of 1.70 MUM behaviours whereas those assessors with a PADS score above the median report a mean of 3.27 MUM behaviours.

Table 6.11: Comparison of mean number of MUM behaviours self-reported by respondents for PADS scores above and below the median.

Means PADS score	Mean number of total MUM behaviours	SD of total MUM
< 2.25	1.70	1.06
2.25 or more	3.27	3.35

This indicates that a greater number of MUM behaviours were self-reported and greater variability seen when participants also self-report more discomfort with performance appraisal (PADS score), and conversely less MUM behaviours and less variation when discomfort is lower.

PADS scores and failure to fail and grade inflation self-reports

Mean PADS scores were compared with assessor self-reports of grade inflation and failure to fail in the preceding 12 months. The PADS scores from those who reported they had given a trainee a higher score than the one that they felt was justified in a SITA (PADS = 2.82) did not differ from those who did not self-report grade inflation (PADS = 2.44, $p = 0.224$) (figure 6.9, table 6.12).

However, a significant difference was found when the PADS scores for those assessors who reported that they had failed to fail a trainee on a WBA or a SITA within the preceding 12 months (PADS = 3.38) were compared with those who did not (PADS = 2.35, $p = 0.004$, 95% CI 0.5 – 1.5) (figure 6.10, table 6.12).

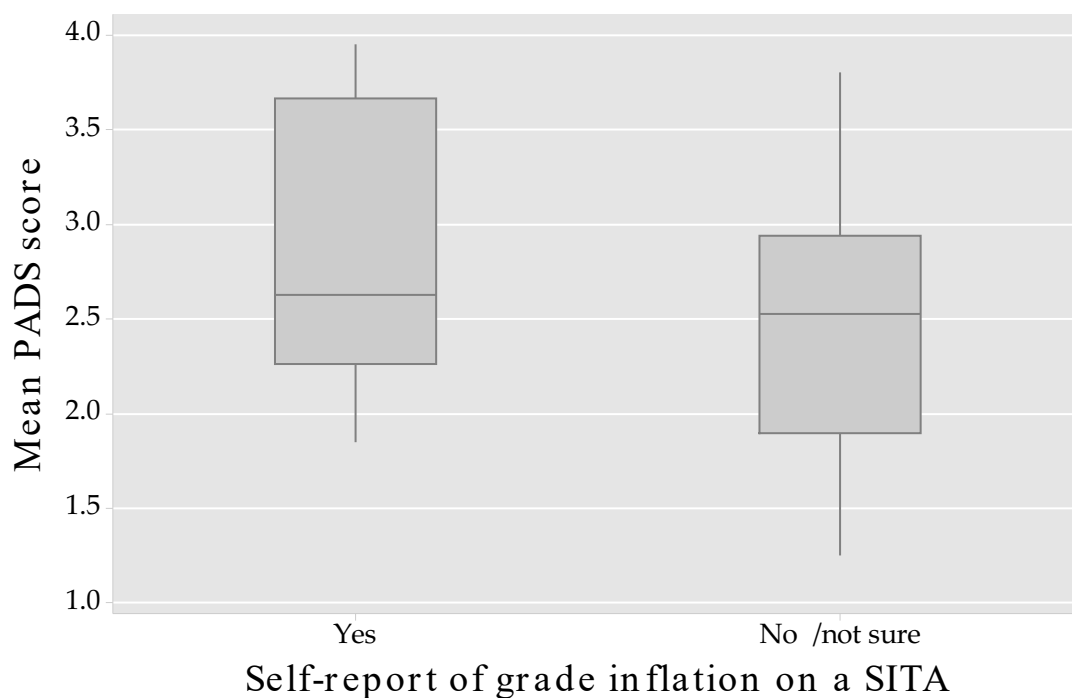


Figure 6.9: Comparison of PADS scores for respondents to self-report grade inflation on a SITA

Table 6.12: Comparison between PADS score and self-report of failure to fail or grade inflation (t-tests)

	Mean PADS score		Mean difference (Yes – no/not sure)		
	Yes	No/not sure	Estimate	95% CI	P-value
Grade inflation	2.82 (n = 8)	2.44 (n = 36)	.377	–.166, .995	.224
Failure to fail	3.38 (n = 6)	2.35 (n = 40)	1.03	.512, 1.495	.004

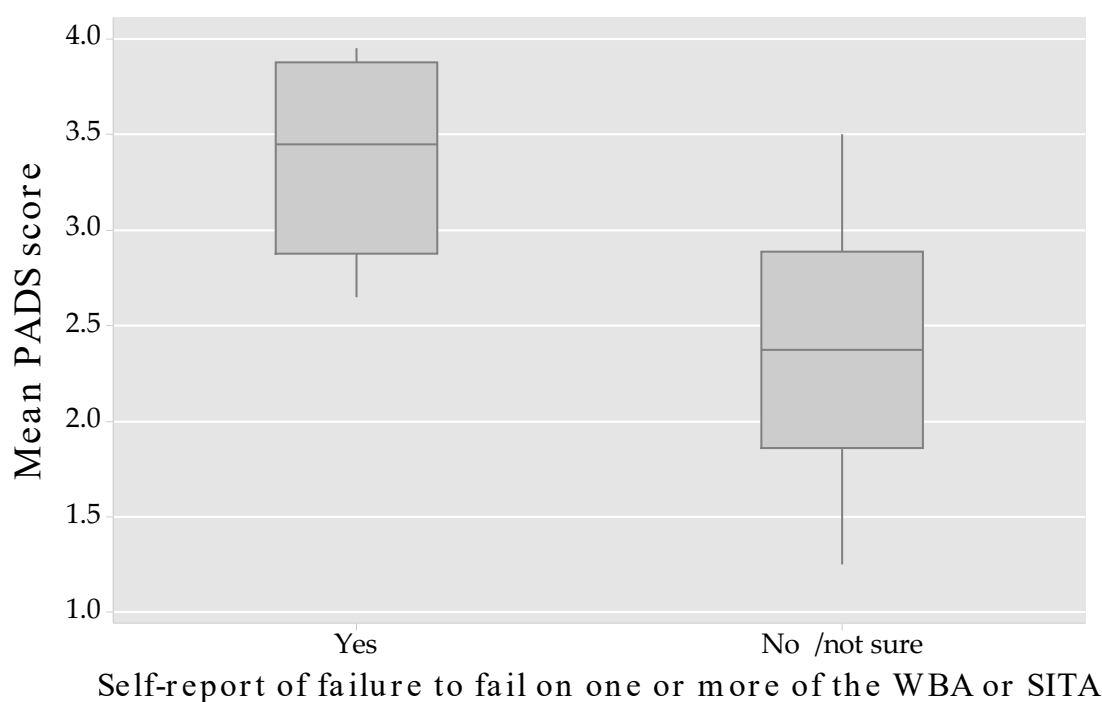


Figure 6.10: Comparison of PADS scores for respondents to self-report failure to fail on a WBA or SITA

Free-text comments

Respondents recorded the following free-text comments on the clinical performance assessments at the end of the questionnaire. These were separated into the following themes (table 6.13):

Table 6.13: Free-text responses to the statement “Do you have any comments you would like to make about performance assessments?”

Theme	Illustrative quote
System problems	“The system is fraught with limitations. It is difficult to give an unsatisfactory assessment when the trainee is borderline” “we cannot justify the fact that our current system allows trainees known to be ‘non-competent’, relying on the examination assessment” “I think that the consequences of negative assessments ... are too great for the trainees. A chat and review should be enough” “To provide an honest rating of a poorly performing trainee is difficult”
Problems with use of assessments	“I was constantly frustrated by SITAs that had little constructive or factual feedback on the trainee’s performance, leaving it up to me to document negative marks” “I think it is very difficult for the people teaching the trainees to also be the assessors. We all want to be kind but we all have a responsibility to the community” “It is difficult to find the time to do these and negative performance assessments do require more time and privacy.” “it is tricky not to shatter [junior trainees] with a terrible report”
Opinions on the assessments	“necessary but stupid” “most are a waste of time” “They are an important part of the training program” “It is a poor tool as most Fellows do not stay long enough with the trainee to fully assess their skills”
Ideal situation	“there should be a culture of regular feedback ... so the trainee sees this as an opportunity for improvement rather than a markdown” “The most important thing we can do for both the trainees and the ACD is to give honest and accurate assessments” “multiple formal assessments by many different supervisors over the course of a trainees’ training is important” “More consultant group collaboration”
Suggested approaches	“keep it private with the trainee only ... have an example ... I find the question ‘is this what you were expecting but did not really want to hear ... useful” “The process needs to be confidential so that assessors will give accurate results”
Regular feedback is ideal	“best to give feedback along the way”
Anonymity is important	“It is much more difficult to give negative feedback in person than it is in a de-identified setting” “anonymity is critically important for honest assessments”
Needs	“Some guidance/training in performance assessments” “assur[ance] of robustness of process, personal indemnification and support from College” “better training for us, and for the registrars about what to expect would be good”
Misunderstandings of process	“I think performance assessments have become more accurate since the SoT chooses the assessors rather than in the past when trainees selected assessors who may give them a more favourable report”
Selection of trainees	“I think we should be tougher in our selection process ... If we select the right trainees at the outset ... then we are less likely to find ourselves in the situation of having to performance manage weak and manipulative trainees”

Discussion

Summary and interpretation of results

To my knowledge, this is the first work to study the level of discomfort experienced by assessors in relation to their assessment duties and to look for the presence of the MUM effect in clinical performance assessments in a medical specialist training setting. The results show that many assessors do experience discomfort in relation to their assessment duties, some to high levels, though there were wide inter-individual variations seen. Further, this study shows that some assessors self-report engaging in MUM behaviours, including failure to fail and grade inflation in this setting, and there is indication that discomfort is one factor that contributes to this. As presented, the performance appraisal discomfort scores for those who self-reported engaging in failure to fail were significantly higher than those who did not.

As seen in the Results, most respondents were male and not SoTs, and most were both experienced clinicians and assessors, with at least five years' experience in each. The number of years practising as a dermatologist and years assessing trainees did not have a statistically significant effect upon responses overall, though some difference in tendencies were noted and both groups generally showed similar patterns. Furthermore, there were no statistically significant differences in relation to self-report of grade inflation or failure to fail based on seniority, as either a clinician or assessor, though senior assessors tended to admit to failing to fail and grade inflation more; and juniors did not report higher levels of direct approach from trainees in regard to assessment rating compared with seniors. In this study, assessor gender did not have a statistically significant impact on PADS score or self-report of MUM behaviours, though males did report a tendency to greater engagement with MUM behaviours, including failure to fail, compared with females.

The assessor SoT status was one demographic feature which did reveal some differences in the questionnaire. Overall, while SoTs showed non-significant tendencies to lower PADS scores than non-SoTs, a significant difference was seen in one of the highest discomfort-inducing situations. While there were no significant differences seen in the proportions who reported having to share negative information with a trainee between SoTs and non-SoTs, SoTs did report engagement in fewer MUM

behaviours, which was statistically significant for distortion. Further, no SoTs self-reported failure to fail or being approached by a trainee in relation to their assessment results. It was of note also that some assessors reported being unsure whether they had engaged in failure to fail or grade inflation or not. This uncertainty with aspects of the assessment process is further discussed in Chapter 9, in interviews with assessors.

Discussion of results

This study reports several important findings in relation to assessor demographics, experiences of discomfort, and the behaviours of both assessors and trainees in the assessment process.

Put together, the findings contribute to those in the literature where some differences in assessment behaviours with assessor demographics are reported. For example, in regards to seniority, Ryan, Mandel, Sama and Ward showed that new or junior assessors gave trainees significantly higher evaluation results than more senior ones did (93); and Phillips, Lim, Madhavan and Macafee concur, reporting that “Registrar (junior) feedback tends to always be positive, but consultants will tell you how it is” (202). Perhaps in explanation, in a study of clinicians’ views of assessment tools in an undergraduate veterinary medicine setting Bok, Jaarsma, Spruijt, van Beukelen, van der Vleuten and Teunissen report that “More experienced teachers reported few difficulties in providing negative feedback” (94). These examples suggest that junior clinicians may have a greater tendency to distort or avoid negative assessment information compared with seniors and whilst the results in this study tended to support the literature, they did not reach statistical significance. However, it is important to note that contrasting views are also found both in the literature and in this study. For example, Smith, Harrington and Houghton reported in a study outside of medical education, that the level of discomfort with performance appraisal was not seen to be modified by the teller’s age and level of managerial experience (156) and this may impact on behaviour. The tendency for a higher reporting of failure to fail and grade inflation by senior assessors in this study is consistent with Smith’s work. It may be that for some, level of experience does not make the task of delivering negative assessment information any easier. The conflicting findings suggest that other factors are also important to the outcome and further study on the impact of seniority on assessment outcomes is needed.

Few other studies have investigated the impact of gender on tendency to MUM, and experimental situations involving news delivery have shown conflicting findings (124,130,132,154) (113,130). In this study, only male assessors admitted to being approached by trainees in relation to their assessment ratings ($p = 0.047$). Further work in this area would therefore be valuable.

In this study, many assessors report experiencing discomfort in relation to clinical performance assessments. As presented, this showed a great inter-individual variation, with some assessors reporting no discomfort for situations that others reported as causing high levels. As shown, the reported levels of discomfort were not significantly associated with assessor demographics, including SoT status; and whether this is in fact a true difference and if so, due to training or self-selection to become an SoT in the first instance would be important to explore further. The fact that some assessors are experiencing discomfort related to their assessment role (some rating this at a high level) is important to address and others have recognised this in the appraisal setting (232). The direct identification of those who are affected would be important, given that demographics did not suggest any group was statistically significantly more vulnerable than another. Having recognised the discomfort experienced by some assessors, the obligation to address this surely exists, particularly as in this context, assessors perform their roles voluntarily and in addition to their other commitments. Options to reduce discomfort should be explored. For example, while proposed to help feedback givers increase the accuracy of their face-to-face feedback delivery “role plays or behaviour modeling” (157) may benefit assessors also.

This study has also documented assessor engagement in MUM behaviours in this setting. It is notable the results showed that less than half of assessors reported having to have shared negative information in relation to an assessment with a trainee in the preceding 12 months. This seems a surprisingly low figure and seems unusual that over half of assessors had no negative comments at all to make about a trainee’s performance in the period of a year. Perhaps in itself, this finding supports the claim that “there is sometimes a culture within medicine where the critical discussions with [trainees] are often avoided” (188). The MUM effect may explain this low rate as some assessors choose to avoid sharing negative information at all. While it is possible that some respondents may have interpreted the question to relate to sharing information about

failure of an assessment, rather than more broadly about giving negative feedback on aspects of performance, the latter aspect was emphasised in the questionnaire. Of those with recent experience in sharing negative assessment information, almost 40% of assessors acknowledged at least one avoidance MUM behaviour and over 80% at least one distortion behaviour. Respondents also tended to greater agreement with MUM statements describing more moderate examples of avoidance (for example “I didn’t go into much detail about the negatives”) than those describing complete avoidance (for example “I avoided it altogether”), and distortion (sugarcoating) statements were agreed with more than avoidance ones. Again agreement was higher for those statements that described more moderate examples of distortion or the practice of softening or sweetening of the news (for example “I tried to make the trainee feel better by emphasising positive things”), where there was almost universal agreement as opposed to distorting or changing the information (for example, “I changed the feedback or rating so it was not so negative”). As SoTs showed less distortion, whether this is an innate trait of those who put themselves forward to be an SoT or due to training would be important to research further.

The results from this questionnaire support and extend the literature by documenting the presence of grade inflation and failure to fail in the clinical performance assessments in a medical specialist training setting in Australia. Six of 48 respondents admitted to failure to fail of trainees in the 12 months prior to distribution of the questionnaire, and 8 of 45 respondents admitted to grade inflation in the same period. Assessors attributed their reasons for this approximately equally between concern for the trainee and concern for themselves. Of further concern, many assessors stated that they kept completely mum and took no further action (6/14), though some gave the trainee accurate verbal feedback (6/14) and fewer gave written feedback (2/14). Both of these latter behaviours are in effect “off the record” as neither contribute to the consequences or outcome of the assessment. As noted by Cohen and colleagues, if underperformance is not labelled officially, remediation will not necessarily occur (28) meaning that such actions may be ultimately of limited benefit to the trainee. Other actions undertaken by assessors in the current study included delegation (speaking to trainee’s SoT), or deflection (questioning trainee’s commitment to dermatology). Though numbers were small, these results suggest that assessors may be more likely to keep mum to the trainee about failing to fail them and not take any further action, but were more likely to give their feedback

verbally to those to whom they have allocated a higher result than they felt was justified on the SITA.

A further major finding of this study is the demonstration of a relationship between discomfort and assessment behaviours. When the number of MUM behaviours assessors admitted to was compared with their PADS scores, a small positive relationship was seen. While engaging in MUM behaviours can only be partly explained by the assessor's level of discomfort with the process, those who admitted to failure to fail did have significantly higher PADS scores compared with those who did not admit to failure to fail. This was not seen in relation to self-report of grade inflation, suggesting that this group may be more comfortable with giving feedback to trainees — but they tend to do this verbally rather than in writing. Where assessors reported they had given written comments reflecting their true assessment judgement but an inflated rating (as seen in table 6.5, one assessor in each group gave written feedback to the trainee, presumably in the comments section of the form), a mismatch between the valence of their ratings and comments would have been seen. However, where assessors gave no (3 in each group) or only verbal comments (1 in the failure to fail group and 5 in the grade inflation group), alongside an inflated rating, that is, they kept MUM totally, a difference between the ratings given and the comments written would not be seen. A comparison of ratings and comments on a sample of real-world assessments will be made in Chapter 8 to consider this further.

Finally, this study also documents the assessors' perspective of being approached by a trainee in relation to their assessment results, either to request a particular result in advance or to modify one already assigned. Five assessors reported being approached a total of six times in such a way. While none of these self-reported failure to fail, three did report grade inflating a trainee, and a significant relationship between those who reported being directly approached and engagement in MUM avoidance behaviours and tendency to engage in sugarcoating behaviours was seen, whereby 100% of those reporting direct approach admitted engaging in avoidance behaviours, compared with 27.8% who denied direct approach ($p = 0.042$). Whether this is a causal relationship or just an association would be important to explore further.

Free-text comments offered by some assessors gave opportunity to express their views further and some expanded on the issues they face, including problems at system and

tool levels, and their opinions, suggestions and needs. Many of these are further discussed in Chapter 9, Interviews with assessors.

The problem with not delivering the assessment message

This study shows that the assessment result recorded and communicated in the clinical performance assessment setting does not necessarily always reflect the assessor's view of the trainee's performance, and that the level of discomfort experienced by assessors may be associated with failure to fail and additional MUM behaviours, though regression analyses show that discomfort is not the only factor contributing to such behaviours.

As discussed, failure to fail and grade inflation have been previously documented in medical education settings (21,67,68) and may have potentially severe consequences. Such issues with inflated results have been identified in medical student settings, both within the United States (233–237) and other countries (10,61) and also in medical speciality training settings (11,238). Studies on the topic are less frequent in the Australian setting, though grade inflation and failure to fail are likely identified here also (22,90). This study contributes to the literature by documenting the issue and additional associations in a medical specialist training setting in Australia. It documents the extent of MUM behaviours including failure to fail, grade inflation and the impact and effect of discomfort on assessors. This has not been demonstrated in this context previously and is important to understand further, not only due to its association with failure to fail, but due to the potential impacts on the wellbeing of the assessors who experience it. Further, this study documents the practice of trainee approaches to assessors in relation to their assessment outcomes, and suggests a possible association with assessor avoidance behaviours as a result. While many interventions specifically targeted at the quality of the assessment instrument or rater training have been tried to date to address such issues; as shown by this work, failure to fail and grade inflation occur for reasons beyond lack of assessor training and issues with assessment forms.

Future directions

This study shows that some assessors experience high levels of discomfort in relation to their assessment duties and that MUM effects, including failure to fail and grade inflation occur in the medical specialist training setting.

The impact that assessor discomfort has on both assessment outcomes and assessor wellbeing are areas for future research. These include, for example the magnitude of the effect of discomfort on outcomes, and further study of the impact of discomfort on the assessor and their family. The PADS scale, modified for local use, may be of assistance in identifying clinicians who are vulnerable to high levels of discomfort in assessment as the demographics measured in this study did not reveal strong associations. This could be used as a self-assessment tool or at an institutional level and it would be important to determine whether an individual's PADS score is modifiable with training or over time. If not, consideration of whether those with high PADS should be precluded from assessment duties, due to both the potential effect on themselves and their higher likelihood of failing to fail trainees, could be made. To my knowledge, no studies have explored methods to modify an individual's PADS score, though other works have demonstrated associations between an individual's psychological traits (core self-evaluation) and performance appraisal discomfort. While Gbadamosi and Ross did not find a significant relationship between perceived stress and reported performance appraisal discomfort, they did find that workers who reported higher core self-evaluations also experienced higher levels of performance appraisal discomfort, leading them to suggest the need for additional support during the appraisal process to these individuals (232). While the association between PADS score and overall MUM behaviours was small in the current study, there was a clear association between discomfort and failure to fail behaviours. While assessor discomfort certainly does not explain all the instances of failure to fail, it represents a further avenue to pursue in an attempt to ensure trainees are provided with the accurate assessment information they deserve.

Limitations

There are several limitations to this study. Given the small pool of eligible respondents, the number of respondents was always to be a limitation. The response rate to the questionnaire of 45.2%, whilst considered satisfactory for an electronic survey and in keeping with response rates from other surveys (239,240) still meant that many possible responders did not contribute their experiences. Attempts to increase this were made by sending reminders and increasing the time available for the questionnaire. Further, as many respondents denied having shared negative assessment information within the

preceding 12 months, a smaller number were asked about MUM behaviours. Questionnaires are intrinsically subject to various biases, including responder bias, whereby those with more extreme views on a subject may be more likely to respond, which has the potential to affect the results, and surveying the whole assessor population rather than a sample, attempted to minimise this effect. All parts of the questionnaire relied on self-report of behaviours, and no independent measures were included. This introduces potential inaccuracies to the results, including the social desirability bias whereby people are more likely to give socially acceptable responses. The use of an anonymous questionnaire is one method of combatting this, but this bias means that the number of issues raised in this questionnaire may be in fact an under-representation of the extent of the problems. Finally, the quantitative data presented here can give information as to the scope of issues, but not explore the reasons behind behaviours. The collection of qualitative data such as with interviews are required to explore this further.

Conclusion

This study shows that assessors experience varying levels of discomfort, ranging from none to high, when conducting clinical performance assessments on trainees. For some, this clearly impacts on the assessment message they deliver to the trainee and may potentially distort it. This study argues that assessor discomfort should be addressed both for the sake of the assessor experiencing it and the trainee, whose assessment results may potentially be affected by it.

The next chapter seeks to explore ACD trainees' views of the clinical performance assessment formats they experience in their training context and in particular, their perceptions of how their assessors manage the delivery of negative assessment messages to them.

CHAPTER 7

*I have hardly ever received
constructive criticism:*

Trainee impressions of assessor behaviours
and the clinical performance assessments

Introduction

This chapter focuses on the trainees' views in regards to the research question of how and why reluctance to deliver negative assessment messages manifests.

Firstly, it considers how reluctance to deliver negative assessment messages manifests by seeking trainees' perspectives on MUM behaviours by their assessors. The systematic review in Chapter 4 showed that some trainees report a lack of honesty in the assessment process. In this chapter, the trainee experience of MUMing by their assessors is sought, which will then permit comparison with the assessors' views.

Further, this study considers why reluctance may manifest by extending the systematic review findings, by seeking ACD trainees' views of the assessments. The systematic review showed that trainees are divided in their views of their messages and in this chapter, ACD trainees' views of their assessments will be compared to their views of MUM behaviours to explore for associations between the two. The chapter concludes by introducing a review and comparison of the written assessment messages conveyed through one of the assessment formats in use at the ACD.

Background

As outlined in Chapter 2, the clinical performance assessments aim to observe trainee performance in the workplace and can have summative and/or formative roles. But as already discussed, these formats face numerous challenges, including the logistical ones of finding the time and an appropriate patient and assessor in order to conduct them within the busy clinical setting. Further, it is suggested that the assessment message delivered is not always the same as the assessor's judgement of the trainee's performance (9,229). As such, it is important to understand the trainees' perspectives on these assessments and in particular, whether they believe the assessment messages they receive from them are valuable for their learning and outweigh the challenges.

The MUM effect literature presented in Chapter 3 showed that people are reluctant to deliver a message which they feel will be perceived as negative by the recipient (113). The systematic review in Chapter 4 highlighted that some trainees have doubts as to the level of honesty in the message delivered to them due to recognition that assessors completed them retrospectively at times (6,186,187) or without having observed the

trainee in all the components of the task (188,189). Some trainees also suspect their assessors are reluctant to deliver negative assessment information (38). While Ginsburg and colleagues have shown that trainees can decode the narrative comments written by assessors about other trainees and determine the rank-order of a group of assessments (38), this does contrast with work outside medicine, where students don't always agree that their feedback comments were consistently understandable (241). One possible explanation for Ginsburg's observation is perhaps that trainees recognise and correct for the MUM behaviour of distortion in the assessors' comments, but it is yet to be established if they can do this for the narrative comments on their own performance, or if they have awareness of other MUM behaviours, including delay. Chapter 3 argued that the MUM effect has potential relevance for the medical education setting and may help us to understand why some assessors are reluctant to deliver their assessment message to trainees, while Chapter 6 showed that assessors self-report low levels of MUM behaviours.

A difference between assessors' and trainees' views may be anticipated, as contrasting views about various aspects of assessments have been previously reported. For example, trainees' assessment of their level of performance often varies from their assessors' views (242) [and differences were also seen when practising doctors were the assessee (243)]; and differences in perceptions around the logistics and practicalities of arranging the workplace assessments (244) and in the amount of feedback given are seen (151).

In order to consider trainees' perspectives on why reluctance to deliver negative assessment messages may manifest, their views on the clinical performance assessments are also relevant. The literature reports that an assessor's behaviour and the judgement they deliver, may be influenced by many factors beyond the trainee's level of performance, and so potentially jeopardise the validity of the message delivered. As one example, Bok and colleagues report that "the student's actual or perceived reaction towards negative feedback" can influence an assessor in their delivery of assessment information (94) and this may result in assessors displaying a reluctance to document negative feedback, exposing a gap between their true evaluation and the judgement they deliver (105). The systematic review concluded that medical specialist trainees' views on the assessment messages they receive through clinical performance assessments vary

widely but are strongly linked to their and their assessors' understandings of the purposes of the assessment — that is whether they are viewed as a learning exercise, valuable to promote development or merely as another task to complete. Some trainees described the difficulties experienced in relation to arranging to have assessments completed and feeling pressured to have them completed. Others openly referred to seeking out assessors who they felt would be more likely to mark them as competent. Hence it is important to determine trainees' views of the assessments in the study context both to further the findings of the systematic review and to permit an exploration of the relationship between their views and their experience of MUM behaviours by their assessors.

This study therefore has two main aims. The first is to explore trainees' perspectives on the MUM behaviours of their assessors and compare their views with those of the assessors' self-reports of MUM behaviours presented in Chapter 6. The second aim is to determine trainee views of the clinical performance assessments, including in relation to logistics of arranging these assessments and feelings of pressure to be marked as competent. It will also analyse whether those who report finding feedback more useful also report experiencing less MUM behaviours from their assessors; and, if those who report greater sensitivity to assessment messages report that their assessors engage in more MUM behaviours.

Research questions

The research questions for this study are:

Do Australian dermatology trainees believe their assessors engage in MUM behaviours in the clinical performance assessment context and how do these views compare with the self-reports of their assessors?

How are Australian dermatology trainees' views on the clinical performance assessments — specifically, their views on the usefulness of feedback, their response/sensitivity to feedback, their preference for confidentiality in a feedback setting and their retention of the feedback given — associated with their impression of MUM behaviours by their assessors?

It is important to note that the term “feedback” is used here as it is familiar to trainees, and in this context, generally aligns with the information (or assessment message) they receive about their performance, as compared with the more relational or process-oriented views of feedback currently in the literature (34).

Methods

A questionnaire was developed and distributed to Australian dermatology trainees to explore their views and experiences with clinical performance assessments in the study setting (the questionnaire is presented in full in Appendix D). All trainees of the ACD with recent experience in clinical performance assessments were eligible to participate. At the time of distribution of the questionnaire, there were a total of 114 trainees. This included 14 post-training candidates (who do not complete WBAs or SITAs), 6 on interrupted training time and 1 who had deferred; however these trainees may have still completed a WBA or SITA within the preceding 12 months, and so were eligible to participate in the study.

Development of the questionnaire

The questionnaire comprised four parts:

1. “Information about you”, which included demographic questions
2. “Trainee feedback orientation scale”, which included questions about feedback in the assessment context
3. “Difficulties giving feedback”, which asked for trainee views of MUM behaviours by their assessors
4. “Your experiences with performance assessments”, which asked about trainee experiences with performance assessments, including items relating to ease of arranging assessments, felt pressure around assessments and self-report of approaches to assessor to request or change assessment ratings.

Sections 2 and 3 used previously published tools as follows. Section 2 comprised the Instructional Feedback Orientation Scale (IFOS) developed by King and colleagues (169) (figures 7.13–7.16, Appendix D). This scale was located during the literature review and was selected due to its ability to measure students’ reactions to feedback and so predict those who may respond positively to it (169). The scale development

involved establishing concurrent and discriminant validity, and authors presented mean scores for each subscale. Other authors have used this scale in experimental settings, including simulation studies with university students in the USA. These were designed to explore the effect of various interventions, such as those to reduce face-threats and increase nonverbal immediacy (such as by smiling, maintaining eye contact) on the communication and perceptions of feedback (164,165). As has been discussed above, assessors' behaviours can depend on expected or actual reactions of students (94). The intention was then to measure trainees' views of their assessors' engagement in MUM behaviours and explore if an association with their views of the assessments could be found. The IFOS comprises four subscales which capture trainees' views of various aspects of feedback: Usefulness of feedback, Response to feedback/Sensitivity, Confidentiality of feedback, and Retention of feedback. The Usefulness subscale is referred to as the "Utility" in the original scale, but termed "Usefulness" here to aid clarity. This scale (comprising 10 items) considers respondent views on the benefits of feedback. The Response/Sensitivity subscale (8 items) relates to the interpersonal context and respondents' emotional reactions to negative feedback. The Confidentiality subscale (5 items) asks for respondent preferences in relation to the delivery of feedback, and the Retention subscale (3 items) asks about respondents' ability to recall the feedback given. The items were slightly modified to suit the specific assessment focus of the study: for example, "I will usually reflect on a teacher's feedback" was modified to

"I will usually reflect on an assessors' feedback"; "My feelings are not easily hurt by corrective feedback from a teacher" was modified to "My feelings are not easily hurt by negative feedback from an assessor"; and "I tend to miss out on the details of what instructors want when they provide me with feedback" was modified to "I tend to miss out on the details of what assessors want when they provide me with feedback". Mean scores for each subscale are presented as they were done in the original work (169); and similar to the MUM effect scale, to permit deeper understanding of results, additional analyses — the number of items participants agreed with and those agreeing with at least one item — are also presented.

Section 3 presented the MUM effect scale developed by Marler and colleagues (132) and used in Chapter 6. While the scale was designed to measure evaluator self-report of engagement in MUM behaviours, it was used here to measure trainee views of the

occurrence of behaviours in their assessors, specifically the avoidance of delivery or distortion (or sugarcoating) of negative information. The statements were modified for the study context as in Chapter 6 and also rephrased to ask for the trainees' perspectives of their assessor's behaviour. A lead-in was given: "In your experience, do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:" and was followed with the scale questions. For example, "I asked others to give any negative information" was modified to "Ask others to give any negative information to the trainee?"; and "I changed the message so it was not so negative" was modified to "Change the message they have to give the trainee so it is not so negative?". As in Chapter 6, the original subscale relating to Norms of the group were not included in the current study as this was not the focus of the research question. Results were presented as outlined in the Methods section in Chapter 6.

The final part of the questionnaire was developed by the candidate to explore trainees' experiences with these assessment formats, including the ease of arranging assessments and their behaviours related to the assessments. The same questions about approaching an assessor in order to request or change a rating were asked of trainees with the intention to compare this to assessor responses.

While the questionnaire was presented to trainees in the order described above, the results are presented in this chapter in a different sequence. This was done as it was felt more appropriate to commence the questionnaire with simpler questions about self-report of personal views, before moving to more potentially sensitive questions such as those involving deviations from expected behaviours. However, the MUM responses are presented first to facilitate easier comparison with the previous chapter.

The questionnaire was piloted with a group of different medical specialist trainees prior to distribution. This was done by sending it to a trainee of the Australian and New Zealand College of Anaesthetists, who distributed it to a further 10 trainees. Minor modifications were made based on their feedback (from approximately three trainees). This included adjustment to the terms used, for example replacing "corrective" with "negative" which was felt to be the term in common usage and so more understandable by respondents (see Appendix D for the full questionnaire).

Distribution

An invitation to participate in the questionnaire was delivered to all trainees by the Honorary Secretary, ACD via the email address registered with them. This invited those with recent (i.e. within the previous 12 months) experience in completing clinical performance assessments to participate in the study. The questionnaire was distributed in an electronic format (SurveyMonkey) and was available for completion between September and October 2016. The platform used allowed ease of distribution, completion and collection of data and for responses to be collected anonymously.

Data analysis

The survey responses were imported from SurveyMonkey into IBM SPSS Statistics Version 25 for statistical analysis. The quantitative data were analysed descriptively to show respondent demographics and the range and mean of variables: namely, their responses to the MUM scale, IFOS scale and other questions. Comparisons of these responses between groups (by seniority as trainee and gender) were made using independent samples t-tests and inferences for proportions (percentages). Responses to scales and questions were separately analysed by trainee seniority and gender for several reasons. Firstly, it was of interest to see whether any differences between the groups were observed based on previous findings reported in the literature. This includes the PADS scale, which considers whether discomfort levels vary with the gender of the receiver (table 6.1); and the MUM literature, which suggests possible differences in the delivery of news with the gender of both the teller and receiver (124,130,132,154). Secondly, the systematic review also found that trainee perspective of assessment messages showed some variation with trainee seniority (172), and the MUM literature has explored the seniority of the teller as a modifying factor to news delivery (132). As such, these analyses were of interest in this study.

Trainee responses to the 19 MUM scale items and the 3 questions about approaches to assessors in relation to requesting or changing a rating were compared with assessor responses to the same questions (presented in Chapter 6). Free-text responses underwent thematic analysis and are presented in the themes to which they relate.

Results

The results of the questionnaire are presented as follows. Demographics are presented first, followed by trainee responses to the MUM scale, item by item. These responses are then compared, at an item and subscale level, with the assessor responses presented in Chapter 6. Trainee views of the clinical performance assessments from the IFOS are presented by subscale, followed by trainee reports of experiences with the assessments. These are then compared with the assessor responses. Finally responses to the IFOS items are compared with the MUM responses.

Demographics

A total of 34 out of 114 trainees completed the questionnaire, giving a response rate of 29.8%. Respondents were from all states and territories in Australia, other than the ACT, Northern Territory and Tasmania. The largest number of respondents were from NSW (32.4%, $n = 11$), followed by Victoria (26.5%, $n = 9$), Queensland (20.6%, $n = 7$), South Australia (14.7%, $n = 5$) and Western Australia (5.9%, $n = 2$). The number of trainees per state were as follows: NSW total of 37 trainees, Victoria 33 trainees, Queensland 25 trainees, South Australia 11 trainees, Western Australia 8 trainees. The majority of respondents were female ($n = 21$, 61.7%) and 2 (5.9%) preferred not to identify their gender. Most respondents were in year 2 ($n = 12$, 35.3%), followed by year 4 ($n = 11$, 32.4%) and year 1 ($n = 8$, 23.5%). Two respondents were from year 3 and one was an International Medical Graduate (IMG). For analyses, trainees have been divided into “junior” (years 1 and 2), and “senior” (years 3, 4 and IMG). As IMGs complete either 6, 12 or 24 months of training (Personal communication, Director of Education Services, ACD, email, 5.4.2017), for the purposes of analysis, the IMG respondent has been classified as a senior trainee.

The MUM effect scale

Trainees used the MUM scale to give their views on how their assessors act in the situation of having negative performance feedback to deliver to trainees in the context of a WBA or a SITA. The mean MUM score (Likert scale 1–5) for the Avoid subscale was 2.89 (n = 30). The mean score for the Distort (sugarcoat) subscale was 3.10 (n = 31) and the mean score was 2.58 (n = 30) for the Other subscale. Trainees' report of their assessors' engagement in MUM behaviours is presented in table 7.1. Trainee responses by subscale are presented following this to facilitate their comparison to the IFOS subscale responses.

Table 7.1: Trainee report of their assessors' engagement in MUM behaviours in relation to needing to give negative assessment messages to them

Possible behaviours in relation to giving negative assessment messages to trainees	MUM subscale	Agree		Disagree or neutral		Total number of respondents
		%	n	%	n	n
Do not point out all the weaknesses?	Avoid	58.1	18	41.9	13	31
Leave out some of the negative information?	Avoid	54.8	17	45.2	14	31
Don't go into much detail about the negatives?	Avoid	54.8	17	45.2	14	31
Try to make the trainee feel better by emphasising positive things?	Distort (sugarcoat)	51.6	16	48.4	15	31
Tell trainees only part of the problem?	Avoid	48.4	15	51.6	16	31
Try to make it sound nicer than it is?	Distort (sugarcoat)	48.4	15	51.6	16	31
Put the negative news between two positives?	Distort (sugarcoat)	41.9	13	58.1	18	31
Change the message they have to give a trainee so it is not so negative?	Distort (sugarcoat)	41.9	13	58.1	18	31
Change the message they have to give a trainee to make it more positive?	Distort (sugarcoat)	41.9	13	58.1	18	31
Try to make the information more positive than it really is?	Distort (sugarcoat)	38.7	12	61.3	19	31
Adopt a very formal manner that doesn't encourage dialogue?	Other	26.7	8	73.3	22	30
Put it at the bottom of their "to do" list?	Avoid	25.8	8	74.2	23	31
Ask others to give any negative information to the trainee?	Avoid	22.6	7	77.4	24	31
Steer clear of circumstances in which they'll have to give a trainee negative information?	Avoid	20.0	6	80.0	24	30
Avoid it altogether?	Avoid	16.1	5	83.9	26	31
Find ways to get out of sharing negative information with a trainee?	Avoid	16.1	5	83.9	26	31
"Unload" the feedback quickly before the trainee has a chance to say anything?	Other	13.3	4	86.7	26	30
Avoid the trainee at work so they would not have to talk to him/her?	Avoid	12.9	4	87.1	27	31
Change the feedback they are giving during a feedback session?	Other	10.0	3	90.0	27	30

MUM avoidance subscale

Analysis of items suggested that the Avoid subscale separates into two parts: those behaviours which described partial avoidance and behaviours that described complete avoidance.

Trainees were approximately equally divided in their opinions of whether their assessors engage in partial avoidant behaviours — such as telling only part of the problem, leaving out detail or information, or not pointing out all the weaknesses — with between 15 (48%) and 18 respondents (58%) agreeing with these statements. Overall, most respondents did not agree that assessors engaged in complete avoidance by avoiding the trainee ($n = 27$, 87%) or avoiding delivering the news altogether ($n = 26$, 84%). Less than one quarter ($n = 7$, 23%) felt that assessors would ask another to deliver the information for them, steer clear of circumstances in which they would have to deliver the information ($n = 6$, 20%) or find ways to get out of delivering the news ($n = 5$, 16%). Percentages agreeing are shown in figure 7.1 (responses shown in Appendix I, table I.1).

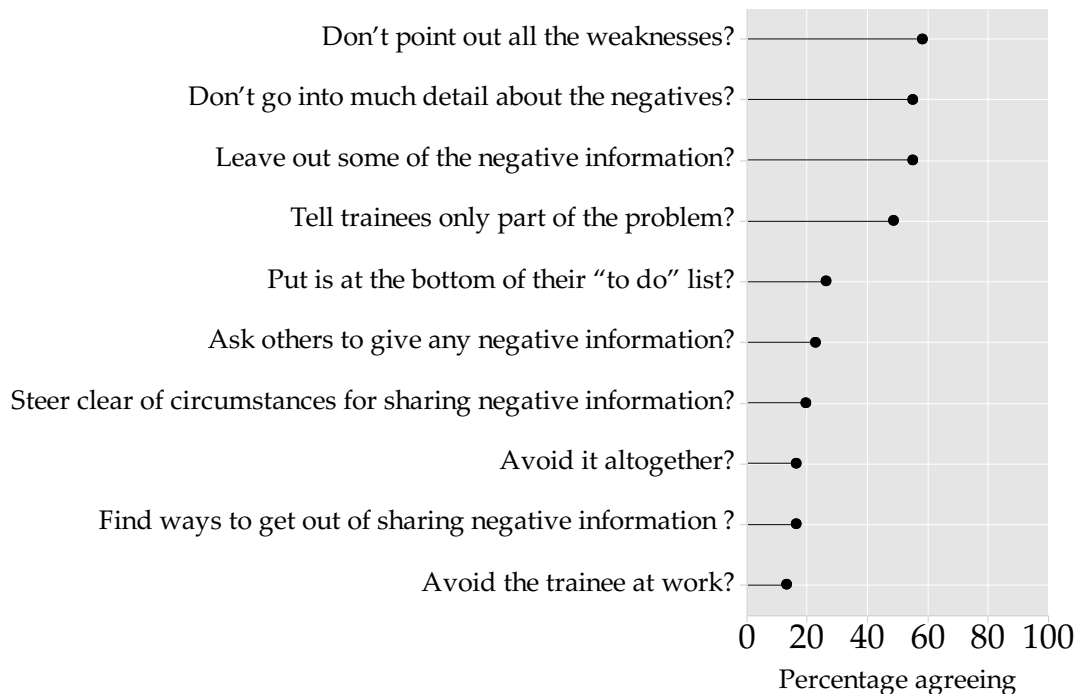


Figure 7.1: Percentage of respondent agreement with MUM avoid subscale

MUM distortion (sugarcoating) subscale

Responses show many trainees agree with statements relating to assessor distortion (sugarcoating) behaviours. Similar to the avoidance subscale, observation of items here seemed to separate the subscale into two parts. There was more agreement from trainees with statements related to trying to make the trainee feel better compared with statements which described changing a message. Almost half of respondents felt that assessors try to make the message sound nicer than it is ($n = 15$, 48%) and just over half ($n = 16$, 52%) felt that assessors try to make the trainee feel better by emphasising positive things. Fewer trainees agreed that assessors put the negative between two positives, change the message to make it more positive or less negative (42%, $n = 13$), or that assessors try to make the information more positive than it really is (39% $n = 12$). The percentages agreeing with statements are shown in figure 7.2 (responses shown in Appendix I, table I.2).

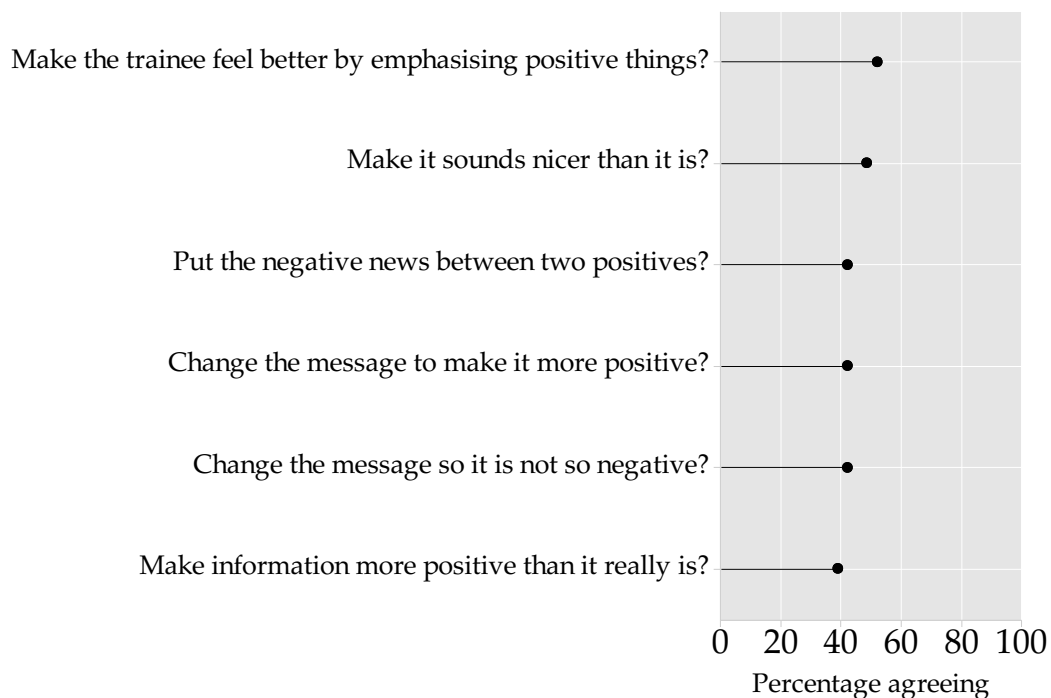


Figure 7.2: Percentage of respondent agreement with MUM distort (sugarcoating) subscale

Other MUM behaviours

Trainees showed much less agreement with Other ways that assessors may behave when faced with delivering negative performance feedback information to trainees. Small

numbers felt that assessors would change the feedback that they are giving during a feedback session (n = 3, 10%) or “Unload” what they had to say quickly before the trainee has a chance to say anything (n = 4, 13%). Just over one quarter (n = 8, 27%) did agree however, that assessors may adopt a very formal manner that doesn’t encourage dialogue when delivering results. The percentages agreeing with statements are shown in figure 7.3 (responses shown in Appendix I, table I.3).

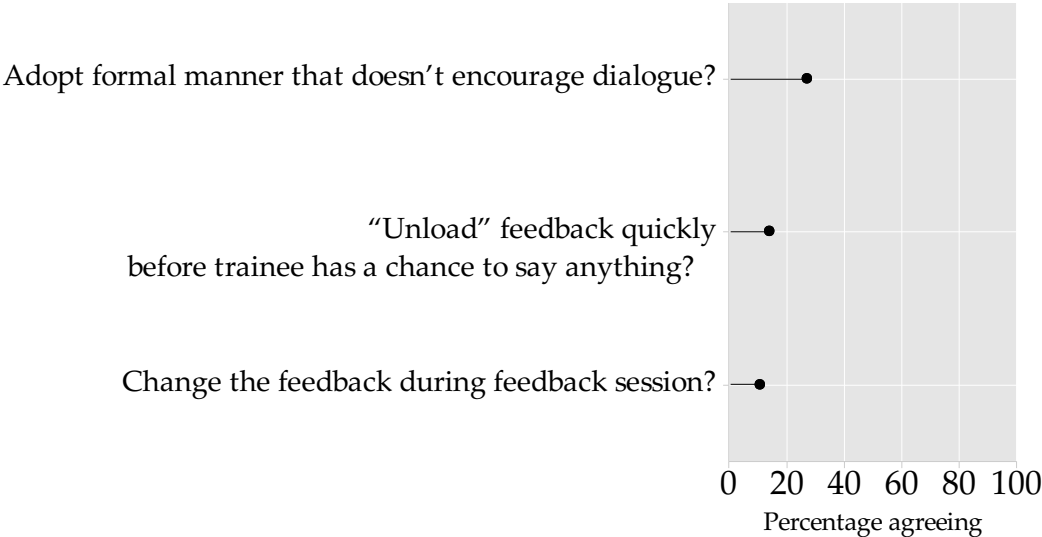


Figure 7.3: Percentage of respondent agreement with MUM other subscale

Figure 7.4 and table 7.2 show the number of items in the subscales agreed with by percent.

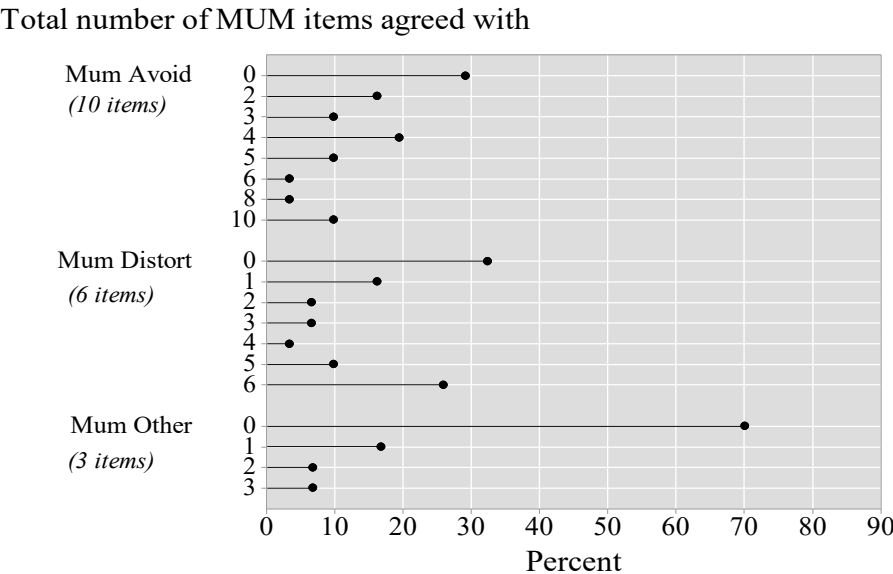


Figure 7.4: Number of items in each subscale respondents agreed with by percent

Table 7.2: Number of items in each subscale respondents agreed with by number and percent

Number of items in subscale respondents agreed with	MUM avoid (10 items)		MUM distort (6 items)		MUM other (3 items)	
	n	%	n	%	n	%
0	9	29.0	10	32.3	21	70.0
1	0	0	5	16.1	5	16.7
2	5	16.1	2	6.5	2	6.7
3	3	9.7	2	6.5	2	6.7
4	6	19.4	1	3.2	n/a	n/a
5	3	9.7	3	9.7	n/a	n/a
6	1	3.2	8	25.8	n/a	n/a
7	0	0	n/a	n/a	n/a	n/a
8	1	3.2	n/a	n/a	n/a	n/a
9	0	0	n/a	n/a	n/a	n/a
10	3	9.7	n/a	n/a	n/a	n/a
Total	31		31		30	

Figure 7.5 shows the proportion of respondents who agreed with at least one item in each subscale by percentage. A total of 71% (n = 22) respondents agreed to at least one item in the MUM Avoidance subscale, 67.7% (n = 21) agreed to at least one item in the MUM Distort (sugarcoat) subscale and 30.0% (n = 9) agreed to at least one item in the MUM Other subscale.

Agreement with at least one MUM item

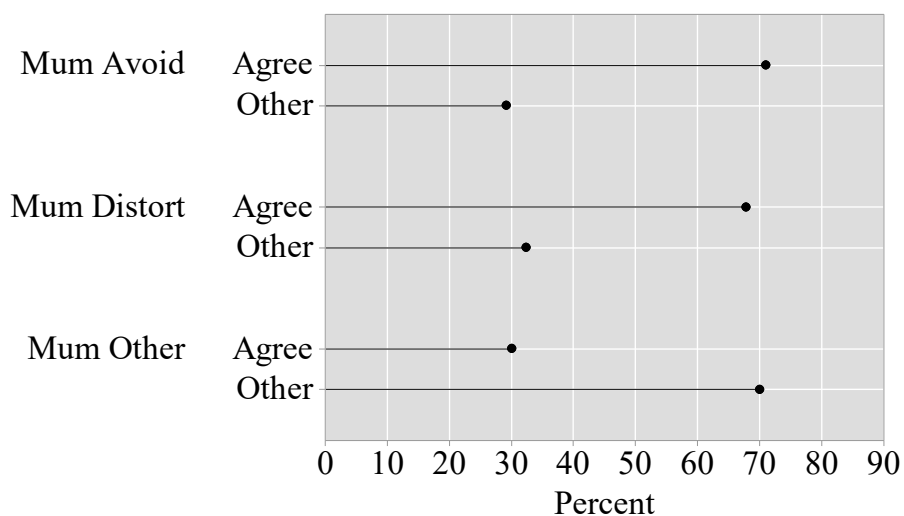


Figure 7.5: Proportion of trainees who agreed with at least one item in each MUM subscale by percent

Comparison of agreement with MUM behaviour items by level of training and gender

Level of training

The number of MUM behaviours trainees agreed with by subscale were compared across trainee level (figure 7.6).

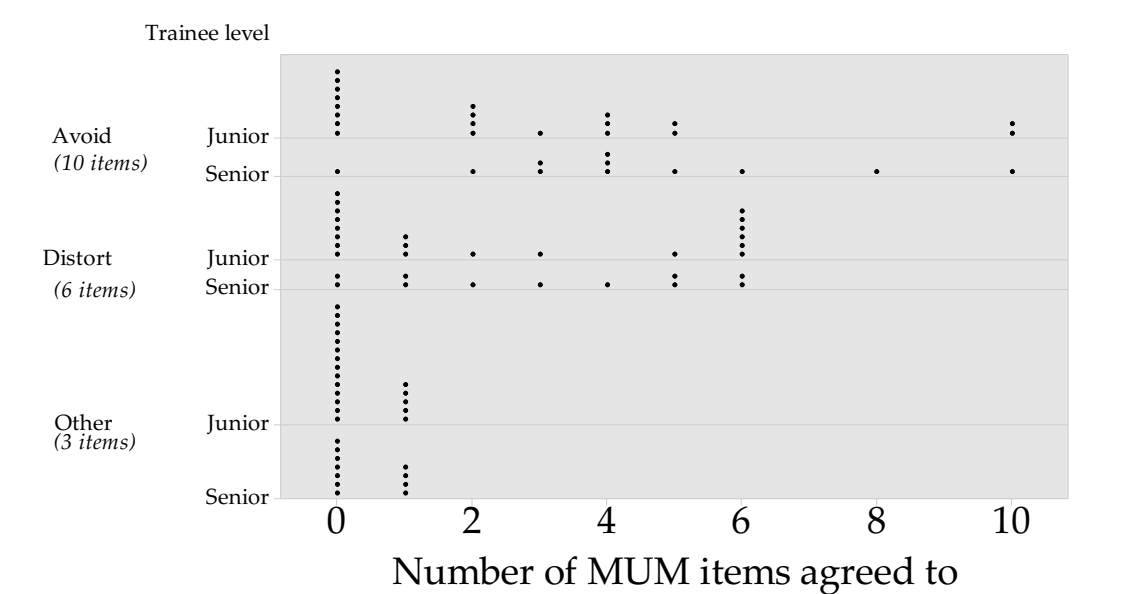


Figure 7.6: Scatter plot showing the number of trainees who agreed with items in each MUM subscale by trainee level

Overall there were no statistically significant differences in the number of items agreed with in subscales based on level of training. Senior trainees tended to agree that they had experienced a greater number of the MUM avoidance behaviours from assessors, but the difference to the degree reported by junior trainees was not significant.

The t-tests showed no significant difference between the mean number of items agreed with by trainee seniority (Appendix I, table I.4).

The proportion of trainees who agreed with at least one item in each subscale was compared with Fisher’s test across level of training (junior compared with senior) and again showed no significant differences (Appendix I, table 1.5).

Gender

Overall, there were no significant differences between the genders in relation to agreement with experiences of MUM behaviours (figure 7.7).

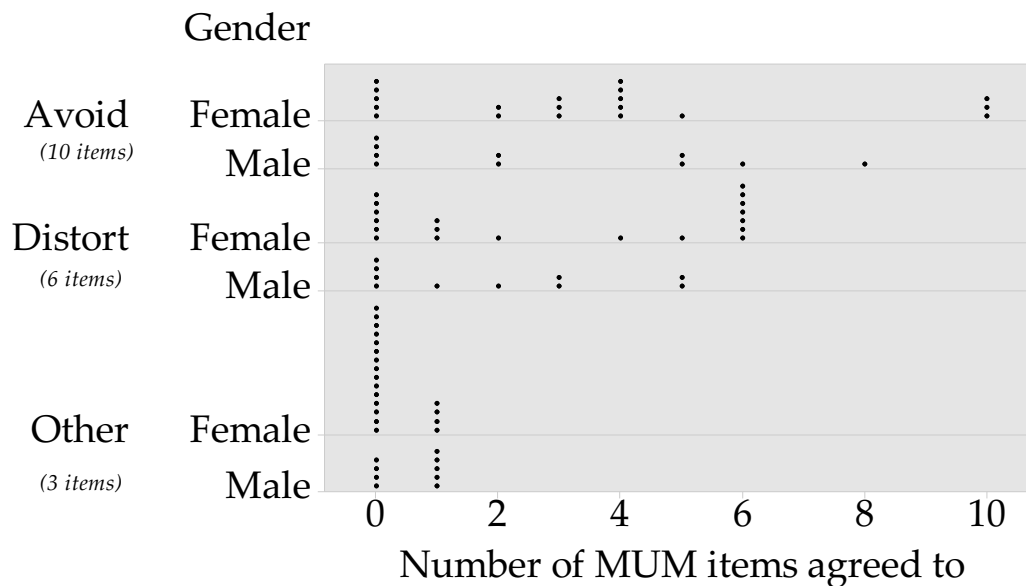


Figure 7.7: Scatter plot showing the number of trainees who agreed with items in each MUM subscale by gender

In relation to MUM avoid behaviours, females tended to agree that they had experienced more MUM avoidance behaviours, but this was not statistically significantly different to the experience of males. Similarly, females were overall more likely to report experience of at least one type of MUM behaviour, but again this was not significantly different to males. In relation to examples of MUM distortion (sugarcoating) behaviours, females were more likely than males to report experience with these behaviours, but the difference was not significant. Finally, male trainees were more likely to report having experienced other examples of MUM behaviour, but this was also not significantly different to females' experience (tables I.16–I.17, figure 7.7).

Comparison of trainee and assessor views on the occurrence of MUM behaviours

Assessor and trainee views of the occurrence of MUM behaviours were compared (assessor responses to the MUM effect scale is presented in Chapter 6, table 6.2). Across all subscales, the mean score for trainee responses was higher than the mean score for assessors (table 7.3).

Table 7.3: Comparison of mean scores by subscale between assessors and trainees

	Assessors	Trainees	Diff in proportion	95%CI Newcombe Wilson	P-value
Avoid	1.78 (n = 23)	2.89 (n = 30)	-1.11	-1.46, -.76	.000
Distort	2.61 (n = 21)	3.10 (n = 31)	-.45	-.90, -.01	.045
Other	1.92 (n = 21)	2.58 (n = 30)	-.66	-1.04, -.27	.001
Total	2.07 (n = 20)	2.92 (n = 29)	-.85	-1.18, -.51	.000

Responses were considered within the subscales (Avoidance, Distortion and Other) and overall. Table 7.4 shows a direct comparison of assessor with trainee responses (percentage and number of respondent agreement).

Trainees showed the highest agreement with zero items in all three subscales (table 7.2, figure 7.4), while assessors showed highest agreement with zero items in the avoid and other subscales and one item in the distort subscale (Chapter 6, table 6.3, figure 6.2). Compared with the assessor results, the percent of trainees agreeing with zero items in the subscale was lower for MUM avoid (29.0 % of trainees agreed with zero items compared with 60.9% of assessors) and MUM Other (70.0 % of trainees agreed with zero items compared with 81.0% of assessors) (Chapter 6, figure 6.2 and table 6.3).

Comparing the percent of respondent who agreed with at least one item, 71.0% of trainees agreed with at least one item in the Avoidance subscale compared with 39.1% of assessors; 67.7% of trainees agreed with at least one item in the Distort subscale compared with 81.8% of assessors and 30.0% of trainees agreed with at least one item in the Other subscale compared with 19.0% assessors (Chapter 6, table 6.3, figure 6.3; Chapter 7, table 7.2, figure 7.5).

Table 7.4: Comparison of assessor and trainee responses to MUM statements (Agree = agree or strongly agree).
Statistically significantly different results highlighted in bold.

MUM statement	Assessors		Trainees		Diff in proportion	95%CI Newcombe Wilson	P-value
	Agree %	n/total n	Agree %	n/total n			
Avoid subscale							
1. I told them only part of the problem	17	4/23	48	15/31	-0.310	-0.507, -0.053	.023
2. I didn't go into much detail about the negatives	4	1/23	55	17/31	-0.505	-0.669, -0.267	.000
3. I left out some of the negative information	17	4/23	55	17/31	-0.374	-0.565, -0.113	.010
4. I did not point out all of the weaknesses	22	5/23	58	18/31	-0.363	-0.560, -0.098	.012
5. I avoided the trainee at work, so I would not have to talk to him/her	0	0/23	13	4/31	-0.129	-0.289, 0.034	.127
6. I asked another consultant to give the trainee any negative information	4	1/23	23	7/31	-0.182	-0.358, 0.018	.119
7. I avoided it altogether	0	0/23	16	5/31	-0.161	-0.326, 0.008	.064
8. I left giving the feedback for as long as I could	4	1/23	26	8/31	-0.215	-0.393, -0.009	.062
9. I steered clear of circumstances in which I'll have to give negative information	0	0/23	20	6/30	-0.200	-0.373, -0.023	.030
10. I found ways to get out of sharing negative information	0	0/23	16	5/31	-0.161	-0.326, 0.008	.064
Distort (sugarcoat) subscale							
11. I put the negative feedback between two positive comments	27	6/22	42	13/31	-0.147	-0.370, 0.114	.385
12. I changed the feedback or rating to make it more positive	5	1/22	42	13/31	-0.374	-0.551, -0.142	.003
13. I tried to make it sound nicer than it was	19	4/21	48	15/31	-0.293	-0.496, -0.027	.042
14. I changed the feedback or rating so it was not so negative	9	2/22	42	13/31	-0.328	-0.513, -0.085	.013
15. I tried to make the feedback more positive than it really was	14	3/22	39	12/31	-0.251	-0.447, -0.003	.065
16. I tried to make the trainee feel better by emphasising positive things	82	18/22	52	16/31	0.302	0.041, 0.502	.041
Other subscale							
17. I changed the feedback during the feedback session due to the trainee's response to it	5	1/21	10	3/30	-0.052	-0.213, 0.138	.634
18. I gave the feedback as quickly as I could and before the trainee could react to it	0	0/21	13	4/30	-0.133	-0.297, 0.041	.134
19. I adopted a very formal or clinical manner when I gave the feedback so the trainee would be less likely to question me	19	4/21	27	8/30	-0.076	-0.287, 0.168	.739

Trainees reported significantly more agreement with partial MUM avoid behaviours compared with assessors. As mentioned, observation of the items suggested the subscale separated into two parts — one describing behaviours of partial avoidance (statements 1–4) and one describing complete avoidance behaviours (statements 5–10). The difference was significant for all the partial avoidance behaviours and one of the complete avoidance behaviours (“I steered clear of circumstances in which I’ll have to give negative information”). Support for the presence of two groups is also provided by noting a difference in the proportion of trainees agreement between these groups. Approximately 50% of trainees agreed with statements 1–4, while less (approximately 10–30%) agreed with statements 5–10. No assessors agreed with Avoid statements 5, 7, 9 and 10. Both trainees and assessors agreed more with the partial avoidance behaviours than more complete ones.

Considering MUM distortion behaviours, a significantly greater proportion of trainees compared with assessors agreed with two of the three statements relating to changing the content of the message (items 12 and 14) and one of the sugarcoating items (item 3). A significantly greater number of assessors than trainees agreed with the statement that they “try to make the trainee feel better by emphasizing positive things” (item 16). The items also seem to separate the subscale into two parts — those which describe the delivery of the message but in a more palatable way (sugarcoating, items 11, 13, 16) and those which involve changing the message (items 12, 14, 15).

There was minimal agreement with any of the items in the Other subscale from either assessors or trainees (items 17, 18, 19). While multiple comparisons between trainee and assessor responses have been made in this section, statistical corrections have not been made. This is consistent with the recommendations of Rothman (245) who warns against such practice, as it may risk masking of differences and so their further exploration. Hence differences seen here may be used to prompt further studies.

Across all items, a greater proportion of trainees than assessors agreed or strongly agreed with all items with only one exception. This was in relation to the statement “I tried to make the trainee feel better by emphasising positive things” (item 16, shown in bold and italics in table 7.4).

In regard to numbers of items agreed with in each subscale, differences between trainees and assessors were also noted. For the avoidance subscale (10 items), trainees showed a difference in their level of agreement with items compared with assessors, with a mean of 2.89 (Likert scale strongly disagree = 1, through to strongly agree = 5), compared with assessors who showed a mean of 1.78 (95% CI 0.758, 1.459, $p < .001$). Trainees also reported higher levels of agreement with distortion subscale items with a mean of 3.07 (across 6 items), while assessors self-reported a mean of 2.61 (95% CI 0.011, 0.896, $p = 0.045$). Trainees had significant difference to assessors in regard to agreement in the Other subscale of three items, where trainees reported a mean level of agreement of 2.58 compared with 1.92 from assessors ($p < .001$) (figures 7.8–7.9).

Figure 7.8 shows a comparison of mean number of items assessors agreed with compared with trainees across the three subscales. Figure 7.9 shows the mean difference between responses across trainees and assessors with 95% confidence interval.

Considering the subscale items as a group, 71.0% (22/31) of trainees agreed that assessors engage in at least one of the avoidance behaviours, compared with 39.1% (9/23) of assessors, giving an estimated difference in proportions of 31.8% ($p = .027$). For the distortion (sugarcoating) subscale, 67.7% (21/31) of trainees agreed that assessors engage in at least one distortion (sugarcoating) item compared with 81.8% (18/22) of assessors, giving an estimated difference in proportions of 14.1% ($p = .348$). The item “I tried to make the trainee feel better by emphasising positive things” was agreed to by the vast majority of assessors while their agreement with other items in this subscale was much less (ranging from 5–27% agreement). As this item appeared to behave differently to the others, it was regarded as an outlying response and comparisons were repeated following its removal. With this item removed, results changed to 58.1% (18/31) trainees agreeing with at least one item in the subscale and 31.8% (7/22) assessors agreeing. This gave an estimated difference in proportions of 26.2% ($p = 0.057$). Figure 7.10 compares agreement with items between assessors and trainees with and without the outlier question “I tried to make the trainee feel better by emphasising positive things”, and figure 7.11 shows the mean difference between responses across trainees and assessors with 95% confidence interval with the outlier

question removed. These are presented for comparison only and the results are discussed with regard to all responses.

There was low agreement with all of the Other subscale items — 30.0% (9/30) of trainees compared with 19.0% (4/21) of assessors agreed with at least one item in this subscale giving an estimated difference in proportions of 11.0% ($p = .518$) (tables 6.3, 7.2, figures 7.8–7.9).

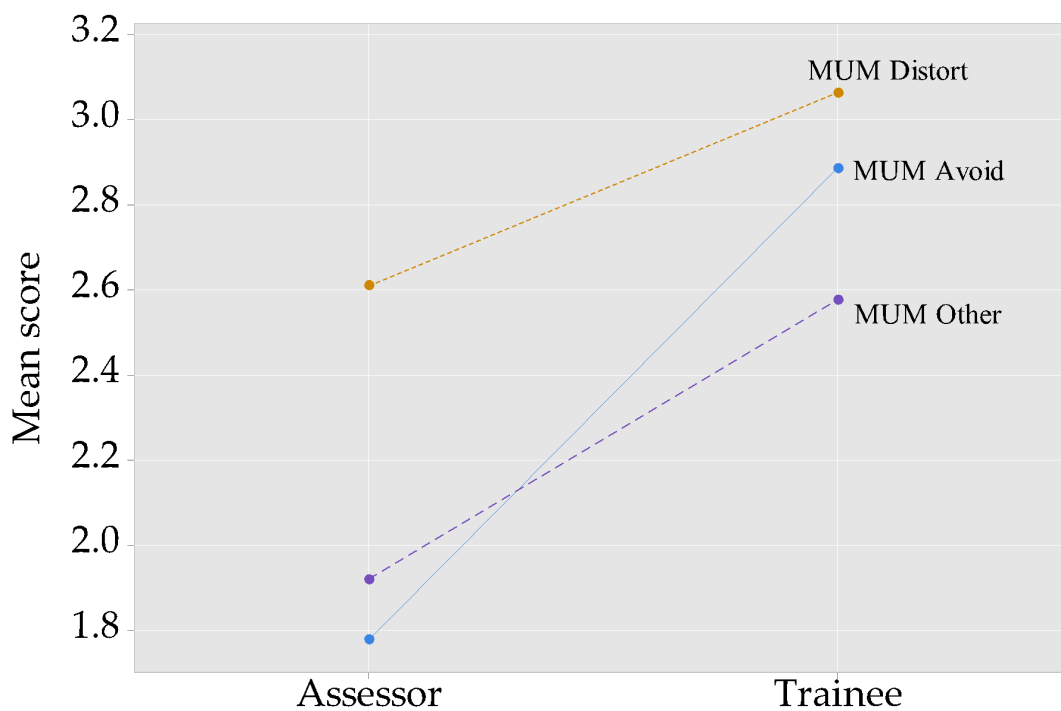


Figure 7.8: Comparison of mean score for assessors with trainees over the three MUM subscales (Likert scale strongly disagree = 1, to strongly agree = 5)

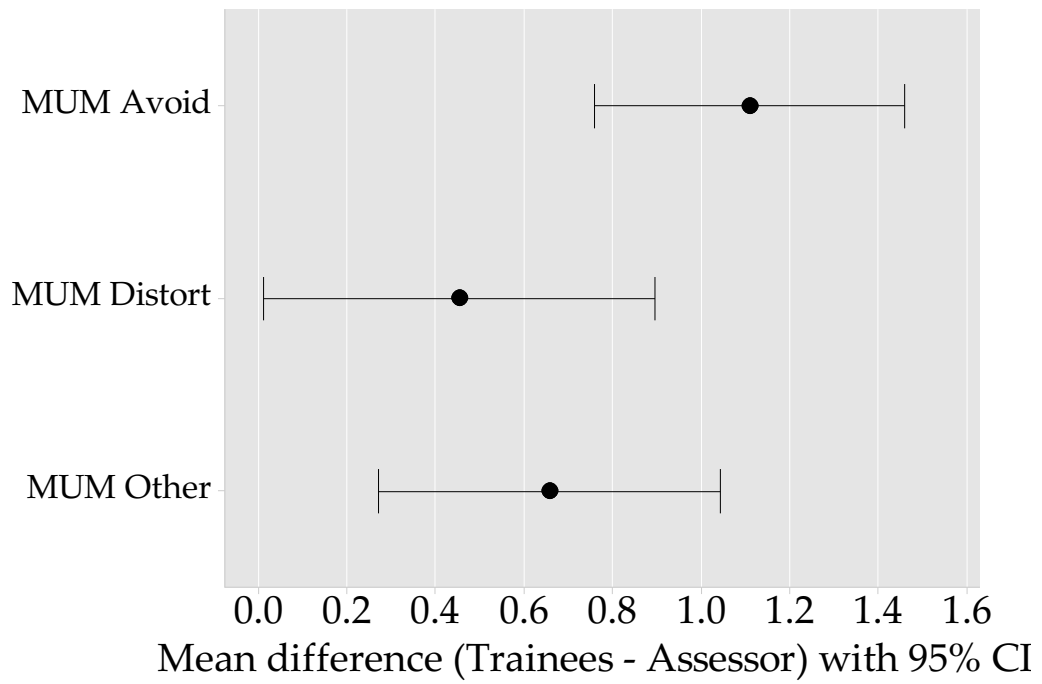


Figure 7.9: Mean differences in scores across the three MUM subscales

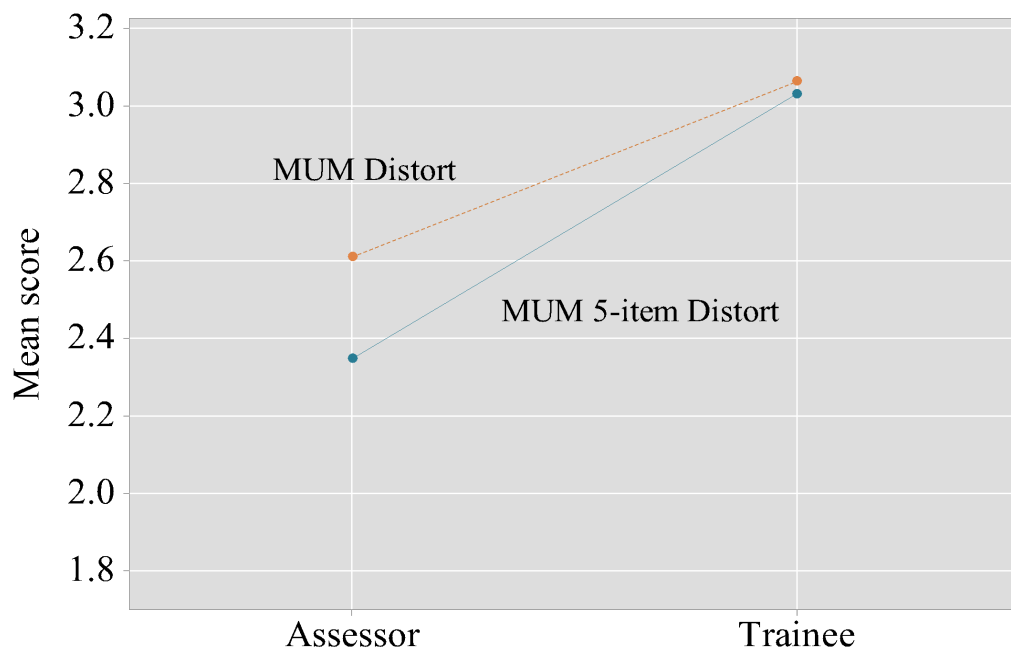


Figure 7.10: Comparison of level of agreement MUM Distort vs 5-item Distort (with the outlier question removed)

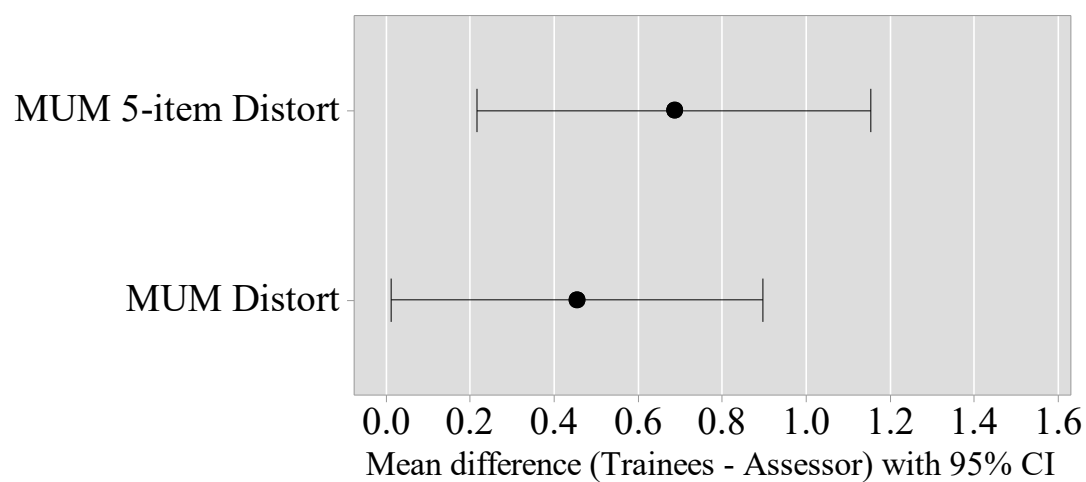


Figure 7.11: Confidence intervals for differences between mean agreement with MUM Distort – full scale vs 5-item (with the outlier question removed)

Instructional Feedback Orientation Scale (IFOS)

Usefulness of feedback subscale

This subscale asked for respondent's opinion on the usefulness of their feedback. Agreement with items in this subscale was high (mean 4.19, $n = 31$) with almost all respondents agreeing with at least one item, with 15 (44%) agreeing with all. In particular, the majority reported that they pay attention ($n = 29$, 94%), listen carefully to feedback ($n = 31$, 97%) and reflect on the feedback given ($n = 28$, 88%). Most also reported feeling relieved ($n = 27$, 87%) and encouraged ($n = 28$, 88%) when receiving positive feedback and that it represents a valuable form of praise ($n = 31$, 97%). Most reported that their feedback motivates them to improve performance ($n = 29$, 94%), however somewhat fewer agreed that it is vitally important in improving their performance ($n = 25$, 78%) and that it provides clear direction on how to improve ($n = 22$, 69%). One respondent (3%) agreed with the reverse coded item that feedback from assessors is a waste of time (it is possible this was an error as the respondent strongly agreed with all other items in this subscale). One respondent (3%) did not agree with any item in this subscale. Figure 7.12 shows the percentage of agreement with statements in the Usefulness subscale (responses shown in Appendix I, table I.8).

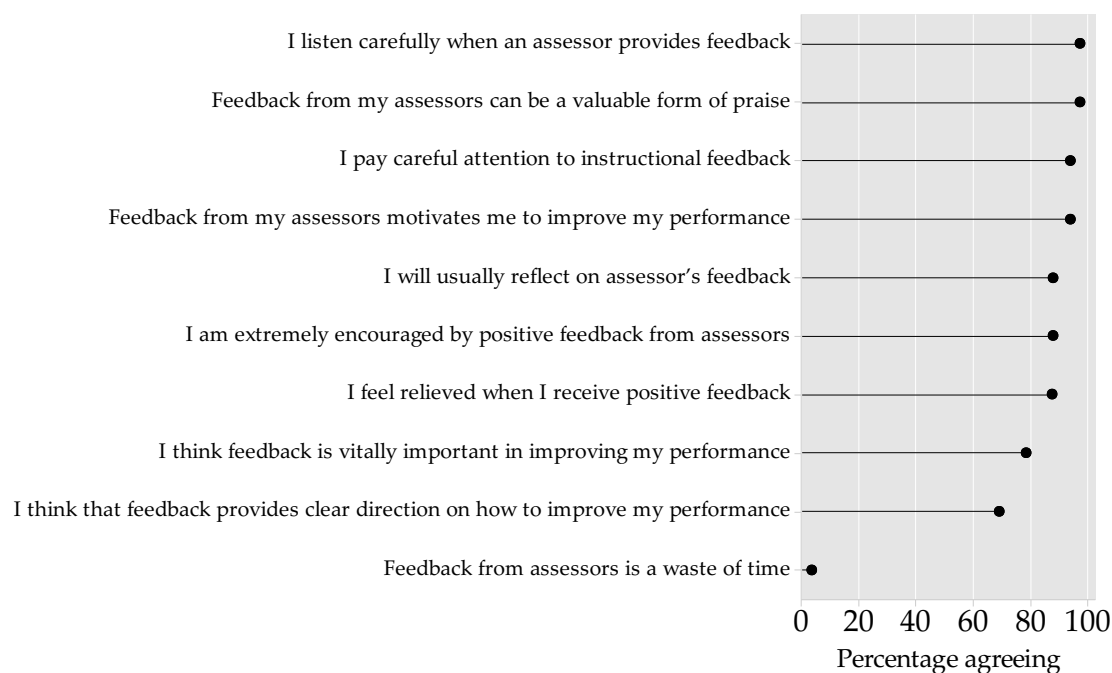


Figure 7.12: Percentage of respondent agreement with Usefulness of feedback subscale

Response/Sensitivity to feedback subscale

This subscale asked about trainees' emotional response or sensitivity to feedback.

Overall trainees showed lower levels of agreement with the items in this subscale (mean 2.91, $n = 31$). While 14 respondents (45%) agreed that negative feedback hurt their feelings, most did not feel their feelings were easily hurt by negative feedback from an assessor ($n = 24$, 77%). Most respondents disagreed or were neutral with the statement that they feel threatened by negative feedback ($n = 27$, 87%) or find it intimidating ($n = 22$, 71%). Similar numbers rejected that negative feedback is difficult to “get over” or is embarrassing or that they dwell on the negative feelings that result from it ($n = 22$, 71%). However, this leaves some 30% of respondents agreeing with these statements. In addition, most ($n = 17$, 55%) agreed that negative feedback from an assessor increases the stress that they feel about future performance. Figure 7.13 shows the percentage of agreement with statements in the Response/Sensitivity subscale (responses shown in Appendix I, table I.9).

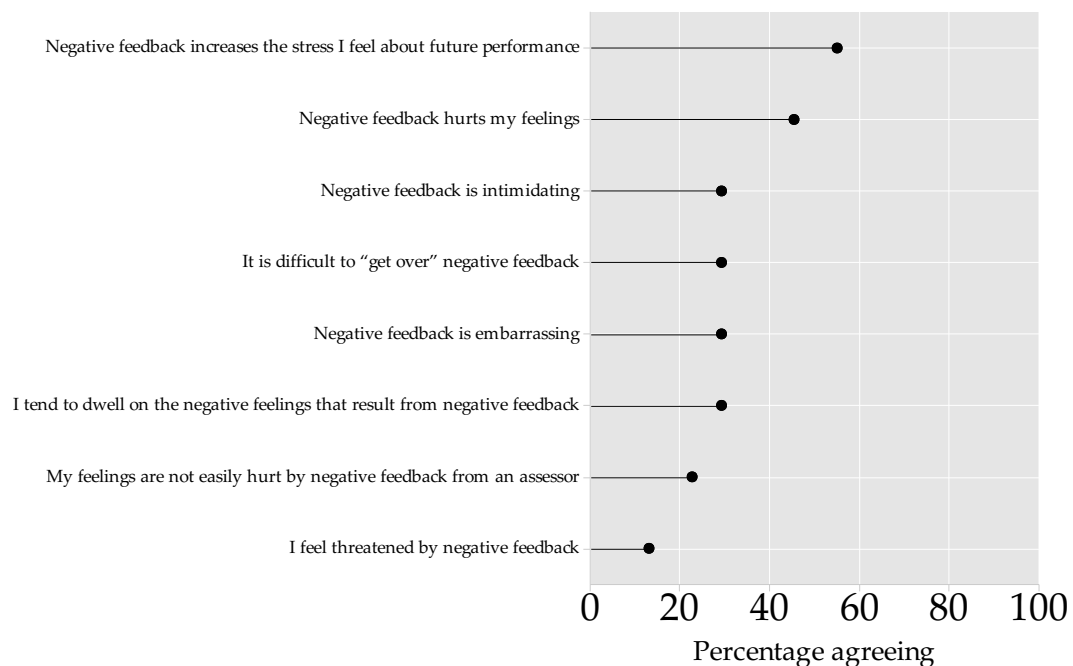


Figure 7.13: Percentage of respondent agreement with Response/Sensitivity subscale

Confidentiality of feedback subscale

This subscale asked about respondent preferences in regards to setting for receiving feedback. The mean for this subscale was 3.90 ($n = 31$). The majority of trainees indicated that they prefer to receive feedback in private ($n = 29$, 94%), although this finding is qualified by the fact that only 74% ($n = 23$) agreed that they don't like others to hear the feedback they are receiving, and 45% ($n = 14$) agreed that they did not mind being singled out by feedback from an assessor. It is possible that trainees answered this question considering all their feedback opportunities, not just those confined to an assessment. Some of these occur in large groups, where they may be singled out, such as in the context of a meeting or ward round. The majority of respondents (81%, $n = 25$) agreed that they don't like receiving negative feedback in front of other people. No respondent agreed that they like others to hear the feedback that they are receiving from an assessor ($n = 31$, 100%), though four gave a neutral response to this. Figure 7.14 shows the percentage of agreement with statements in the Confidentiality subscale (responses shown in Appendix I, table I.10).

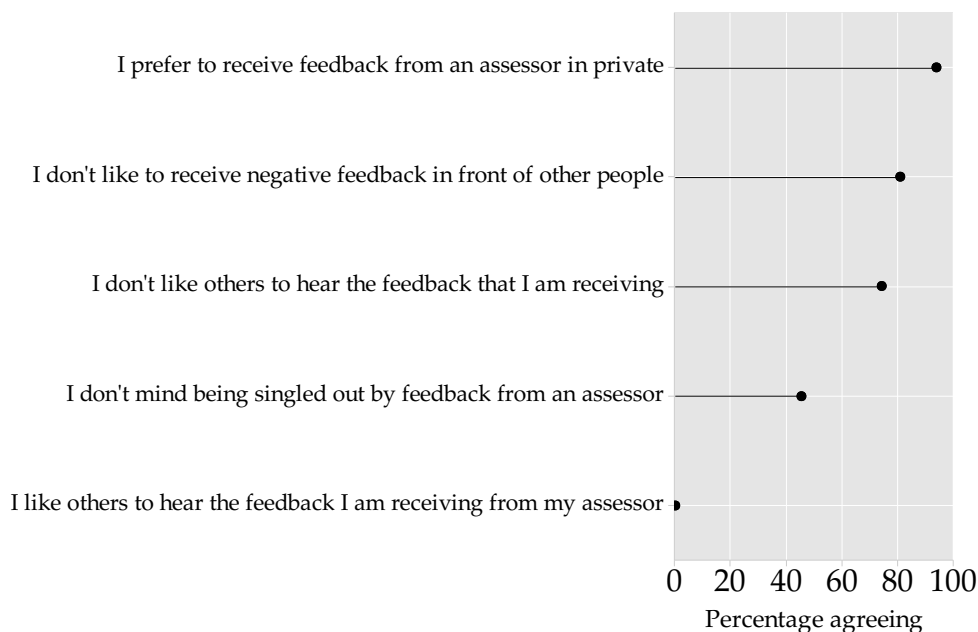


Figure 7.14: Percentage of respondent agreement with Confidentiality subscale

Retention of feedback subscale

This subscale asked about trainees' ability to retain feedback. The mean for this subscale was 3.82 ($n = 31$). Most agreed that they can remember the feedback that they are given, with only a small number agreeing that they cannot remember it ($n = 5$, 16%) or miss the details that they are given ($n = 3$, 10%). The minority also reported not making notes of feedback ($n = 3$, 10%). Figure 7.15 shows the percentage of agreement with statements in the Retention subscale (responses shown in Appendix I, table I.11).

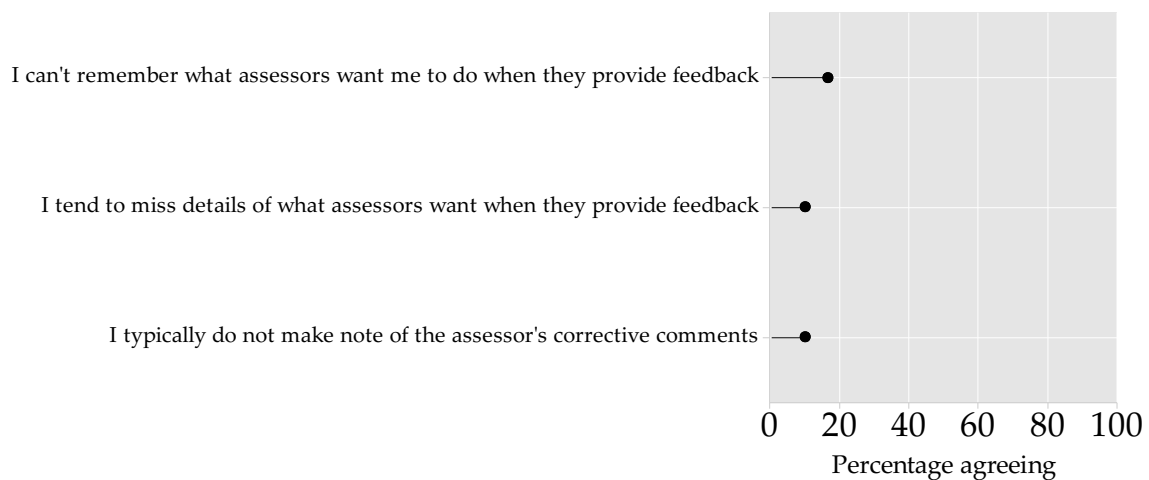


Figure 7.15: Percentage of respondent agreement with Retention of feedback subscale

Comparison of agreement with statements in IFOS and period of training (junior or senior) and gender

Comparisons by trainee seniority

The scatter plot below (figure 7.16) illustrates this large spread of agreement across the individual subscale statements and trainee seniority (level of training).

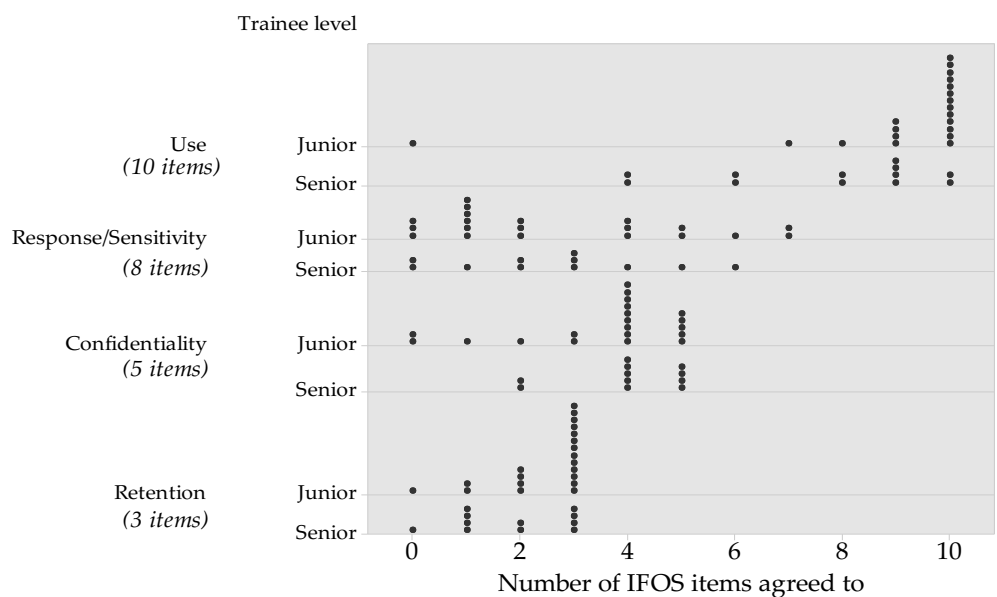


Figure 7.16: Scatter plot of number of IFOS subscale items agreed with by level of training

Independent samples t-tests (with bootstrapping due to non-normally distributed data) show that there were no statistically significant differences between the level of training (junior compared with senior) and the mean number of items respondents agreed with in any of the four IFOS subscales (Appendix I, table I.12). A non-significant tendency to less agreement with the usefulness of feedback and less retention of feedback was seen as trainees became more senior, while preference for confidentiality increased (figure 7.16).

Similarly, Fisher’s test and the difference of proportions (percentages) showed there was no statistically significant difference in the proportion of senior versus junior trainees who agreed with at least one item in the subscales. Though usefulness of feedback did not vary with stage of training, trainee reports of usefulness was overall very high with most agreeing with several statements (Appendix I, table I.13).

Comparisons by trainee gender

Similarly, when respondents were separated by gender, a large spread of agreement across the individual subscale statements was seen (figure 7.17).

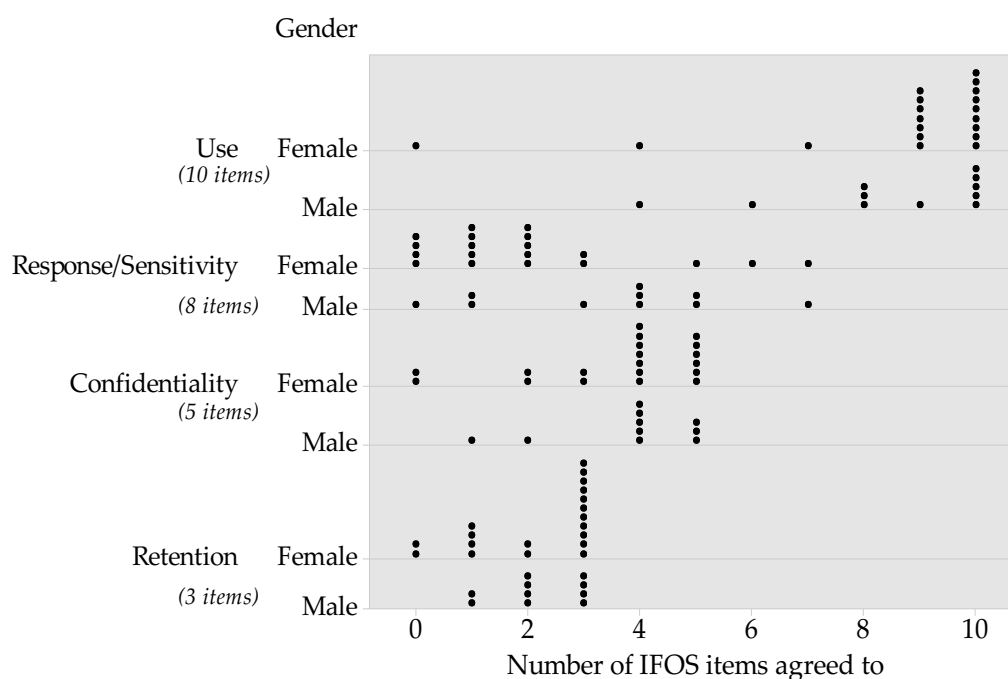


Figure 7.17: Scatter plot of number of IFOS subscale items agreed with by gender

Independent samples t-tests showed that there were no statistically significant differences between the mean number of items trainees agreed with on any of the subscales based on gender (Appendix I, table I.14). Note that two trainees who did not wish to specify their gender have been excluded from these results.

Fisher's test and the difference of proportions (percentages) also show that there were no statistically significant differences between the trainee's gender and agreement with at least one of the items in the IFOS subscales. Specifically, there was no difference in the number of males versus female trainees who agreed with at least one item in the subscales. Response to feedback and confidentiality preferences were more marked between genders, with males tending to report greater sensitivity to feedback, but this did not reach significance (Appendix I, table I.15, figure 7.17).

Experiences of clinical performance assessments

The majority of responding trainees (n = 22, 73%) agreed that it can be a challenge to arrange to complete a work-based assessment (figure 7.18, responses by trainee level and gender are shown in Appendix I, tables I.16–I.17).

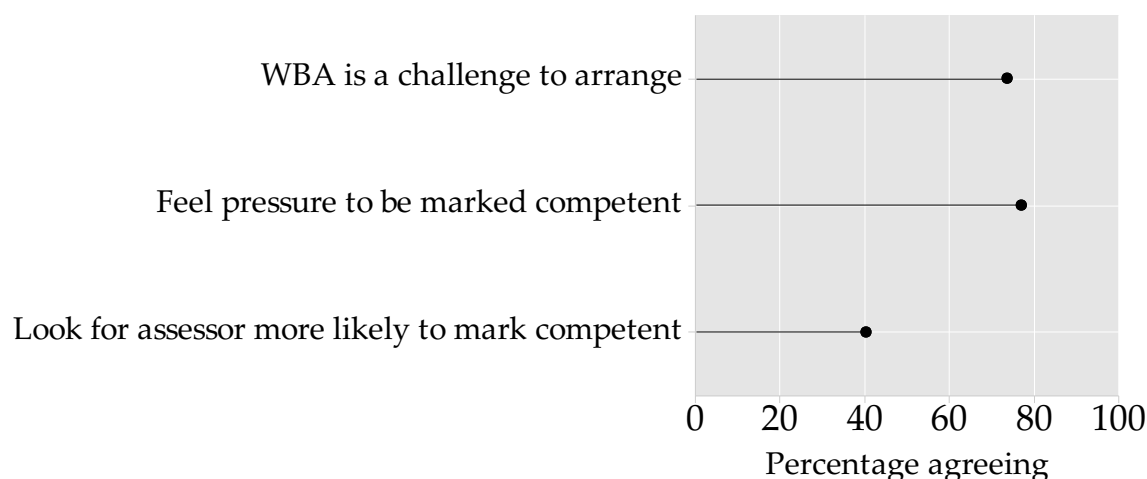


Figure 7.18: Percentage of respondents agreeing to statements about WBAs

In explaining the challenges in arranging a WBA, trainees nominated that it is difficult to find the time (n = 16, 53%) to complete these assessments, it is difficult to find an assessor (n = 3, 10%) and that it is difficult to find a patient (n = 2, 7%). Other issues noted by one respondent were anxiety and embarrassment about asking, stating that they “frequently put it off because I feel my performance needs improvement so by practising longer my performance will be better”.

Most respondents (n = 23, 77%) reported feeling pressure at times to be marked as competent when they complete a WBA (figure 7.18).

In explaining these responses, trainees described themes relating to time pressures and reluctance to waste this and a desire to show that they are performing competently. Some also noted that assessors vary, and one trainee reported asking an assessor who they had more rapport with to assess them (table 7.5).

Table 7.5: Explanations given by trainees who agreed that they do feel pressure at times to make sure that they are marked competent in a WBA

Theme	Illustrative quote
Time pressure	“It is a compulsory part of training so of course you need to be marked competent. It is rushed enough to find the time let alone having to fail and repeat it.”
Reluctance to waste time	“yes, should only flag to do an assessment once you feel you are competent or it wastes time” “A WBA can be time consuming to arrange and inconvenient for assessors. The pressure is there so that you can complete it by the due date without having to spend too much time on it.”
Outcome focus – desire to show competence, impress assessor	“I feel that I don’t want to let the consultants that have invested time in training me down. Approaching the end of my training, I also want to appear competent at all times in case it impacts my future employment prospects as a consultant.” “I study beforehand so I know all about the topic so I don’t sound silly”
Outcome focus – tick-box	“It’s important to get the necessary items during the year to move on to the next task.” “Isn’t that the point?”
Choice of consultant	“you would tend to ask an assessor that you have better rapport with” “Because most of them have different forms of assessment. It is not a standardised assessment”
Consequences	“If I’m marked as incompetent I’ll need to do additional assessments to show I am competent”

Forty percent of respondents (n = 12) indicated that due to feeling pressure to be marked as competent on a WBA they may look for an assessor who is more likely to do so (figure 7.18). By way of explanation, respondents gave the following reasons when they answered “yes” to this question (table 7.6):

Table 7.6: Explanations given by trainees who agreed that they may look for an assessor who is more likely to mark them as competent in a WBA

Theme	Illustrative quote
Tick-box formality	“I am aware of my own level of competency. I know what I need to improve and what am good at. These assessment tasks don’t help me– they are just a formality. Perhaps in the younger years, there were more effective for me but now, I don’t feel they add anything. Therefore, I just need them done by someone who will do them quickly and with no fuss.”
Desire for positive reinforcement	“I think this is just human nature!”
Difference among assessors	“Different assessors have different approaches. Some assess as competent for level of training and others assess at fully competent as a dermatologist.”
Choice of consultant	“I just look for someone I’m comfortable asking irrespective of whether they’re in easy or hard marker”

There were no associations observed between trainee gender or level of training and reported challenge in arranging a WBA (estimated difference in proportions 0.070, $p = 1.000$ and 0.153, $p = 0.417$ respectively, Appendix I, tables I.16–I.17). There were no statistically significant differences in reports of feeling pressure to be marked competent between trainee genders (estimated difference in proportions 0.123, $p = 0.794$, though a tendency for females to feel more pressure (table I.17). Similarly, there was no statistically significant difference based on trainee level of training (estimated difference in proportions 0.349, $p = .068$, table I.16). There were no statistically significant differences between trainee level of training (estimated difference in proportions 0.057, $p = 1.000$, table I.16) or gender (estimated difference in proportions 0.129, $p = 0.228$, table I.17) and reporting to look for an assessor who is more likely to mark the trainee as competent. Trainees were asked whether they had encouraged or asked an assessor to mark them as competent on a WBA. One trainee out of 30 responded “yes” to this question but did not describe the details. Trainees were asked whether they had encouraged or asked an assessor to give them a certain rating on a SITA to which all 30 responding trainees responded “no”. Trainees were finally asked whether they had encouraged or asked an assessor to change a score that they have been given on a WBA or a SITA. One trainee out of 30 responded “yes” to this and described the situation. The trainee reported:

“I completed a case discussion where my assessor said I did an excellent job on a case of high complexity, and said my answers to questions were correct. Yet he marked me as incompetent and gave me all ‘ones’. I pointed out that he made a mistake and obviously ticked the wrong boxes based on verbal feedback, he commented that he intentionally failed me. I was too surprised to ask for clarification, he didn’t volunteer any.”

Comparison of trainee and assessor responses to statements about direct approaches

Comparison of assessor and trainee responses to statements about direct approaches are show in table 7.7.

Table 7.7: Comparison of self-reports of trainees asking assessors to give or change a rating in an assessment with assessor reports of being approached by a trainee in relation to an assessment

Have you or has a trainee:	Trainees (n)				Assessors (n)			
	Yes	No	n/a	Total	Yes	No	n/a	Total
Encouraged or asked to be marked competent on WBA	1	29	0	30	2	40	6	48
Encouraged or asked for particular SITA rating	0	0	0	30	3	44	1	48
Asked to change a previously assigned rating	1	29	0	30	1	46	1	48

This table presents responses from assessors alongside those from trainees. In relation to the first two questions, assessors report being approached by trainees more than trainees report approaching them. In relation to the third question, while both one assessor and one trainee report this situation occurring, it is likely that these responses reflect two different situations rather than the same one.

To summarise, in regard to all items where trainee and assessor responses to the same statements were compared, a greater proportion of trainees agreed with the (16) MUM scale statements and the additional (6) statements (Other MUM subscale and questions relating to direct approaches to assessors) with three exceptions. These were in relation to the MUM distortion (sugarcoating) item of trying the make the trainee feel better by

emphasising positive things; a trainee encouraging or asking to be marked competent on the WBA; and a trainee encouraging or asking for a particular rating on a SITA, where a greater proportion of assessors than trainees agreed.

Other free-text comments

Trainees were asked to provide any free-text comments they wished at the end of the questionnaire. Table 7.8 below shows the responses entered, arranged by theme. As shown, trainees highlight their views of the usefulness of the clinical performance assessments, though believe that level of training, logistical issues and the trainee and assessor relationship can influence them. Comment around the themes of the importance and potential helpfulness of the assessments were made and also that assessors should be assessed. Some comments reflected perceptions around awareness of avoidance of the delivery of negative messages at times and the limitations of feedback.

Table 7.8: Comments made by trainees in response to the question “Are there any comments that you would like to make about the WBA and/or SITAs as they currently are?”

Theme	Illustrative comment
Usefulness of assessments varies with training level	“More useful for junior years of training. I am already aware of my strengths and weaknesses”
Logistical difficulties	“WBA are tedious and difficult to complete during work hours (usually time pressures of supervisors)” “forms are too long”
Impact of relationship	“I think there is perhaps a conflict in that your future close colleagues (with whom you are developing a professional and possible personal relationship with) are in a position of grading your progress”
Importance of assessments	“I think the feedback process is absolutely vital for trainees ... negative feedback MUST be relayed to trainees” “This sort of feedback is extremely valuable ... however often assessors are going through the motions because of an obligation. This sort of feedback is too important to be left to supervisors who aren’t passionate and proactive in medical education”
Helpfulness potential	“WBA are really helpful from a trainee perspective if helpful advice can arise from a negative WBA”
Assessors should be assessed	“Registrars should have the opportunity to provide feedback on their supervisors”
Avoidance of delivery of assessment messages	“Most people that need to improve don’t receive the appropriate feedback in what to do better because the assessor doesn’t have the strategies to help improve the deficits of the trainees” “The supervisors MUST give negative feedback for people to have an opportunity to improve and do better. A task that is often very difficult and frequently avoided” “I think it is difficult in any situation to give negative feedback. I have hardly ever received constructive criticism ... and I am sure this is not because I am a perfect trainee!”
Limitations of feedback	“There is not really room for positive feedback” “Simply making non-specific observations, especially if subjective is unhelpful feedback”

Comparison of trainee IFOS responses to trainee reports of MUM behaviours by their assessors

Trainee IFOS responses to the Usefulness and Response/Sensitivity subscales were compared to their reports of MUM behaviours by their assessors to look for associations. These subscales were tested as they were felt to be the most likely to be associated with MUM behaviours. For example, if a trainee reported high usefulness of their feedback, they may behave in a way that supports assessors delivering their feedback and not engage in MUMing; while conversely those who report high sensitivity to feedback may discourage assessors from delivering their message (94).

There were no significant associations between reports of experiencing MUM Avoidance behaviours and agreement with one or more items on the IFOS subscales Usefulness of feedback and Response/sensitivity to feedback (estimated difference in proportions 0.045 and 0.086 respectively, $p = .688$ and $.458$ respectively, table 7.9).

Table 7.9: Comparison of percentage agreement with MUM avoid behaviours and IFOS subscales.

Percentage agreeing			Difference in percentages (Agree with at least one MUM behaviour occurring – do not agree/neutral to at least one MUM behaviour occurring)		
	MUM Avoid agree (n = 22)	MUM Avoid not agree/neutral (n = 10)	Estimate	95% CI	<i>P-value</i>
IFOS Usefulness agree	95.5% (n = 22)	100% (n = 10)	–4.5%	–.218, .235	.688
IFOS Response/Sensitivity agree	86.4% (n = 22)	77.8% (n = 9)	8.6%	–.167, .423	.458

Similarly, there were no significant associations between those experiencing MUM distortion behaviours and agreement with at least one item on the IFOS Usefulness of feedback and Response/Sensitivity to feedback subscales (estimated difference in proportions 0.091 and 0.057, $p = .344$ and $.528$ respectively, table 7.10).

Table 7.10: Comparison of percentage agreement with MUM distort behaviours and IFOS subscales

Percentage agreeing			Difference in percentages (Agree with at least one MUM behaviour occurring – do not agree/neutral to at least one MUM behaviour occurring)		
	MUM Distort agree (n =)	MUM Distort not agree/neutral (n =)	Estimate	95% CI	<i>P-value</i>
IFOS Usefulness agree	91% (n = 21)	100% (n = 11)	–9.1%	–.377, .081	.344
IFOS Response/Sensitivity agree	85.7% (n = 21)	80% (n = 10)	5.7%	–.192, .381	.528

Finally, there were also no significant associations between those reporting to experience the Other MUM behaviours and agreement with one or more items on the IFOS Usefulness of feedback and Response/Sensitivity to feedback subscales (estimated difference in proportions 0.043 and 0.071, $p = .719, .542$ respectively, table 7.11).

Table 7.11: Comparison of percentage agreement with MUM other behaviours and IFOS subscales

Percentage agreeing			Difference in percentages (Agree with at least one MUM behaviour occurring – do not agree/neutral to at least one MUM behaviour occurring)		
	MUM Other agree (n =)	MUM Other not agree/neutral (n =)	Estimate	95% CI	<i>P-value</i>
IFOS Usefulness agree	100% (n = 9)	95.7% (n = 23)	4.3%	–.258, .21	.719
IFOS Response/Sensitivity agree	88.9% (n = 9)	81.8% (n = 22)	7.1%	–.271, .294	.542

Further analyses were performed to look for correlations between variables. Bivariate correlations analyses showed a very small and non-statistically significant correlation between Usefulness and Response/Sensitivity scales and the number of MUM statements respondents agreed with (table 7.12). The scatter plots show the correlation of the IFOS scales compared with number of MUM items agreed with separately (figures 7.19–7.20).

Table 7.12: Pearson correlation – IFOS scales compared with number of MUM statements respondents agreed with

IFOS subscale	Number of MUM statements agreed with	
	Pearson correlation	P-value
Usefulness	−0.074	.699
Response/Sensitivity	.023	.902

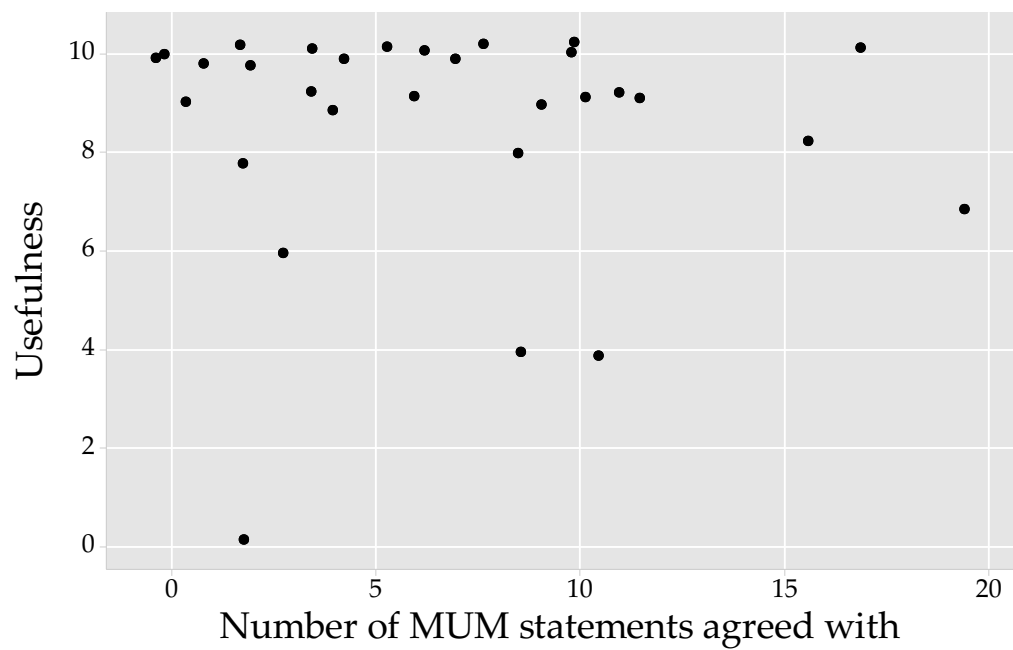


Figure 7.19: Scatter plot showing correlation between trainee response on Usefulness scale and the number of MUM statements agreed with

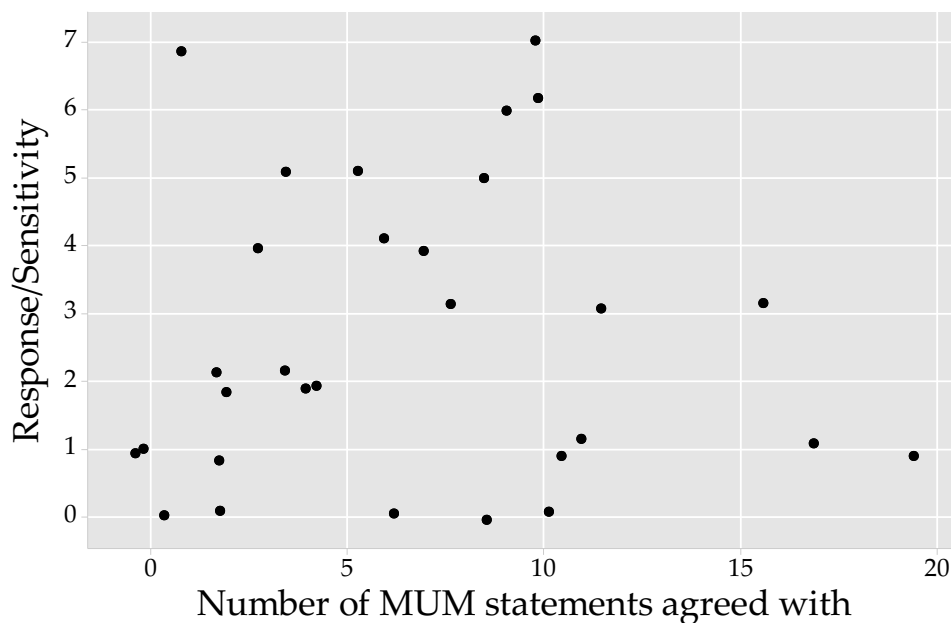


Figure 7.20: Scatter plot showing correlation between trainee response on Response/Sensitivity scale and the number of MUM statements agreed with

There were no statistically significant differences between the reported challenge of arranging a WBA and reported agreement with one or more items on the IFOS Usefulness of feedback and Response to feedback subscales (estimated difference in proportions 0.45, 11.4, $p = .733, .405$ respectively, Appendix I, table I.18); nor with any of the MUM subscales. However, it is noted that there was overall very high agreement with the Usefulness subscale, with all but one respondent agreeing to at least one item on it. Further, there were no statistically significant differences between reported pressure to be marked competent on a WBA and reported agreement with one or more items on the IFOS Usefulness of feedback and Response to feedback subscales (estimated difference in proportions 0.43, 15.5, $p = .767, .329$ respectively), nor with any of the MUM subscales (Appendix I, table I.18). There were no statistically significant differences between seeking an assessor more likely to mark as competent on a WBA and reported agreement with one or more items on the IFOS Usefulness of feedback and Response to feedback subscales (estimated difference in proportions 0.83, 0.139, $p = .4, .318$ respectively), nor with any of the MUM subscales (Appendix I, table I.18).

Discussion

Summary and interpretation of findings

This study has several key findings. Firstly, it shows that many trainees believe that their assessors do engage in MUM behaviours. A large minority to over half of trainees [between 39% (n = 12) and 58% (n = 18)] believe that their assessors engage in partial avoidance and distortion (sugarcoating) behaviours, and some believe they engage in more complete avoidance behaviours [13% (n = 4) to 26% (n = 8)]. Trainee level of training or gender was not associated with their views of MUM behaviours by their assessors, and this study did not find an association between trainee views of feedback and their experience of MUM behaviours. That is, an association between higher opinions of the value of feedback and lower experience of MUM behaviours, or a greater sensitivity to feedback and higher levels of MUM behaviours were not found. Even so, trainee perceptions that their assessors are engaging in MUM behaviours has important implications for their understanding and interpretation of the assessment messages they receive. As discussed, while recent work confirms that trainees can decode the narrative comments that are presented in ITAs (38), it is not known if all trainees can do this to the same extent or whether they can do this for their own assessments. Similarly, it is not yet known whether trainees can factor in missing comments (assessors' MUM avoidance), nor how they interpret a delay in receipt of assessment messages.

Secondly, in addition to showing that many trainees agree that their assessors are engaging in MUM behaviours, this study also shows that trainees differ from their assessors in the extent to which they believe such behaviours are occurring. As noted, with only one exception, a greater proportion of trainees than assessors agreed with the individual MUM statements. These differences were statistically significant for five of the ten avoidance subscale items and four of the six distortion (sugarcoating) subscale items.

It was seen that assessors and trainees generally agreed with the items which described partial MUM avoidance — items such as not going into too much detail or leaving out some of the detail. Much less agreement was seen with items which describe complete avoidance — such as avoiding it altogether or asking someone else to give the message.

Overall, assessor agreement with distortion items was greater compared with their agreement of avoidance subscale items, and their agreement was higher for sugarcoating as compared with distortion ones. This suggests that assessors admit to trying to make the message sound as positive as possible, but reject the notion that they alter the result.

The only distortion subscale item that a majority of assessors agreed with was that they tried to make the trainee feel better by emphasising positive aspects. Over 80% (n = 18) of assessors agreed with this item, compared with just over 50% (n = 16) of trainees. This statement may have been viewed by assessors as being supportive practice and it is arguably the least distorting item of all in the scale. Even so, it is important to examine why assessors choose to do this and the possible implications that this may have. For example, while small amounts of sugarcoating may be beneficial to performance (136), excessive levels may be counterproductive. However, the counter-argument is that given trainees have some awareness of the occurrence of distortion, the message will still get through as trainees “read between the lines” (38). Assessors and trainees both showed much lower levels of agreement with items in the other MUM subscale items but trainees agreed with them more than assessors. While few assessors admitted to adopting a formal or clinical manner in order to minimise questions from the trainee, over one quarter of trainees reported having experienced this. While as presented, corrections for multiple comparisons were not made, the observed differences in results between assessors and trainees should prompt further exploration.

Considering trainees’ views of their assessment, the results of this study show that they do find their clinical performance assessments overall to be useful, with high agreement with the Usefulness subscale. Agreement was high around items which described paying careful attention to and reflecting on the feedback. However, agreement was also high to items which described relief and encouragement from positive feedback and that feedback is welcomed as a form of praise. This seems reminiscent of the work of Klein, who suggested trainees’ desires were for “feeding” not “feedback” (98) and may explain the slightly higher agreement with items relating to positive reinforcement, compared with constructive (and therefore potentially challenging) feedback. As will be discussed below, some trainees also recorded comments which reflected the perception of limited usefulness of the assessments. Hence, similar to the systematic

review findings, trainees differ in their individual views of the usefulness of the assessment messages they receive.

Trainees reported low levels of agreement with Response/Sensitivity scale items, indicating a low level of emotional response to feedback. However even so, it is important to note that large minorities (around 30%) did find that feedback could be intimidating, embarrassing or difficult to get over. In addition, it is an important finding that over half the respondents agreed that negative feedback from an assessor increases their stress about future performance. In light of the systematic review findings, this may be particularly relevant for trainees who value the summative aspects of the assessment over the formative ones. This finding also adds to reports from others about trainees' experiences of stress and anxiety in the clinical performance assessment setting (186,196,197,199,204) and its potential to influence performance (197). Perhaps unsurprisingly, trainees almost universally wished for confidentiality in regards to the delivery of results and assessors should be aware of these results. While retention of feedback was reported to be high amongst this cohort, again assessors should appreciate that approximately one in ten trainees report that they struggle to collect and retain all their feedback. Though respondent numbers were small, there were no statistically significant differences demonstrated in regards to trainee responses to the feedback scale items when examined by level of training or gender, suggesting that these results are relevant across all trainees.

In this study, over three quarters of respondents admitted to feeling pressure to be marked as competent on the WBA. While it may be surprising that not all trainees reported feeling this pressure, those that didn't may be the ones who value the formative aspects of the assessment over the summative ones. The fact that 40% of those trainees who felt pressure to be marked competent may go on to select an assessor who is more likely to mark them as competent also echoes reports in the literature (187,200) though these results contrast with those from the first part of the questionnaire, where trainees indicated in large numbers that they value the feedback that they obtain. Seeking out assessors who they feel are more likely to pass them demonstrates the "vicious" as opposed to "virtuous" attitude to WBAs held by some trainees as described by Castanelli and colleagues (187), likely promoted by the as yet unresolved practical and logistical issues, including difficulties in arranging them to be done, the number of

assessments required to be completed annually, and the consequences of failing to achieve this (including potential for lack of progression). The pressures involved seem to lead some trainees to seek out MUMing assessors in order to have their assessments ticked off, with the consequence of missing important developmental opportunities. These factors warrant consideration if these assessments are to achieve their purpose and be seen as valuable contributions to a trainee's learning and development, and not merely a "tick-box" exercise (186).

In comparison, few trainees reported approaching an assessor in relation to assigning or changing a certain rating. Though again the numbers were very low, when these results were compared with that of their assessors; more assessors reported being approached by trainees for this purpose. This may be explained by different interpretations of the question, different perspectives or recollections of the same event, or reluctance to admit to such behaviour by trainees. The fact that even small numbers of trainees did report such an occurrence, however, suggests this is of importance.

Some of the free-text comments written by respondents tell a somewhat different story to the numerical results, and show the variation of views held by trainees. Respondents explained about the pressures that the WBA place on them, especially in terms of the logistics involved in arranging them (having noted that over three quarters reported finding them a challenge to arrange). The consequent pressure this places on them to be marked competent was also explained by some, both so to not waste the assessor's or their own time, and also to avoid possible future negative effects, such as on their employment prospects. While some respondents recognised the importance of the assessments, others alluded to awareness of lack of standardisation between assessors as a negative, and some comments suggested the assessments are viewed as tick-box tasks, just to be completed so trainees could move on. Others were more frank, stating that the assessments were of no help, merely a formality, while others were aware that at times assessors avoided delivering their assessment message. Such factors may have influenced some trainees to behave in a more performance goal-oriented than learning goal-oriented manner, as discussed in Chapter 4, the systematic review.

Implications

The results from this study show that trainees believe their assessors are engaging in MUM behaviours and to a greater extent than their assessors believe they are. As Ende highlighted, when there are concerns about the assessment message to be delivered, the use of indirect statements (seen now as the distortion of the message) may “obfuscate the message entirely” with the result that “nothing of any real value gets transmitted or received” (3). While both the results from this chapter and the preceding one are based on self-report only, they do give some indication of the extent of such behaviours. As the social desirability bias predicts that the assessor reports to be an underestimation of MUMing, its actual occurrence may be more at the scale noted by trainees. The importance of the delivery of truth in the assessments in order to minimise any miscommunications arising from them is clear. Where results are not being delivered, the trainee is denied information on their performance, with possible implications for their future development. This leads to potential confusion as how to interpret what they are told — “I don’t know if they’re not saying things. I don’t know if they’re avoiding saying things they want to say” (38). This may foster a sense of mistrust in the process and so reduce engagement further, and the importance of engagement in the assessment process has been highlighted earlier (102).

Future directions

There are several important implications of this work. Future work should look to develop an understanding of how trainees reconcile their impressions of MUM behaviours by their assessors, especially avoidance, and the impact that it has on trainee interpretation of their assessment messages, so that ultimately the perceived usefulness of the assessments may increase further. The difficulties reported in arranging and performing these assessments cannot be ignored, given most trainees found them to be a problem. Such pressure may foster other behaviours, such as the two examples where trainees admitted to approaching assessors in relation to giving or changing an assessment result. As these may be counterproductive and work against the aims of the formative value of assessment, further evaluation in order that these assessments can truly provide trainees with quality assessment messages to help their professional development would appear necessary.

Whilst it is positive that trainees generally report they find the feedback received in the context of clinical performance assessment to be useful for their learning, this needs to be considered in context. This includes the fact that some trainees seem to value praise above feedback and that some admit to strategically choosing an assessor in order try to secure a positive outcome. The logistical difficulties in arranging assessments highlighted earlier is given as one possible explanation for this.

Trainees reported relatively low levels of sensitivity to their assessment messages and this may be lower than what some of their assessors expect. Knowledge of trainees' desire to hear assessment messages, even if negative, may encourage some assessors to communicate their assessment message more completely (150), and so decrease their inclination for MUM avoidance and distortion (sugarcoating) behaviours. Open acknowledgement of the MUM effect and how it may present could help promote more honest feedback exchanges and so decrease the need for "reading between the lines", with its possible misinterpretations.

Some trainees reported their impression of issues with subjectivity and differences in the rating practices of their assessors, the importance of being rated as competent, and the lack of constructive criticism or difference between ratings and comments. This provides a rationale for the next study in this project — a comparison between the ratings and comments on a sample of real-world assessment forms.

Limitations

The response rate to the questionnaire was low at 30%, despite attempts to increase participation through reminders. Subsequent information showed that the trainees have relatively low levels of monitoring of email communication from the ACD, and this may have contributed to the low response rate, given it was the method of delivery of the invitation to participate. Potential respondents may have also been concerned that the invitation to participate came from the ACD itself, and so had concerns about the confidentiality of their responses. Methods to attempt to maximise the response rate were made, including keeping the questionnaire as short as possible (220), but incentives for completion were not offered as it was felt these could possibly influence responses. The social desirability bias, though of less influence in an anonymous questionnaire, may still operate and influence responses obtained. Of course, the fact

that some respondents did admit to behaviours such as requesting an alteration to a mark, shows that such issues do occur and may have been under-represented in the results. Reliance on self-report of behaviours, as was the case here, is less objective than observation and subject to recall bias, in addition to other biases. Where trainee responses have been compared to those of assessors, it is important to appreciate that the samples were not paired. As such, the assessors to whom trainees refer, may be different from those who responded to the assessor questionnaire. Finally, the questionnaire presents largely quantitative data, which though informative, needs to be complemented with qualitative exploration in order to explore the reasons behind the results obtained.

Conclusion

Clinical performance assessments have an important position in the assessments in many medical speciality training programs, including this study context. While trainees largely report that they value the constructive or developmental assessment information they receive from them, they recognise assessor MUM behaviours are occurring and at higher levels than their assessors report. MUMing has the potential to mitigate the value of these assessments. A deeper understanding of the influence of the MUM effect in the delivery of assessment messages is needed in order that trainees receive the most valuable developmental information possible. Highlighting to assessors the lower than perhaps expected levels of trainee sensitivity to feedback may help to promote the delivery of more honest communication of results by them.

The next chapter will consider how the numerical ratings and written comments from a real-world sample of one assessment format (the SITA) compare, and whether there is suggestion that assessors differ in the assessment messages they deliver through the written word, and which they deliver by numbers.

CHAPTER 8

A review of the use of the summative in-training assessment (SITA)

Introduction

This chapter presents a review of the real-world use of the in-training assessment form used within the ACD, the SITA. Since some reports in the literature suggest that MUMing may affect the numerical ratings and comments parts of an assessment differently, this study aims to explore the relationship and degree of alignment of the message that assessors deliver in these parts through a review of a sample of SITA forms. The chapter reviews how the assessment form is used and then introduces the next part of the project to explore these and other findings further.

Background

As discussed in the Assessment Background introductory chapter, clinical performance assessments typically provide trainees with assessment messages in two formats: numerical ratings and written comments. While these convey different types of information (246), it would be expected that they will tell a similar story with respect to the trainee's performance. For example, in a study of ITAs submitted on a group of junior doctors, Ginsburg and colleagues found a strong concordance between the numerical scores and comments submitted (247). However, the literature also shows that such a relationship does not always occur and that in some cases, and often in relation to the poorest performers, the comments and ratings in an assessment may appear to send different messages. For example, some studies have shown instances where comments are more negative than the accompanying ratings (28,73,248). While not using MUM terminology, some authors have concluded that this difference may occur due to faculty reluctance to give negative assessment messages, for reasons including potential role conflict for the assessor (28). Such MUM avoidance has also been identified in the ACD previously, where the practice of assigning a "borderline" instead of the appropriate "unsatisfactory" rating and putting a negative comment in the relevant section instead, has been defended as "sometimes a useful strategy" (12).

The MUM effect literature shows that the consequences of a message may affect its delivery and so offer an explanation for these observed differences between numerical ratings and written comments. That is, the lack of direct consequences from comments may help to explain observations including that qualitative data can reveal greater

numbers of poor performers than the quantitative (249), and descriptions of performance are seen to be less susceptible to grade inflation (250).

However, the MUM effect may also predict the opposite effect: where distortion or sugarcoating lead to positive comments being recorded alongside more negative ratings. In a wider higher education setting, Mulliner and Tucker have described how both assessors and students “find it confusing to give/get positive feedback alongside a poor grade” (241). Further, as noted in the trainee questionnaire chapter, one trainee described receiving positive comments on their WBA, accompanied by a “not competent” rating. An assessor’s desire to deliver their message, but soften it by sugarcoating, may see them deliver appropriate negative ratings but accompany them with more positive “softening” comments.

As such, the relationship between the numerical ratings and the written comments and other features of the assessment is a key interest in this study. It is important to see if a discrepancy between comments and ratings occurs in the study context, and how other features of written comments such as word count, number of themes and suggestions for trainee development relate to the ratings assigned. Where discrepancies do occur due to MUMing, this would be important to identify both to further our understanding of how the MUM effect impacts assessment, and for the development of strategies to minimise the inclination of assessors to MUM.

Research questions

The aims of this chapter are to review the use of the SITA and explore the relationship between the written comments and numerical ratings. The guiding research questions for this chapter are:

To what extent do the numerical ratings and written comments in a sample real-world SITA summary forms align?; and

How do the characteristics of written comments, such as the word count, number of narrative units and suggestions relate to the ratings?

It should be re-emphasised at this point that the SITA involves more than the written document alone, and that a feedback exchange typically takes place between the SoT

and the trainee about the results. This study does not seek to explore this verbal exchange, rather looks at the written component of the assessment only.

Methods

The forms

A random sample of 158 (from a total of at least 382 [personal communication, Brett O’Neil, ACD]) anonymised SITA summary forms completed between 2014 and 2015 was supplied by the ACD. The staff who provided these collected the completed forms from both years for the sample, but did not use a particular random format. The SITA is the only clinical performance assessment tool that is submitted to the ACD hence the only one for which central records are kept. The WBA result only (that is when the trainee has been deemed competent at the particular assessment) is recorded by College, but the individual forms are not submitted and these remain with the trainee. In the case of the SITA, the summary form — which is prepared by the SoT using the SITA forms submitted by the CS — is submitted to the ACD after it has been discussed in a meeting with the trainee. This SITA form has been in use by the College since 2006, but logistical issues (difficulties accessing stored forms) and subsequent form modifications restricted review of the forms to 2014 and 2015. As previously described in Chapter 3, Research Methods, this paper-based form has two components — a numerical rating scale and a written comment (see Appendix E for the SITA summary forms). Assessors (usually three per trainee) independently and anonymously complete a SITA form on trainees on a six-monthly basis¹. The ratings component consists of two domains: clinical expertise (comprising 5 criteria) and professional qualities (comprising 20 criteria over 7 subdomains). The clinical expertise domain covers areas related to the trainee’s clinical skills and knowledge, while the professional qualities cover the trainee’s skills and knowledge in the areas of communication, quality and safety, cultural competency, leadership and management, health advocacy, teaching and learning, and ethics. Trainee performance in each criterion is rated on a Likert scale from 1–5 as follows:

¹ Some rotations at times have involved more frequent assessments, e.g. four monthly.

1 – unsatisfactory significantly below expectation for stage of training

2 – below expected standard for stage of training

3 – borderline requiring some development

4 – meets expectations for stage of training

5 – beyond expected standard for stage

N/A – insufficient opportunity to observe

Assessors may also enter free-text comments into two boxes — one entitled “Summary of the main strengths of this Trainee’s performance (based on the above ratings 4/5)” and the second “Summary of the Trainee’s areas most in need of further development (based on the above ratings 1-3)”. The completed forms are collated into a SITA summary form by the SoT and the trainee’s overall performance for the rotation is recorded as either “satisfactory” or “unsatisfactory” by checking the relevant box. The summary form is presented to the trainee during the meeting with the supervisor of training. A trainee is deemed unsatisfactory if they receive a rating of “2” or below from two or more assessors on two or more criteria, or if they have not completed the required number of WBAs commensurate with their year of training. This leads to the preparation of an SSP and a plan for increased supervision of the trainee as outlined in Chapter 5 (222).

Coding

Data from the SITA forms were entered into a spreadsheet in Microsoft Excel for Mac Version 15.30 (Microsoft Corp., Redmond, WA, USA). These data included the trainee year and state of training, number of assessors contributing to the form and the ratings data for each of the 25 criteria over 8 subdomains (1 clinical expertise and 7 professional qualities). The mean ratings for each domain (clinical expertise and professional qualities) and the mean total rating overall were calculated for each assessment. While the definition for unsatisfactory performance on the SITA is not directly related to the mean total rating, the mean total rating was used to summarise the overall valence of the ratings to allow for comparison with the written comments. Total word counts, the number of suggestions for development and the number of themes in

the comments, referred to here as narrative units, were also recorded (see box 8.1 for a summary of the information coded from the SITA summary form).

Data coded:

- Trainee demographics – year and state of training

Information from numerical ratings component of form:

- Ratings from 1–5 over 25 criteria

Information from written comment component of form:

- Word count in each area – Summary of main strengths, Summary of areas most in need of further development
- Number of suggestions for trainee development
- Number of narrative units in recorded comments

Box 8.1: Information coded from the SITA summary form

These aspects of the qualitative comments were coded to assess how they related to the mean total ratings. For example, the word counts were coded as some authors have reported that word count may show association with trainee level of performance (251).

Narrative units were determined by the candidate as follows: the text from each comment section was imported into a table in a Microsoft Word document for Mac Version 16.24 (Microsoft Corp., Redmond, WA, USA) and then separated into sections, where each represented a discrete theme or idea about the trainee (252). For example, the comments written on the summary form of:

- “A good 1st year registrar with good communication skills, good team member, caring, compassionate”

was separated into four narrative units:

- A good 1st year registrar
- Good communication skills
- Good team member
- Caring, compassionate

These are described as coding units as by Crano, Brewer and Lac (253) and comment units by Frohna and Stern (254). A method to compare the quantitative (ratings) and

the qualitative (comments) parts of the form was located in the literature (252). Here, Smither and Walker described identifying separate themes in the comments written on multirater feedback reports and assigning a rating from 1 (unfavourable) to 5 (favourable) to them (252).

Thus to facilitate comparison of comments and ratings in this study, a numerical score, termed the “favourableness score”, was assigned to each narrative unit according to its valence as follows:

1 = unfavourable

2 = slightly unfavourable

3 = neutral

4 = slightly favourable

5 = favourable

To increase the reliability of the process of assigning favourableness scores, two additional Fellows of the ACD with experience in the SITA format were purposefully recruited by a member of the PhD Advisory Panel (Associate Professor Sam Zagarella) to independently assign favourableness ratings to each narrative unit in addition to the candidate. Thus, in total, three scores per narrative unit were determined. Linear progression across the Likert scale was assumed and mean favourableness scores were calculated — following Smither and Walker (252). The mean favourableness score was used in the analyses. Cronbach alpha was calculated to measure the inter-rater reliability across all three reviewers who assigned favourableness scores to the comments. The Intraclass Correlation was 0.899 (95% CI = 0.766, 0.947).

During the coding process, it was observed that a printing error was present on 31 of the forms and the criterion for unsatisfactory performance was recorded incorrectly as being “A trainee is deemed unsatisfactory if they receive ‘3’ or below from two or more CSs on two or more criteria” instead of “if they receive ‘2’ or below”. Two of the forms had been manually changed to the correct criteria, but the remaining forms which displayed the incorrect criteria were excluded from this study as it was felt that this error could have impacted upon the assessor ratings. A separate review limited to the ratings

assigned on these excluded forms was conducted, with comparison made to the main sample, as it appeared that they showed some differences.

Data analysis

The SITA summary forms yielded quantitative and qualitative data related to trainee performance as described above. Quantitative data from Microsoft Excel were then imported into IBM SPSS Statistics Version 25 for statistical analysis. These data were analysed descriptively to show the distribution of forms across trainee year of training and home state, and the range and mean of variables including ratings, favourableness scores and features of comments. Independent samples t-tests were used to compare the mean ratings of junior (years 1 and 2) and senior (years 3 and 4 and IMG) trainees on the SITA for clinical expertise, professional qualities and mean total rating, as well as the mean favourableness score, word count, narrative units and suggestions for development. As outlined in Chapter 5 Methods section, whether results varied by trainee seniority was of interest in this study also. A Bonferroni correction was applied to a p-value of 0.05 to adjust for multiple comparisons, setting the threshold for statistical significance at 0.007.

Results

A total of 158 SITA summary forms were obtained from the ACD. Removing the forms with the incorrect criteria for “Unsatisfactory” left 129 forms available for analysis.

The 129 forms reviewed included assessments on trainees in all years of training (38% in year 1 (n = 49), 37% in year 2 (n = 48), 23% in year 3 (n = 29) and 2% in year 4 (n = 3). Most trainees were in full-time training (94%, n = 121) and trainees from all states were represented in similar proportions to the total number of trainees.

Assessments on trainees rotating to overseas posts were also included (UK and Singapore). Most assessments were completed by between 2 and 4 assessors, with a mean of 3.13 assessors per SITA summary form (with minor exceptions having more than 4 assessors).

Review of the use of the SITA summary form

Numerical ratings component of the SITA

Overall, the numerical ratings across the domains in the SITA were high. The average rating across all 129 SITAs for the clinical expertise domains was 4.19 and 4.31 for professional qualities domains, with a mean total rating across both of 4.25 (range 3.60 to 4.99) (descriptive statistics are presented in table 8.1).

Table 8.1: Summary of information recorded on the SITA summary form overall and by year of training (ratings and favourableness scores)

		Overall	Year 1	Year 2	Year 3	Year 4
Clinical expertise rating	mean range	4.19 3.31–5.00	4.21	4.22	4.17	3.66
Professional qualities rating	mean range	4.31 3.71–4.98	4.31	4.35	4.29	3.95
Mean total rating	mean range	4.25 3.60–4.99	4.26	4.28	4.23	3.80
Favourableness score	mean range	3.97 2.08–5.00	4.05	4.00	3.88	2.98

A total of 23 (17.8%) forms had a mean total rating of less than 4. A very small number of forms recorded any ratings indicating performance below the expected standard: 2 ratings of “1” and 23 ratings of “2” were recorded (25 out of the total of 9975 or 0.25%). These were in the areas of clinical expertise (2 ratings of “1” and 10 ratings of “2”), leadership and management (7 ratings of “2”), communication (3 ratings of “2”), teaching and learning (2 ratings of “2”) and quality and safety (1 rating of “2”). No ratings of “1” or “2” were given for the domains of cultural competency, health advocacy or ethics (table 8.2). Seven trainees received these ratings of “1” or “2” across 8 SITAs. Three trainees received 2 ratings of “2” across 3 SITAs (one received both “2”s for clinical expertise and two received one each for clinical expertise and teaching and learning), and three trainees received 1 rating of “2” across 4 SITAs (one trainee received this on two separate SITAs). These were for the domains of clinical expertise, communication and teaching and learning (2 separate SITAs). One trainee received 2 ratings of “1” (both in clinical expertise) and 13 ratings of “2” across the domains of clinical expertise, communication, quality and safety and leadership and management. This was the only overall unsatisfactory SITA in the analysis group, in

accordance with the criteria of a “2” or below from two or more CSs on two or more criteria. The mean total rating for this form was 3.61 and the mean favourableness score was 3.85. Neither box where the overall performance of the trainee was to be determined as “satisfactory” or “unsatisfactory” was ticked. A total of 348 ratings of “3” were recorded across all forms.

Table 8.2: Number of ratings of “1” or “2” across reviewed forms

Domain	Rating of “1” or “2”	No. trainees
Clinical Expertise	12	5
Leadership and Management	7	2 over 3 SITAs
Communication	3	2
Teaching and Learning	2	2
Quality and Safety	1	1
Cultural Competency	0	0
Health Advocacy	0	0
Ethics	0	0
TOTAL	25	7

Written comments component of the SITA form

Several other aspects of the qualitative comments, including the number of words, narrative units and suggestions for development were noted to allow for their association with the mean total ratings to be explored (table 8.3). There was an average of 40.84 words per SITA recorded in the comments section, 22.15 in the strengths area and 18.70 in the areas for development area. The minimum word count in the strengths area was 2 (in 1.6%, n = 2 of SITA forms) and the maximum 96 (in 0.8%, n = 1 of SITA forms). The minimum word count in the areas for development area was 0 (in 7.0%, n = 9 of SITA forms) and the maximum was 123 words (in 0.8%, n = 1 of forms). A total of 26 forms had either “nil” or “none” (n = 25, 19.4%) or a comment suggesting the trainee had nil areas for development (n = 1, 0.8%) entered in this section. This means 35 (27.1%) forms had no written message about areas for development.

There was a mean total of 6.43 narrative units per SITA, 4.58 recorded in the strengths area (minimum of 1 in 6.2%, n = 8 of SITA forms, maximum of 11 in 0.8%, n = 1 of SITA forms) and 1.84 recorded in the areas for development area (minimum of 0 in 7.0%, n = 9 of SITA forms, maximum of 6 in 0.8%, n = 1 of SITA forms). The mean

number of suggestions for development averaged 1.45 per form overall (range 0–5). There were no suggestions for development on 21 forms² (16.2%) through to the maximum reported of 5 on one form (0.8% of forms) (descriptive statistics are presented in table 8.3).

Table 8.3: Summary of information recorded on the SITA summary form overall and by year of training (characteristics of comments).

		Overall	Year 1	Year 2	Year 3	Year 4
Word count/SITA	mean range	40.84 3–219	33.22	40.29	52.28	63.67
Word count – strengths	mean range	22.15 2–96	19.96	22.54	25.21	22.00
Word count – areas for development	mean range	18.70 0–123	13.27	17.75	27.07	41.67
Narrative unit count/SITA	mean range	6.43 2–11	6.73	6.15	6.52	5.00
Narrative units – strengths	mean range	4.58 1–11	5.02	4.33	4.41	3.00
Narrative units – areas for development	mean range	1.84 0–6	1.71	1.81	2.10	2.00
Number of suggestions for development	mean range	1.45 0–5	1.33	1.47	1.66	1.44

In total, 196 of the 831 (23.6%) narrative units recorded on the reviewed forms were assigned a mean favourableness score across the three evaluators of less than or equal to three. The mean favourableness score per SITA was 3.97 (range 2.08–5.00) (table 8.1) and a total of 10 (7.8%) SITA forms had an overall favourableness score of less than or equal to 3. Of the 10 forms with a score of less than or equal to 3, three had a mean total rating over 4. Seven forms (5.4%) had both favourableness scores and mean total ratings of less than 3 and 4 respectively. Table 8.4 shows that across three different ranges of favourableness scores, the range of mean total ratings was highly similar.

² Suggestions for development were not always recorded in the areas for development section of the form.

Table 8.4: Mean total rating ranges across a range of favourableness score ranges.

Written comments	Numeric ratings
Favourableness score range	Mean total range
2.0–3.0	3.60–4.26
3.1–4.0	3.61–4.89
4.1–5.0	3.80–4.99

A wide range in favourableness scores was also evident for SITAs which recorded at least one rating of 1 or 2. Table 8.5, “SITAs with at least one rating of ‘1’ or ‘2’”, shows that these eight forms had favourableness scores ranging from 2.38 to 4.11, with two having scores of less than or equal to 3 and six having scores above 3. Mean total ratings for these forms ranged from 3.61 to 4.34. Table 8.5 also shows a wide range in the mean total ratings for the eight trainees who had favourableness scores of 3 or below but did not have any ratings of 1 or 2. The mean total ratings for this group ranged from 3.60 to 4.26. In comparison, SITAs which did not record any ratings of 1 or 2 (n = 121) had favourableness scores ranging from 2.08 to 5.00 and mean total ratings ranging from 3.96 to 4.99.

Table 8.5: Favourableness scores and mean total ratings for SITAs where 1. At least one rating of “1” or “2” was recorded and 2. No ratings of “1” or “2” were recorded (favourableness score ≤ 3)

SITAs with at least one rating of “1” or “2”		SITAs with no ratings of “1” or “2” and favourableness score ≤ 3	
Favourableness score	Mean total rating	Favourableness score	Mean total rating
2.38	3.71	2.08	3.70
2.89	3.77	2.20	3.94
3.19	3.64	2.40	4.26
3.70	3.70	2.60	3.83
3.85*	3.61*	2.64	3.72
3.92	3.99	2.83	3.60
3.96	4.34	2.93	4.26
4.11	3.80	3.00	4.09

*-indicates results from the unsatisfactory SITA

In order to explore if trainee seniority impacted on SITA results, results from junior trainees (years 1 or 2) were compared with those of the senior trainees (years 3 or 4). Levene’s test was used to test for the equality of variances for each of the comparisons.

Equal variances were assumed in all but two comparisons and in these cases the statistics for unequal variances are reported. Independent samples t-tests showed there were no statistically significant differences between the groups (table 8.6).

Table 8.6: Comparison of SITA variables between junior and senior trainees – independent samples t-tests

Variable	Junior trainee (n = 97)	Senior trainee (n = 32)	Mean difference (Junior – Senior) estimate	t-value	95% CI	P-value
Clinical expertise rating	4.21	4.12	.089	1.199	–.061, .239	.237
Professional qualities rating	4.33	4.26	.070	1.188	–.048, .189	.241
Mean total rating	4.27	4.19	.080	1.242	–.049, .209	.221
Favourableness score	4.03	3.80	.230	1.756	–.033, .492	.085
Word count – strengths	21.24	24.91	–3.699	–1.084	–10.494, 3.156	.284
Word count – areas for development	15.48	28.44	–12.953	–2.448	–23.668, –2.238	.019
Total word count	36.72	53.34	–16.622	–2.173	–30.35, –2.38	.036
Narrative units – strengths	4.68	4.28	.399	1.086	–.336, 1.135	.282
Narrative units – areas for development	1.76	2.09	–.331	–1.278	–.851, .190	.207
Total narrative unit count	6.44	6.38	.068	.143	–.890, 1.027	.887
Suggestions for development	1.40	1.64	–.240	–1.004	–.720, .240	.320

Comparison of the numerical ratings and written comments for each SITA

A scatter plot of mean total ratings against favourableness scores is presented in figure 8.1. This shows a modest trend for higher mean total ratings to be associated with higher favourableness scores, though there is considerable variation in favourableness scores for each mean total rating score. Area A on the plot shows a small number of cases (n = 3) where the mean total rating was more positive (> 4) than the accompanying comments (favourableness score less than 3), and area B shows 16 cases where the mean total rating was less positive (< 4) than the accompanying comments (favourableness score greater than 3). Results from the unsatisfactory SITA fall in this area.

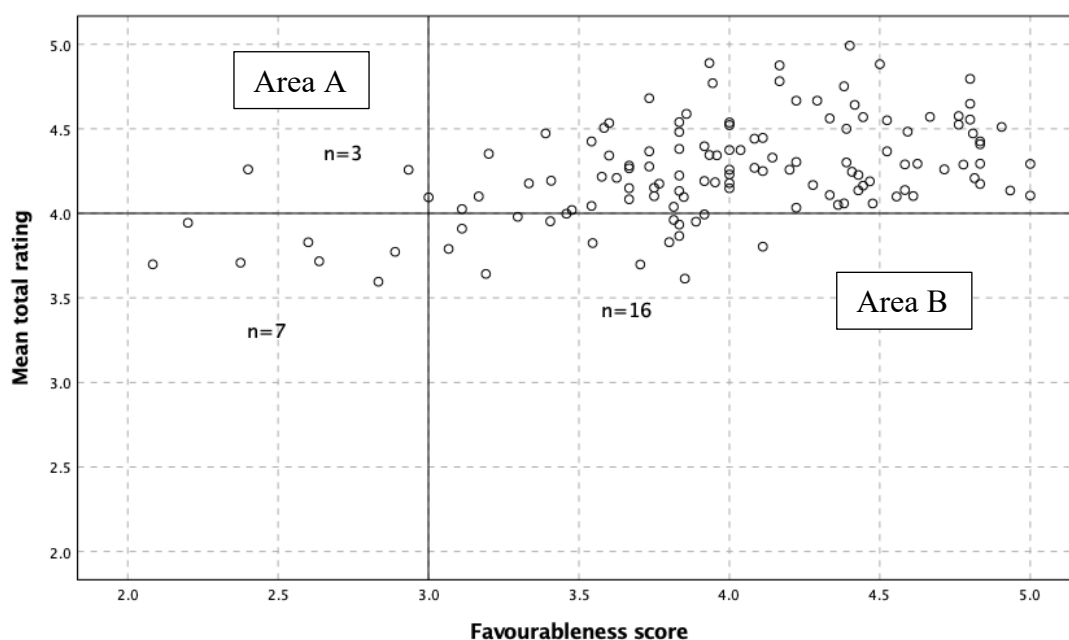


Figure 8.1: Mean total rating compared with favourableness score

Linear regression models were used to explore the relationship between the mean total numeric rating and variables related to the qualitative comments. A simple linear regression model showed mean total ratings and favourableness scores were moderately positively related with a Pearson correlation of 0.492 (0.459 for Spearman's rho) (table 8.7). The small effect was likely partly due to the positively skewed nature of the results. The favourableness score contributed 23.0% to the variance of the mean total rating (figure 8.2).

Table 8.7: Pearson correlations between mean numeric rating and characteristics of written comments.

	Mean total numeric rating	Sig. (2-tailed)
Favourableness score	.492**	.000
Total word count	-.449**	.000
Total narrative units	-.235**	.007
Number of suggestions for development	-.426**	.000

** correlation is significant at the 0.01 level (2-tailed)

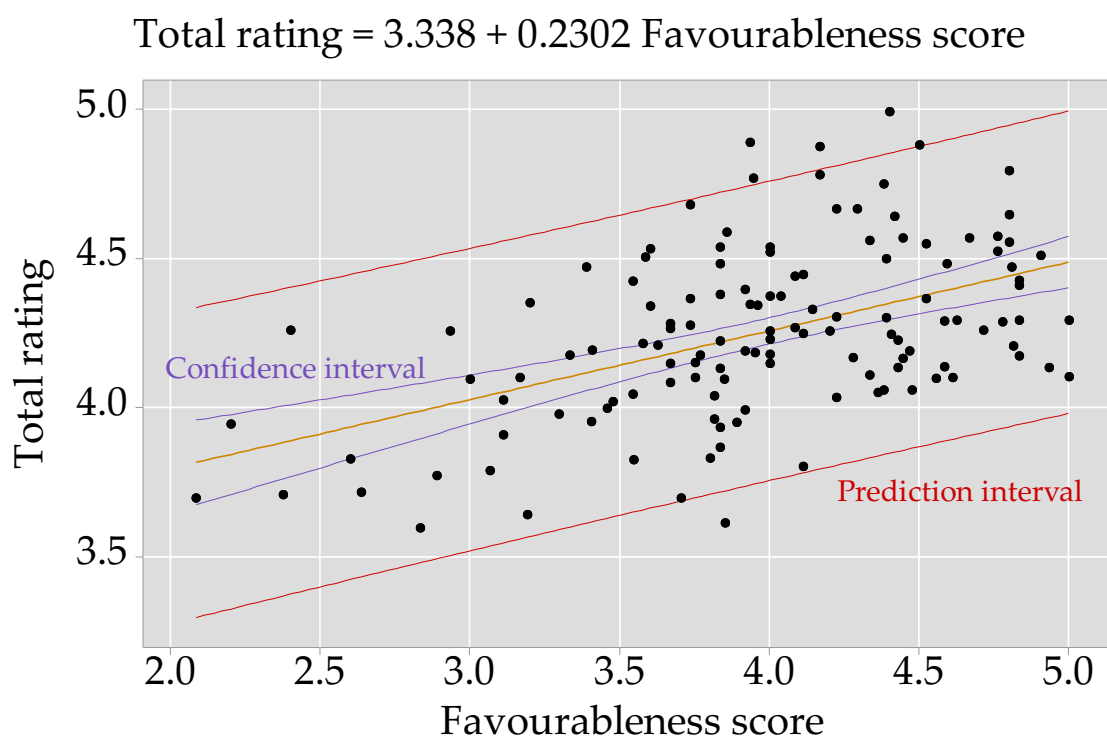


Figure 8.2: Scatter plot of mean total rating compared with favourableness score

The relationship between favourableness score and mean total numeric rating for the poorer performing trainees was not more diverse than for other trainees, though there were fewer data points at this end of the scale. The range of favourableness scores were much greater than the range of mean total ratings and a wide prediction interval was noted across all the results. In summary, the results show that the comments recorded on the SITA forms correlate with but only partly explain the variation in mean total ratings.

Other variables contributed very small amounts only to the variation in the mean total rating and the scatter plots below show the comparison of total word count (figure 8.3) and narrative unit count to the mean total rating (figure 8.4). These show that as the word count increases the mean total rating tends to decrease and as the number of narrative units increases the mean total rating tends to decrease. These data indicate that there is a trend for examiners to write longer qualitative comments (with a greater number of narrative units) for candidates who receive lower overall ratings.

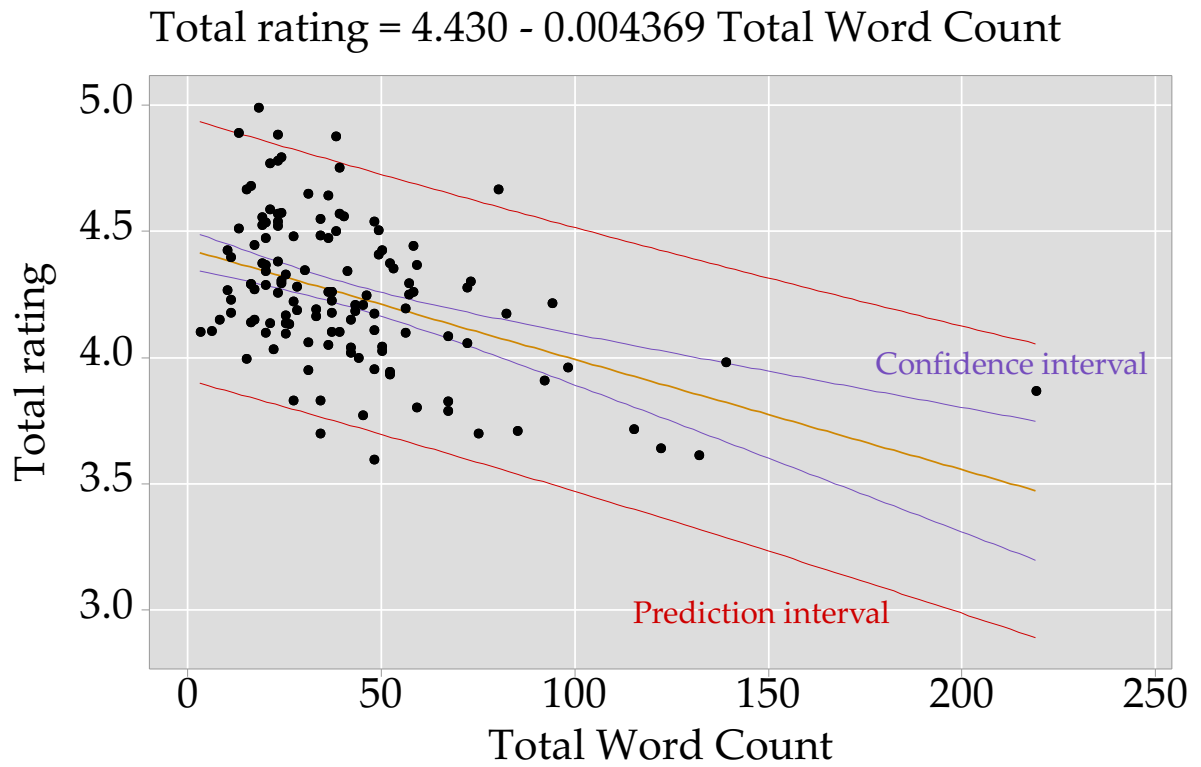


Figure 8.3: Scatter plot of mean total rating compared with total word count

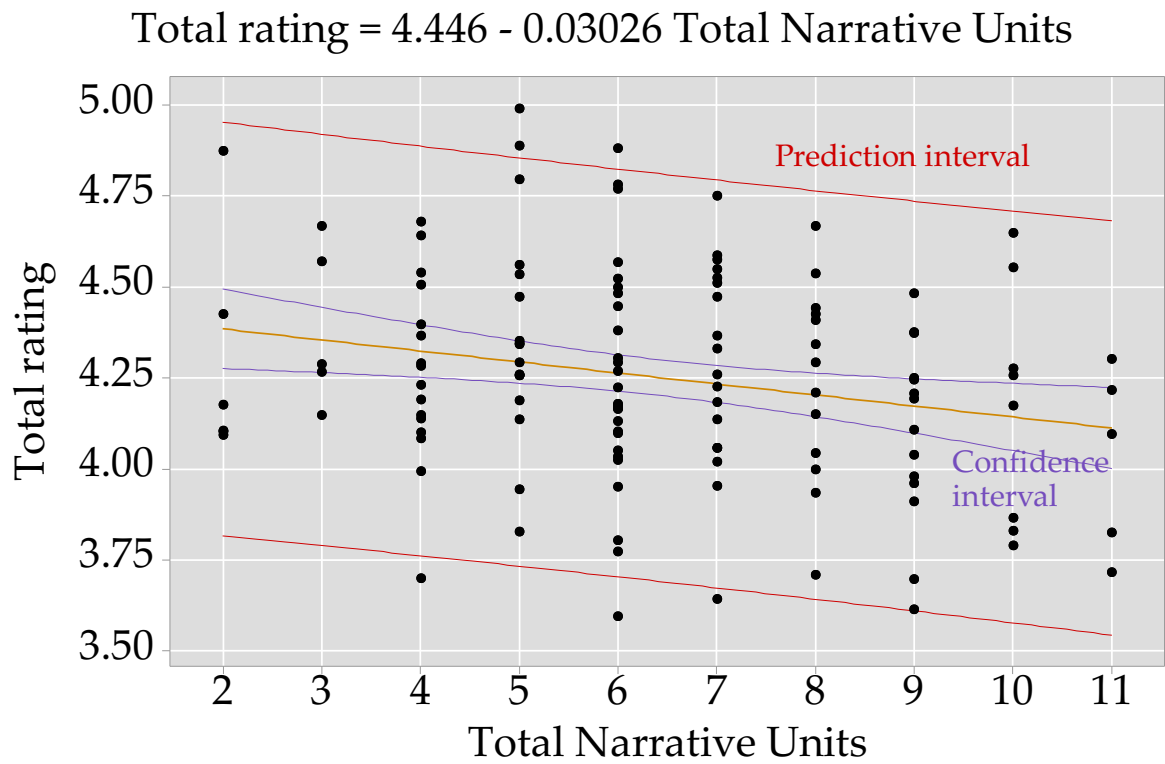


Figure 8.4: Scatter plot of mean total rating compared with total number of narrative units

Review of forms displaying incorrect criteria for unsatisfactory performance

A review of the 29 forms where a higher criteria for unsatisfactory performance was incorrectly printed (referred to as the “incorrect group” from here) suggested there were some differences to the main group, so comparisons between the two groups were made. For assessments in the incorrect group, 24 recorded no borderline ratings, and one each recorded 1, 2, 3, 4 and 5 borderlines giving an average of 0.52 borderline ratings per form. Forms in this group had a mean total rating of 4.30 and a mean favourableness score of 4.24. In comparison, forms where the criterion for unsatisfactory was correctly printed had an average of 2.70 borderlines per form (range 0 to 113), a mean total rating of 4.25 and a mean favourableness score of 3.97. When the incorrect and main groups were compared using independent samples t-tests with equal variances assumed, the mean number of borderline ratings in the incorrect group was less than the mean number in the main group (mean difference 2.18, 95% CI .23, 4.13, $p = 0.028$). When an outlier assessment form with a large number (19) of assessors (which had 113 borderline ratings and was the only failed SITA in the group) was excluded, a difference in mean number of borderline ratings between the two groups was still observed (table 8.8).

Table 8.8: Comparison of mean number of borderline ratings, mean total ratings and mean favourableness scores between main and incorrect groups

Variable	Criteria for unsatisfactory = 2* (n = 129)	Criteria for unsatisfactory = 3** (n = 29)	Mean difference	95% CI	P-value
Excluding failing SITA form:	(n = 128)	(n = 29)			
Mean number of borderline ratings recorded per form	1.84	.52	1.32	.37, 2.27	.007
Mean total rating	4.26	4.30	-.04	-.15, .07	.446
Mean favourableness score	3.97	4.24	-.27	-.47, -.08	.007

* two ratings of “2” or below from two more supervisors on two or more criteria

** two ratings of “3” or below from two more supervisors on two or more criteria

Discussion

Summary and interpretation of findings

This study aimed to explore the real-world use of the in-training assessment used in the ACD through a review of previously submitted forms, and to investigate the relationship between the assessment messages conveyed through the written comments section with those in the numerical ratings section.

Review of the use of the SITA summary form

The review of the use of the form shows that on average, trainees were rated as having performed above the standard expected and this was seen across both the clinical expertise and professionalism domains. Further, most individual ratings were positive, with comments showing more variability than ratings. The unsatisfactory performance rate of 1 in 129 (0.8% of assessments) seen here is in keeping with rates in recent years (personal communication June 2019, Brett O'Neill, ACD). It was noted that on this unsatisfactory assessment form, the box indicating the trainee's overall performance in the rotation (i.e. either "satisfactory" or "unsatisfactory") was left blank, perhaps also suggesting MUM avoidance.

Overall, the total word count was approximately evenly distributed between the area for strengths and the area for development, but the number of themes these words related to differed with more themes conveyed in the strengths area using fewer words as compared with the areas for development. Most narrative comments recorded on the assessment forms, as judged independently by three assessors, were positive. In sum, results at the lower end of both the numeric and written scales were uncommon.

Overall, relatively few developmental comments were seen on the forms in this study. Over one quarter of assessments contained either no comments in the areas for development section, or the comments indicated that the trainee had no areas for development. When the results were separated and trainees compared by level of seniority, no significant differences were noted.

Comparison of the numerical ratings and written comments

The results showed a small but statistically significant association between the mean total rating and the favourableness score, suggesting that the written comments and numerical ratings assessors assign were generally aligned. While both the overall mean total rating and the mean favourableness scores for the sample were higher than the midpoint of their respective scales, more individual forms showed a mean total rating of less than four, than a mean favourableness score of less than or equal to three. The only unsatisfactory SITA showed a mean total rating of less than four but a favourableness score of over three. It is important to recall that the definition for unsatisfactory performance on the SITA is not directly related to the mean total rating but that this calculation was used to summarise the valence of the ratings overall for comparison with the comments. Ratings which may potentially lead to a consequence (a “1” or a “2”) were rare, although individual unfavourable comments were more common than negative ratings by a factor of approximately 100. The exceptional use of “1” or “2” ratings compared to corresponding written comments may be explained by the MUM effect: assessors may have felt less reluctant to deliver a negative message through comments, which had no immediate summative impact, rather than through numerical ratings, which in the ACD context, may carry significant consequences including impact on trainee progress.

Implications

This was an exploratory study which sought to review and document the use of an in-training assessment format in actual use and examine the relationship between the numerical ratings and the written comments sections to look for discrepancies as possible further indicators of MUM behaviours by assessors. This study did show an association between the mean total rating and the favourableness score, though the small size of this was likely due in part to the positively skewed nature of the results. This means that overall, there is not a strong difference between numerical ratings and written comments in this context, so a focus on improving this relationship is not likely to be key in addressing the issues of MUMing in this context.

When forms were considered on an individual basis, some (almost 15% of forms) showed a misalignment between written comments and numerical ratings, and contrary

to some findings in the literature (28,73,248,255), this was mostly in the direction of comments being more favourable than ratings. While these isolated examples may represent distortion (sugarcoating) of results, this is not possible to determine as an attempt to match comments to a domain was not made in this study, however this has been done in some other studies (248,254). As such, it is not known whether written comments and numerical ratings addressing different domains was an explanation for any lack of concordance in this study as has been reported elsewhere (246,248,249,254,256). Another possible reasons for a mismatch between numerical ratings and written comments is the effort that is required to construct comments. Discussing a similar performance appraisal system in business settings, Brutus emphasised the “profound differences between agreeing to a performance-related statement by ticking a box and actually writing a statement about someone’s performance” (257). In addition to the cognitive challenges that writing comments presents, motivational ones are also important and the impact of motivation on behaviour is well described, both in appraisal (112) and other settings (258). When an assessor has positive comments to make about the performance of a trainee, the task may be easy and motivation levels high. However, when the message is negative, assessors may find it more challenging to deliver and potentially lead to MUMing and a possible discrepancy between numerical ratings and written comments. A culture where giving written feedback is not the norm, combined with a lack of training in how to provide comments (259), may thus combine to limit the written comments on performance given and so potentially result in a paucity or even absence of comments that then appear out of alignment with ratings.

In this work, all parts of the assessment form were coded to see their relationship to the other variables, but again few practical associations were seen. For example, while the word count was seen here to increase with decreasing mean total rating, this differs from some reports in the literature. For example, Ginsburg and colleagues found longer comments (more words) were seen in cases of both excellent and struggling trainees (251). They noted the possible implications of this — for example, some participants in their interview study suggested the length of written comments reflected the credibility of the comments, the amount of effort put into them or how well the assessor knew the trainee (251), showing that the length of comments is not solely related to the performance standard of the trainee. In the current study, while increasing word count

was associated with lower ratings, it was not able to be determined whether these comments were as similarly negative as ratings or whether they were more positive and perhaps in keeping with distortion (sugarcoating).

The paucity of developmental comments noted in the current study limits the benefit of the assessment for the trainee, while others have noted that comments may be limited in their value. For example, a participant in Sabey and Harris' study commented that "... 'Needs more clinical experience' is generally what everyone writes" (189). For forms where no comments were recorded, it seems surprising that a trainee in a training program would have no areas to work on, so assessors keeping MUM in this area may be a more likely explanation. However the form design itself may also have contributed to this outcome, as it instructed assessors to provide a "Summary of the Trainee's areas most in need of further development (based on the above ratings 1-3)". Some assessors recorded a comment noting that the trainee had no ratings below 4, and recorded no further comment, however others did write comments on the trainee's areas for development, even if they did not have ratings below 4. When the limited number of constructive comments is considered with the relatively few suggestions for development recorded on forms, at approximately one and a half per form on average, there is little documented for trainees to work on. While this study did not attempt to account for the verbal interaction that may have occurred in the SITA interview, it is salient to recall that, as reported in Chapter 7, Trainee Questionnaire, 16% of trainees agreed they can't remember what assessors want them to do when they provide feedback, and almost 10% agreed that they miss out on the details of what assessors want when they provide them with feedback, emphasising the importance of documenting feedback.

While the low failure rate observed in this study may reflect the high standard of trainees and the training program, it also needs to be considered in light of both the self-reports of failure to fail in the SITA presented in Chapter 6, Assessor Questionnaire, and the views on the occurrence of MUMing expressed by both assessors and trainees. While the failure rate in this study is in keeping with that of recent years in the ACD, the literature also supports that this figure is an under-representation. For example, while Kendal and colleagues noted in their review of 50 consecutive ITAs on their medical specialist trainees that performance was rarely documented as "unacceptable"

or “in need of improvement”, they concluded their grading scale to be “inefficient in identifying residents in need of help”, implying it underestimated failure rates (73). Similarly, on their review of end of rotation assessments for prevocational trainees, Bingham and colleagues reported low levels of underperformance or borderline performance and quote evidence to suggest that the real rate is “much higher” (22), and in the medical student setting, Wilkinson and colleagues found prior to their intervention, only 1.1% of students were identified as being in need of assistance, with many more identified subsequently (61). Thus it is very likely that the low failure rate in the study setting represents an underestimate of the true rate, and so failure to fail.

The importance of the instruction documented on the assessment forms is also demonstrated from the observation of the effect that the incorrect criteria for unsatisfactory performance had. Though small in number, those summary forms where the criteria for unsatisfactory was incorrectly higher showed fewer borderline ratings. A difference in the comparison of favourableness score to mean total rating was also seen, however, this was in the opposite direction than perhaps would be expected: forms where the criteria for unsatisfactory performance was higher (two ratings of “3”), had more favourable comments on average than when the criteria was lower (two ratings of “2”). In the context of high-stakes judgements in an environment where MUM behaviour has been shown to exist, this raises important questions about whether the different criteria could have altered assessor behaviour.

Future directions

This chapter has shown the presence of largely high ratings and positive comments, and a significant relationship between the comments and ratings from a sample of ITA forms.

For the minority of forms where the ratings and written comments did not appear to align with respect to the valence of the message that they sent to trainees, the potential for confusing or missing the message exists. Many trainees report they value and desire narrative feedback in performance assessments (102) and some find it more useful than ratings (38,188). As highlighted in the systematic review in chapter 4, written comments are frequently reported by trainees as being very important, if not the most important part of their clinical performance assessments — some stating “the written

comments... are the most helpful because you get a sense of what the staff person was actually thinking as opposed to just the tick boxes ..." (188), and some in fact would prefer to avoid ratings altogether (7). The strength of such comments is highlighted in one pilot study of an MSF form, where it was noted that the comments could show improvement in trainee performance over time, while ratings showed no significant change (260). Where ratings and comments don't align, it is unknown if all trainees have the ability to "crack the code" behind what is written (38). So the potential exists for them to miss the important assessment message, should they "overlook or misinterpret assessors' intended meanings" (259), especially if they are too positive (136). As seen in these results, most of the trainees who received a mean total rating of lower than 4 still received favourable comments, which may lead to confusion. The situation highlighted by the trainee in chapter 7, who queried their favourable comments accompanied an unsatisfactory numerical result, echoes other reports in the literature, whereby some "may read comments that seem positive ... and not understand why their grades do not directly reflect what appeared to be positive comments" (211). While as presented in the MUM literature review chapter, experience from business settings suggests that some amount of sugarcoating may help promote an improvement in performance (136), it may also risk trainees not fully understanding their developmental needs or disregarding them if they are contradicted by the ratings. If MUM avoidance leads to accurate written comments but inappropriately high ratings, delay or denial of remediation to struggling trainees may result (28); for example, Kendal et al. have reported feeling "unable to act upon verbal comments alone" (73). And even if accurate verbal information is given in the feedback exchange, some trainees may struggle to recall the detail if it is not documented.

The MUM effect may offer both a new explanation for cases where a difference between numerical ratings and written comments are seen and so the potential for development of new approaches to address the issue. Of course, assessors may also avoid communicating their opinion on a trainee's performance for reasons other than a reluctance to do so. Furthermore, the presence of an association between the written comments and numerical ratings in an assessment does not of itself necessarily indicate an absence of the MUM effect, as assessors may keep mum in both parts of the assessment and either deliver their assessment message verbally or not at all. The results from Chapter 6, Assessor Questionnaire, showed that where assessors admitted

to grade inflation or failure to fail on the SITA, in most cases they did not document any comments reflecting this, which would not result in a misalignment with ratings. It is very important that assessors understand the implications of the words and numbers they choose to combine, and are provided with support in how to provide meaningful narrative comments (259). While this study showed that numerical ratings and written comments were aligned in the sample reviewed, other information suggests the situation is not simple and further exploration is required.

Limitations

The study design meant that answering the main research question for this study in more than general terms proved to be challenging due to several factors. Firstly, the mean total rating calculated is not directly related to the outcome (overall satisfactory or unsatisfactory performance) of the assessment, thus a SITA could have a low mean total rating, but still be regarded as satisfactory. For example, a trainee may score several ratings of “1” or “2” and so a low mean total rating, but unless these are distributed in accordance with the definition of unsatisfactory, the assessment result would be deemed satisfactory. The mean total rating is therefore a somewhat blunt measurement of trainee performance and also obscures the fine detail present in each form, such as the variations in trainee performance between the different domains. However, given the rarity of unsatisfactory outcomes, it was not possible to compare comments on unsatisfactory compared with satisfactory SITAs in a meaningful way, thus the mean total rating was employed as a summary of performance. Further, the subjective nature of determining the favourableness scores and the fact that the midpoint of each scale differed were further factors which made a direct comparison of comments to ratings difficult.

Further, as outlined above, the observed statistically significant relationship between numerical ratings and written comments does not exclude the presence of the MUM effect, if MUMing affected both components of the assessment form and the results from the assessor questionnaire study support that this could be an explanation (whereby some assessors admitted to grade inflation or failure to fail and either took no further action or gave only verbal comment only, which would not be detected in this study).

As this review considered submitted assessment forms only, and as highlighted by Liao and colleagues, analysis of the written comments only does not account for the total feedback encounter, and may therefore underestimate the amount of feedback given to the trainee (52). Logistical issues with access to SITA forms prevented the examination of a broader range of assessments which may have influenced the results. Similarly, there were smaller numbers of senior trainee SITA forms supplied, which was beyond the candidate's control.

The allocation of favourableness score to comments was a subjective process and limitations exist in relation to this. As Ginsburg and colleagues have noted "it is not always easy to decide whether a comment is positive, negative, or somewhere in between" (247), and other authors have used a different method to the one used here [for example (211,254)], and so results may have differed with use of an alternate technique. While three evaluators were used to determine scores and the intraclass correlation was strong, it is possible that different evaluators would have come to different conclusions and therefore different results.

Conclusion

Clinical performance assessments are potentially valuable sources of assessment messages for trainees and can provide numerical as well as written and verbal commentary on their performance. In this study, overall the numerical ratings and written comments on the SITA summary form were seen to be statistically associated. However, differences were seen in some forms and as the self-reports of assessors and trainees in the previous studies in chapters 6 and 7 have indicated that MUMing occurs in the study context, further information is needed to understand this more. As such, the next chapter presents an interview study where assessors describe their experiences with the clinical performance assessments of dermatology trainees to further extend and explain these findings and those of the previous chapters in greater detail.

CHAPTER 9

We probably don't fill them out truthfully enough:

Assessors' experiences with clinical performance assessments

Introduction

While the preceding chapters have described and captured some of the issues with clinical performance assessments in a quantitative manner, this chapter presents a qualitative study. This seeks to explore the assessors' personal experiences with the assessments and in particular with the delivery of negative assessment messages and as such, present additional insights to the issues. This chapter therefore offers new and expanded perspectives from the data already collected, which in the discussion chapter which follows, will be synthesised in order to answer the research question.

Background

The preceding chapters have revealed much about the use of the clinical performance assessments in the study context. The self-reports of assessors and trainees presented in chapters 6 and 7 have shown that MUMing does occur, though trainees perceive it occurs more than assessors report. The MUM literature allows us to now expand our understanding and views of behaviours associated with difficulty with delivery of negative assessment messages beyond those of failure to fail and grade inflation, to include delay, avoidance and distortion of messages also. Assessor discomfort with the assessment process was found to be a relevant contributing factor in some cases. When the numerical ratings and the written comments on a sample of previously submitted assessment forms were compared to see whether MUMing affected these parts of the assessment form differently, the two were seen to be relatively aligned, however, there are indications that this is not the full story. The preceding chapters have approached the overall research question from a largely quantitative perspective and while this is of importance in documenting and quantifying aspects of the issue, and so for example allowing the measurement of changes over time or comparisons between settings, it also poses limits on the information collected. As quantitative methods are often aligned with a positivistic epistemology, and the view that a single truth awaits discovery and is independent of the researcher, these methods search largely for observable and measurable data. Considering failure to fail and related issues from a purely quantitative perspective is limiting and as the incidence of failure to fail is low, may risk it being considered insignificant when viewed from a purely quantitative stance. In contrast, a qualitative approach permits the exploration of the subjective and lived experiences of the individuals involved, and so the construction of a different type of

knowledge (217) and arguably a deeper understanding of the issues. The different perspectives obtained may permit the development of new insights and ideas about the phenomenon: illumination rather than direction (190). Thus it is necessary and consistent with the theoretical framework of this study to employ additional methods to explore this complex and persistent issue.

Many options for collecting qualitative data are available; for example, many studies in this area have employed interviews or focus groups [for example (10,11,261)] to explore issues around the difficulties of delivering negative assessment results. Various types of analyses have also been employed; for example, those relating to the content of data and what was said (10) compared with processes and how it was said (262). In this study, interviews were chosen to explore the issues with assessors of the ACD and develop a deeper understanding of the issues facing assessors in relation to clinical performance assessments. The research questions for this component of the study were:

How do assessors' experiences and perspectives on the clinical performance assessments inform their delivery of negative assessment messages?

How do assessors manage their delivery of negative assessment messages, in particular:

What facilitates delivering their message; what hinders it?

How do assessors deliver their negative assessment messages?

What do assessors report that could assist them in their assessment duties?

Methods

Qualitative approach and research paradigm

This chapter is prepared with reference to the Standards for Reporting Qualitative Research (SRQR) (263), with the relevant parts presented in each section. Those relevant to the methods are presented here, with the exception of 9, Ethical issues; 20, Conflicts of interest; and 21, Funding, which are outlined in Chapter 5, Research Methods. As described, while the research project overall is situated in a pragmatic paradigm, this part of the project uses a constructivist/interpretivist approach to seek to answer the relevant research question.

Research characteristics and reflexivity

As the candidate and interviewer, it is important to acknowledge aspects of my background that may influence this research. As a dermatologist, a previous trainee of the ACD and a previous CS and SoT at three health services, I bring many experiences, perspectives and beliefs to this project. I work alongside, but not in the same clinical areas as the Victorian participants in this study and have served on committees with four of the participants. I have presented my emerging research at the Annual Scientific Meeting of the ACD and at Supervisor Training sessions, which some of the interview participants may have attended. As described in Chapter 1, my own experiences with the clinical performance assessments in part led me to pursue this research project from two main perspectives. The first was that while an SoT, I observed that at times, the assessment message documented on the assessment forms differed from what the CSs would express verbally about a trainee's performance. The second was the recognition of my own experiences of the delivery of assessment messages and how these differed depending on how well or poorly the trainee had performed. In order to recognise the impact of my personal views on the analysis and synthesis of the data, I have kept a journal to document and explore my observations. I have also ensured that at all stages of the project, including data collection and analysis, discussions were held regularly with my supervisors to challenge and question the findings and conclusions I formed, and ensure that I was open to exploring different and contrasting interpretations and meanings within the data.

Context, sampling strategy and participants

Participants for the study were recruited via advertisement in the weekly email newsletter distributed by ACD as per the Ethics application. All interviewees self-identified as assessors of the ACD (either a CS or an SoT), with recent (within the preceding 12 months) experience with assessments. This recency was required in order to ensure currency to the data collected. There was no limit placed upon the number of interviewees and eight participants had volunteered after the advertising period had concluded. All who volunteered were interviewed and no incentives were provided for participation. To facilitate participation (for example, as some participants had child-care responsibilities and others were in states other than Victoria, where the candidate was located), most interviews were conducted by phone. As such, of the eight

interviews, seven were conducted by phone and one in person. This indirect format may have facilitated the disclosure of sensitive information, though in comparison, the face-to-face format did not appear to inhibit the sharing of information. Interviews were scheduled according to the preference of the participant, and six of the eight interviews were on separate days (one interview was scheduled immediately after a preceding one). All interviews were recorded with a handheld dictaphone with a microphone adapted for phone use for phone interviews (Olympus DM-720). The candidate also took notes during the interviews. The plain language statement and consent forms were distributed to participants prior to the interviews and these were completed prior to or around the time of the interview and stored electronically in protected files. All interviews were downloaded to the candidate's password-protected computer and stored in protected files. Prior to the interviews commencing, the candidate interviewer gave a brief introduction to the participant including a reminder that the interview would be recorded and transcribed. In some cases, a voice check for the transcriber was conducted.

Data collection methods, collection instruments and technologies

The overall research question for the project guided this part of the work and the specific questions are detailed in the Background section of this chapter.

The interviews were conducted between October and December 2018 and were semi-structured in design (264). Extensive reference to methods sections of articles and texts [in particular (217)] was made prior to conducting the interviews. An interview schedule was devised with reference to the literature and based on the results of the previous data collected (Chapters 6–8) with the intention of further exploring the results obtained thus far. The development of questions proceeded as described by Braun and Clarke (217). Open questions were used initially and in addition, scenarios depicting examples of instances of failure to fail and grade inflation were developed and presented to participants as prompts to promote and facilitate discussion about topics including sensitive areas (such as where behaviours had deviated from ideal practice). See Appendix F for the interview schedule. The interview schedule was developed and underwent several iterations after reflection and discussions with the candidate's project supervisors. Pilot interviews were conducted with two of these supervisors (one a

general practitioner and one a physiotherapist) and one other health professional (an anaesthetist). This led to several further changes in the schedule including a shortening in its length (by reduction in the number of questions), rewording of questions to improve clarity, and reduction in the planned number of scenarios presented. All interviews were conducted by the candidate. Rapport was easily achieved and the prior knowledge of participants and insider status of the researcher did not appear to overtly affect the interview process. Given most of the interview participants were known to the candidate, efforts were made to avoid glossing over information that was common knowledge to both but instead discuss and therefore record them so they could contribute to the dataset. Power or hierarchy issues were not felt to be encountered in the interviews as none of the participants had a direct working relationship (including committee work) with the candidate. One participant did however appear to be uncomfortable discussing some issues and at times gave brief answers only. These related particularly around discussions about the processes of delivering assessment messages to trainees. All other interviewees spoke easily and freely during the sessions.

Further modifications to the interview schedule were made after the first interview — in particular, questions about the differences between telling bad news to patients in clinical practice and about performance to trainees were added to the schedule.

Several participants proved to be quite challenging to keep on track and needed to be guided back to the interview schedule. In some cases, I gained the impression that the interviewee was keen for their own agenda to be discussed and these often did not align with the research topic. For example, some interviewees were keen to discuss their views on the trainee selection process or professional political issues, such as working conditions. As such, some interviews felt like they were an opportunity for participants to voice their grievances, while others felt like debriefing sessions, where participants acknowledged their experiences, at times including regret and uncertainty about their behaviours, and seemed to be seeking approval or reassurance about these. Two participants in particular expressed high levels of personal frustration and distress, and at the end of the interviews were re-offered the opportunity to speak with the main project supervisor (as outlined in the ethics application).

Transcription and trustworthiness

The interview recordings were professionally transcribed using the orthographic or verbatim method immediately after they were conducted [for example as with studies by other authors (102,109,189)] by Pacific Transcription, Milton, Queensland, Australia. The verbatim method was chosen as it was felt to provide sufficient information for the thematic analysis which followed (217). The decision to have audio files professionally transcribed was made in order to obtain a high-quality document within a more reasonable time period than would have been possible had the candidate undertaken this. Upon their return, participant demographic data was removed and stored in a separate file from the body of the transcription. Transcripts were made anonymous by the candidate and checked back to the original audio in their entirety (226). Minor alterations to the transcript were made, including the corrections of errors and addition of unclear words where these could be recalled or worked out from the context of the discussion or notes taken during interviews. Transcripts were returned to the participants [as in (6)] by email for their checking. Most participants made minor changes, such as corrections where the transcript was incorrect, clarifications to more fully explain what was written in the transcript, and addition of unclear words from the recording. One participant made several deletions and rewordings of their transcript due to concerns that the content made them identifiable. These processes occurred while new interviews were being conducted.

Data analysis

An inductive thematic analysis (185,265) was performed on the verbatim transcripts as follows. This was in the tradition of a qualitative description as outlined by Sandelowski (266), meaning the analysis is less interpretive and findings “closer to the data as given” (267). After completion of all interviews, all transcripts were read again and then reread several times by the candidate. The supervisors read between two and three transcripts each in detail (all transcripts were available to all). The candidate made initial noticings of areas of interest and recorded these separately and on hard copies of the interviews. The candidate open-coded all interviews and the supervisors coded two or three each and these codings were discussed as a group. The candidate then developed focused codes based on the first three interviews and further meetings were held to discuss these. Several diagrammatic representations of this work were

developed, and a focused coding framework including a description of codes and examples was formed by the candidate after several iterations. All transcripts were then imported in to NVivo 12, QSR International to facilitate coding with reference to this framework. Modifications to codes and additional codes were added during the coding of all interviews and these changes were again discussed with all supervisors. This led to the development of themes and subthemes, with note of the relationships between. Finally, the themes were defined and named and supporting quotes determined for each one. These were checked back to the focused codes, the initial noticings and the transcripts to ensure completeness of the process.

Results

Demographics

The characteristics of the eight assessors interviewed are presented in table 9.1. To summarise, there were five females and three males interviewed who practised across three states. Two of the group were overseas trained. Half of the interviewees had practised and assessed between 5 and 10 years, with the remaining more than this.

Table 9.1: Demographic features of interviewees

Demographic feature		Number
Gender		
	Male	3
	Female	5
State		
	Queensland	4
	Victoria	3
	South Australia	1
Training		
	Australia	6
	Overseas	2
Years of practice		
	0–5	0
	5– <10	2
	10– <15	2
	15– <20	2
	>20	2
Years of assessing		
	0–5	0
	5– < 10	4
	10– <15	0
	15– <20	3
	>20	1
Role		
	CS	4
	SoT	4

Interviews ranged in duration from 37 to 75 minutes.

Synthesis and interpretation

Assessors discussed many aspects of their experiences with the clinical performance assessments which the thematic analysis developed into the following themes and subthemes (table 9.2):

Table 9.2: Themes and subthemes developed from the thematic analysis of the interviews

Theme	Subtheme
What being an assessor is like	
“Everyone knows?”: assumptions and beliefs about assessment	
Where it’s easier	
“I mean it’s not easy for the assessor”	Ideas around failure: “fear of failure runs deeply” Pressures of being an assessor Previous bad experiences: “Bloody takes it out of me” Fear of potential outcomes “I don’t know if this is okay?”
“So, what I do ...”	DAD behaviours: delay, avoid or distort the message Give the message without the negative consequences Methods and messages
What is needed	Practical support: “ <i>I would like help</i> ” The whole is greater than the sum of the parts Reframe a negative to a positive: “ <i>spun in a positive way</i> ” Emotional support

These are presented separately in the next section with supporting illustrative quotes. The theme entitled “So, what I do ...” may be viewed as a key theme in this analysis as it describes how assessors engage in MUM behaviours. The theme entitled “I mean it’s not easy for the assessor” describes some of the reasons assessors MUM, and the theme “Everyone knows?” includes some ideas about the culture or the way things are done that foster this.

What being an assessor is like

This theme describes the views that assessors held about their role, including their views on the scope of their role and their opinions of the clinical performance assessments.

Interviewees’ range of views on being an assessor varied from the positive to more negative, though some were neutral about the role. Many valued the opportunity to be involved and help trainees to develop professionally:

“Mostly it’s just really enjoyable to see them develop as dermatologists” (A5)

Some felt it led to a greater understanding of the trainees:

“I think you have a closer relationship with the registrars because you need to be aware of what's happening with them and what's happening in their sort of psychosocial background and you also need to be aware of their sort of academic progress.” (A3)

Several discussed the responsibility they felt and the requirement to keep their own skills and knowledge up to date:

“It’s mainly really rewarding but I suppose it can be challenging and you do feel the responsibility of it ... It’s also an opportunity to keep your own skills and knowledge fresh, because we have a responsibility to do that.” (A5)

Many expressed a deep commitment to the assessment process and a sense of obligation to contribute to the training of the next generation.

Assessors spoke about the work they do beyond assessing — including providing trainees with general guidance and motivation, such as encouraging them to complete training requirements and scheduling assessments, though some left this up to the trainees. Many spoke of trainees’ need for pastoral care that extended beyond what they felt was expected of the role and recognised the mental health issues facing trainees, particularly in relation to the Fellowship exams:

“the extreme stress and anxiety they experience around their exams, which you see them suffer with enormously and I often feel they need more in the way of pastoral care than a typical supervisor because it’s so, so stressful.” (A5)

“these days they just have generally a lot of personal stuff going on, a lot of anxiety and then also depression around the exam” (A1)

One assessor spoke specifically about their role in helping to take focus off the exams:

“I have to take the focus off passing the exam to, we are training you to a higher level to present some service to the Australian community” (A8)

The relationships that assessors had with trainees differed among those interviewed. Though there was overlap, some viewed themselves in a more professional

colleague/mentor role, others a more pastoral care/supporting role and others a parental or gatekeeper role.

In regards to the assessment tools in use, several assessors believed them to be valuable:

“The idea is beautiful and very well worked out. Obviously from people who understand what they’re doing, all the people who have a lot of experience in more than pedagogy” (A6)

though differing opinions were noted:

“I think they’re patronising actually. You know, the good trainees, it’s just wasting my time filling out a piece of paper, the poor trainees, well, it’s wasting time filling out a piece of paper” (A7)

One assessor questioned the purpose of the tool — whether they were an end in themselves or should lead to other outcomes. Some doubted the ability of their colleagues to provide an accurate assessment and another questioned whether it was their right or responsibility to potentially halt a trainee’s progress in their chosen career. Reservations about the ability of the assessments to detect, identify and then manage underperformance were expressed by several and some held the view that the Fellowship exam would step in when the tools failed. Assessors also held polarised views on the potential of trainees to complete training successfully. For example, one assessor believed that all those selected into the training program were capable of becoming “excellent dermatologists” while another had no faith in the selection process. Others doubted the benefit of the assessment outcomes anyway:

“they all get through in the end anyway, because then they’re coached and then they’re mentored and then they’re fudged.” (A7)

“Everyone knows?”: assumptions and beliefs about assessment

In most interviews, assessors discussed holding beliefs about aspects of the clinical performance assessments and gave the impression that they felt these beliefs were shared among or held by other assessors too. These related to beliefs about individuals, such as the standard of trainees, and to beliefs about processes, such as around the way

things are done, as with a folk belief (268). This included that the assessment rules are sometimes “flexed” or unofficially and silently adapted or even circumvented — by both themselves and by the trainees.

Almost all of those interviewed indicated that consultants hold a collective, though often off-the-record awareness of the standard of a trainee:

“everyone knows they weren’t really a great candidate” (A2)

“we know after year 2 or after year 3 if people will be able to go ahead or not” (A6)

And they described that it often did not take long to come to this conclusion:

“You immediately know it.” (A6)

There was a sense that this awareness occurred alongside recognition that the clinical performance assessment system was not working as it should and was therefore felt to be responsible:

“We’ve had some trainees who were appalling at certain areas of their performance and that was unanimously agreed upon by consultants. They subsequently failed the exam but they were never failed on the SITA or not enough or not repeatedly failed to the point where they should have been” (A1)

There was awareness, that at times, they and others avoided giving trainees appropriate negative feedback:

“I imagine there’s not nearly enough negative feedback provided in a constructive way” (A5)

and the implications that this may have, such as not passing the Fellowship exams, and the cost and effect on the trainee that accompanies this:

“there’s no value in saying they’re all great during that training and knowing they’re not going to pass.” (A2)

Assessors discussed having knowledge or a folk belief that their assessments are influenced by many factors, such as relationships with the trainees — that the lens

through which they viewed the trainee could have an impact, especially if they were “willing them to do well” (A5):

“If I really like a person and I’m very impressed for a variety of reasons then my assessment will probably always be slightly better I mean, in reality. Theoretically of course it will be completely independent and everyone will say of course it’s completely independent, but I still think that this subjective note is very difficult to avoid.” (A6)

Some assessors voiced their perception that such issues extended beyond individuals to whole departments:

“it is my opinion, that there are some centres of training that never criticise people. They are too nice. Lovely to work for, but they send us rubbish. Then we’re expected ... to polish them off for the fourth year, or third year” (A7)

While the participant did not present specific examples of this, their comment highlights the impact that this view has on them.

Many knew that difficulties with giving trainees performance information extends beyond the formal assessments also:

“many consultants don’t give feedback at all even when they think that things are not going well. They just turn up, see the patients and then they leave and they don’t do any form of feedback.” (A1)

Some conveyed a knowledge of trainee involvement in the avoidance of the delivery of assessment messages. Assessors discussed being aware or having beliefs in regard to strategies employed by some trainees, including their avoidance of assessors identified as “hard markers” or ones who conducted assessments by the book or their avoidance of more challenging cases:

“but this year, no one has asked me to do any. And I didn’t fail anybody so I don’t know why they’re not asking me ... My hunch is because I make them do it properly and I sit in the room” (A1)

“There are trainees that deliberately avoid complicated stuff” (A7)

Another also felt at times there was unintentional masking of underperformance by trainees and that assessors had awareness that this occurs:

“I don't think anyone's intentionally masking anything, but there's a level of sophistication I think in some trainee's presentations that does mask in a way sometimes less than desirable level of either professionalism or competence”
(A5)

Where it's easier

This theme describes circumstances where assessors either found their message was easy to deliver or factors which helped to make the delivery of a difficult message easier for them.

None of the interviewees reported problems with delivering positive assessment messages and some reported it was a nice experience. They felt confident to deliver a positive message without feeling the need to collaborate with others and found it took less time and work than assessments where there were performance issues:

“I think delivering good things is really easy” (A2)

Another felt somewhat saved from having to deliver negative messages as in their view, trainees are usually very competent.

Other situations made the delivery of negative assessment messages somewhat easier for assessors. These included trainee factors — for example, the knowledge that a trainee had insight into a lack of progress or wanted to receive constructive feedback helped assessors with delivery:

“I had one trainee who [s/he] said to me, it's really good that you've given me this feedback because I need to learn to get better. So [s/he] was you know exceptional in [his/her] acceptance of criticism” (A2)

The seniority of the trainee was relevant for some assessors. For example, some felt that borderline ratings should really only be given to junior trainees, thus making this somewhat easier to do for them. However, others expressed the opposite view.

Assessors identified factors related to themselves which influenced how easy it was to deliver a negative assessment message. For example, some assessors described feeling an intrinsic lack of pressure or stress in relation to assessment duties. One identified them self as the “bad cop”, which may have facilitated the easier delivery of negative messages. Some assessors described methods they employed to help make the delivery of what they perceived to be a negative message easier. These included keeping the negative feedback light and offering to complete an assessment again in the near future. Being engaged with the assessment process and feeling an obligation to be honest or to uphold the profession helped make it easier for some to deliver their message:

“I do not think it [not giving appropriate feedback] serves them and it doesn’t serve us as colleagues ... Even though it’s hard you know it shouldn’t stop us from doing it.” (A2)

Similarly, recognition of their obligation to patients and the community helped others:

“there’s a real responsibility to do it for the patients” (A5)

A desire to avoid giving a trainee a mixed message about their level of performance encouraged some assessors; for example, they could see the problem when trainees were given passing marks during training but then poor ratings towards the end of training or if they went on to fail the Fellowship exam. Several assessors highlighted their motivation to give honest feedback about performance stemmed from the limited time period for training and the financial cost of the Fellowship exam.

The assessment format was the final area identified which affected the level of comfort assessors felt in delivering messages, with some feeling it easier to deliver assessment messages in the context of a WBA than for a SITA:

“The SITAs as I say is harder because they’re sort of – it’s kind of more a reflection of them ... as a person. Whereas in the clinic it’s kind of about – you know it’s about teaching them about dermatology and being a dermatologist. So it’s much easier. It doesn’t seem as personal” (A2)

Some felt that other features of the assessments potentially helped them to deliver their message: in particular when assessors were anonymous; though doubts were expressed

that the SITA is truly anonymous. Where a trainee had completed multiple assessments with an assessor, some interviewees felt this facilitated the delivery of a negative message, compared with doing so after a single assessment or where the trainee had not done one for a while:

“if they’ve done it 12 times this year and I’m like: ‘Holy crap that was terrible’, that’s a different scenario to: ‘Holy crap that was terrible but that’s the first time they’ve been assessed this year’.” (A8)

“I mean it’s not always easy for the assessor”

This theme, “I mean it’s not always easy for the assessor” (A6), describes the reasons why assessors found their role was challenging at times with the use of several subthemes:

- Ideas around failure: “fear of failure runs deeply”
- Pressures of being an assessor
- Previous bad experiences: “Bloody takes it out of me”
- Fear of potential outcomes
- “I don’t know if this is okay?”

Difficulties for assessors often appeared to stem both from a firm belief that failure in these assessments is a very negative thing and by the fact that the decision to award an unsatisfactory outcome or not, is ultimately theirs. Assessors described aspects of their role and around the delivery of negative assessment messages that increased the challenge, including feeling sorry for the trainee and knowing they have very little work-life balance. Some assessors reported previous negative experiences of delivering a negative message and also discussed their fears of the potential outcomes of doing so. Many also highlighted uncertainties or unknowns about the processes which contributed to making the process more difficult for them.

Ideas around failure: “fear of failure runs deeply” (A5)

Assessors described that failing an assessment was an unusual event and not to be taken lightly. The effort involved in giving a negative assessment, or even a borderline rating was apparent and completely different to delivering a positive message:

“I wouldn’t ever put people down as a three [in a WBA] until I’ve had a real think” (A8)

Assessors spoke of failure as being “quite devastating” (A5) and “distressing” (A2), especially for such a group of “*highly achieving people who have never failed anything in their life*” (A1). Failure was seen as a big deal or as a punishment for a trainee, which would lead to them feeling shame, embarrassment or awkwardness. They also recognised consequences for themselves:

*“I feel like you’re perceived as just being a pain in the butt if you don’t just do it all so they can sign it off. Unless they’re really hopeless and you’re **really** going to make a point of it.” (A1)*

Failing a trainee in a SITA was viewed as a major event, requiring a lot of evidence from different sources and involving “enormous amounts of time” (A1):

“They had some major and significant problems and multiple patient complaints. Problems in knowledge, problems in communication, problems in work performance and then multiple patient complaints.” (A1)

Some assessors were aware of the other pressures that trainees face, from both within and outside the training program, and particularly around the Fellowship exam and did not want to compound this stress by delivering a negative assessment message. With the scenarios as discussion points, most assessors agreed that failure to fail and grade inflation occur currently.

Some assessors described their feeling of personal responsibility for the outcome of documenting the situation:

“How tough will you be?” (A6)

“will I or won’t I give this person this negative mark here?” (A4)

They discussed that failure, as distinct from passing, generally also requires support from colleagues, such as other assessors, to both confirm results and provide them with reassurance that they were being fair. Other hospital departments such as Medical Education, Human Resources and equivalents were also often enlisted by assessors for

support, especially as some felt disempowered to discuss performance issues with trainees:

“The terminology and the language has to be perfect otherwise you know we’ll upset people” (A7)

Pressures of being an assessor

Many assessors discussed the challenges and pressures that they experience with their role. A lack of time for observation and assessment of trainees was noted by most:

“it’s all very time pressured. We don’t have training and we don’t have allocated time. So as I’m meant to leave for the day to pick my kids up from school or to get out or to get onto the next patient they say, they’ve got this form under your nose, ‘oh I need to get these ones done for my logbook’” (A1)

Some felt clearly overwhelmed by their workload, and factors unrelated to assessment — such as perception of loss of autonomy over clinic practices, and procedures around the referral of patients — contributed to this. Some expressed profound frustration and a sense of vulnerability with the situation:

“I feel unsafe and the reason I feel unsafe is that I have to see too many patients with junior registrars, or not even registrars, and the senior ones are more focused on their exams et cetera, et cetera, and there have been mistakes. So, there are lots and lots of reasons that I’ve been unhappy. I’m struggling to find positive in any of it at the moment” (A7)

Assessors described facing pressures from other staff at times, such as nursing staff due to their frustrations with a poorly performing trainee:

“the nurses were getting really angry with me because they were like ‘it’s up to you to sort this out, we can’t’” (A8)

Some found the pressures of aspects of the supervisor role to be greater than those of the clinician role. For example, the delivery of negative assessment news to trainees was seen by some to be more challenging than delivering bad clinical news to patients. Though this was appreciated as a difficult situation also, they felt comfortable with the

latter, largely due to their experience and training, which meant that they also knew how to manage the situation:

“If you’re in a clinical situation dealing with patients clinically, this is what you’ve been trained to do. I feel like that’s where I’m comfortable, that’s what I’m trained to do. Like I can help the patient, I know how to help the patient, I feel like I can do that in a way that helps, that makes the patient’s life better. Whereas maybe with this kind of thing I actually don’t feel - I’m out of my comfort zone a bit, or partly that I think.” (A5)

Previous bad experiences: “Bloody takes it out of me” (A8)

While many assessors highlighted parts of the role they valued, several also discussed the impact that assessing has on them individually. This was seen to extend beyond the time and effort required, to include the associated emotional toll of the role. While not all assessors felt that delivering negative assessment information to trainees had an impact on them, many spoke, some reluctantly, about feeling anxious when delivering assessment results, disliking confrontation and their need to spend huge amounts of time on negative assessments, which took them away from their own family time and for which they ended up feeling “terrible”:

“So dealing with registrars in trouble can be a very stressful thing which takes up a lot of your time and emotional energy depending on how difficult the situation is ... but I have had a number of times where it’s been a very difficult journey with certain registrars and I found it quite draining” (A3)

Some discussed the challenge of dealing with trainees with “difficult personalities” and those who do not take advice or criticism or respect their authority or experience. This was even more frustrating when they felt they knew how to help the trainee:

“You need to listen to us; you need to have faith and trust in us” (A7)

Others expressed the problem of dealing with a poorly performing trainee when they were either unsure of what the problem really was or how to help them. While being involved in failing a trainee on an assessment had a huge impact on the assessor, so did facing a trainee who was upset with their results. Some assessors described the

emotional reactions from trainees — being tearful and upset. Some described feeling “shocked” to learn that a trainee had complained to others about the negative feedback that they had given them, and that this resulted in them changing their assessment practices in the long term. For example, some reported this made them very nervous to say anything negative or to have all such interactions witnessed or to obtain external help to deliver assessment messages. Others described receiving written responses to a negative assessment from trainees. Some summarised the bad experiences of having delivered negative assessment messages in the past:

“I’m terrified to give a one or a two because I just know the chaos that it causes” (A7)

“They don’t take it well” (A1, 2)

Legal challenges to previous failing decisions were also briefly mentioned.

Fear of potential outcomes

Several assessors expressed deep reservations about giving a trainee a negative assessment message as they feared the outcomes, for both the trainee and for themselves.

Assessors were mindful of the potential consequences of a poor assessment for the trainee. This ranged from awareness that they would have to do it again:

“[s/he’s] just trying to get [his/her] logbook signed off at the end of the year. [S/he] just doesn’t want me making [him/her] do it again.” (A1)

They also feared offending, upsetting or hurting trainees or damaging their sense of self, especially when they felt sorry for the trainee; for example, when they knew they had been under personal stress. Some SoTs recalled being contacted by other assessors who were concerned about implications their assessment result would have for a trainee. Assessors were aware of trainee mental health issues, in particular anxiety and depression, and noted that at times this led to a modification of their assessment:

“You know, there have been some registrars who have had personality disorders and had mental health issues which with very minor feedback have really taken

that very poorly and just given their emotional reaction, I haven't felt that I could continue bringing anything further negative into the discussion because it wasn't likely to be helpful at that stage. So I think you do have to be mindful of their reaction and you know, how fragile they are, just how you frame things. You can't just have a speech and just give it. You need to really be aware of you know, the impact that that's having on the registrar.” (A3)

Two assessors talked about their concerns that assessments could potentially lead to suicide — one spoke of this in the context of the Fellowship exam and one in relation to clinical performance assessments.

“It takes a lot of performance effort on my behalf, at night ... And then it takes me a couple of days to think about, did I represent everything appropriately? Do I think that [s/he's] taken something, is [s/he] committing suicide right now? You know like, you know you just sort of, it takes a few days to calm down from that. And then the next week and the registrar comes back and says ... what you said was so great, thanks. It's only until a week later when they say that that you're just like, oh thank god [s/he's] okay.” (A8)

The tremendous burden these concerns place on the assessor were evident. As mentioned, assessors felt such issues were compounded by the fact that they feel it is ultimately their decision “*whether to put somebody in the red so to speak or not. It seems like such a large barrier to cross*” (A4). Assessors also feared harming a trainee's future career prospects or being responsible for them being exited from the training program.

In regards to concerns for themselves, assessors feared the delivery of a negative assessment would lead to more work, in addition to possible complaints, allegations of bullying, fights and being thought of as mean:

“I don't want to have all the paperwork. And you don't want the person to hate you.” (A1)

Fears of being sued and complained about were also raised and that the fail would make their future relationship difficult:

“The truth is, most of them are going to become dermatologists and you’ve given them a big fat fail, that doesn’t help our community because dermatologists hate dermatologists.” (A8)

Some appreciated the difficulty of the situation for assessors:

“so I don’t criticise anyone because I see the human aspect” (A6).

“I don’t know if this is okay?” (A3)

Assessors described uncertainties and at times, misunderstandings around some aspects of the assessment process which contributed to making the assessment process challenging. Difficulties determining the appropriate standard and benchmarking trainee performance was noted by most assessors:

*“I’m usually basing them on what I thought was expected of me as a registrar or I base it on registrars who I think are the highest performing but I don’t **really** know what the assessment standards are other than that.” (A1)*

At times they wondered whether they were able to detect poor performance, especially when assessment results are not discussed as a group:

“I haven’t yet had an experience, and I don’t know whether I’m one of those supervisors that’s sort of unable to - or are not very good at picking up problems, but I’ve not had an experience where I’ve, anyone has been really off ... they’ve all been on, you know, scored as competent. So does that mean that the test is sort of unnecessary or that it’s not discerning enough? Or is it the operator of the test?” (A5)

In addition to recognising the challenge of assigning the trainee to a particular category, the difficulty of providing formal feedback, especially in the context of the SITA was raised. Some questioned whether their assessment message would even have the required or desired impact:

“is it does depend on the insight of the trainee as to how well they may respond, if at all ... One thing is to identify it, another to deliver it, which I think is a challenge, but the third really is in just thinking, well, how can we effect change? I think, yes, it is a little frustrating thinking that I’m not really sure how much I can do.” (A4)

Even when issues with trainee performance were recognised and the assessor motivated to act, at times they struggled knowing how to proceed:

“I was in a position where I had to tell a registrar that [s/he] was bad, on many fronts ... What does [s/he] do with that and how do I deliver that to [him/her]?” (A8)

Some questioned their responsibility and how far their role as an assessor should extend; for example, in relation to issues such as professional behaviour or personality:

“You don’t want to change people’s personalities but sometimes it is in their actual personality ... That’s simply their personality is interfering and it’s got nothing to do with their judgement or their abilities or anything” (A8)

Some assessors expressed uncertainties about practical aspects of the assessment processes; for example, when assessments were required to be submitted, how SITA comments were to be included in the summary form, and the criteria for an unsatisfactory SITA and the processes that followed *“can you just remind me ...” (A2)*. In one case, uncertainty about the outcome for the trainee prevented them giving a failing assessment:

“But I think having some reassurance that this is what should be done, but these are the clear consequences, not to worry there’s a process in place and you’ve clearly made this decision and you need to sort of stick with that.” (A4)

Several assessors discussed uncertainties around whether they were required to have attended supervisor training courses (and some expressed a reluctance to do so), and some revealed misunderstandings of processes through mentioning that they believed a change to the allocation of SITA assessors had occurred in recent times, when none had in fact occurred.

“So, what I do ...”

This theme, which reports the MUM behaviours discussed by participants, has three subthemes:

- DAD behaviours: delay, avoid or distort the message
- Give the message without the negative consequences
- Methods and messages

It describes the way in which interview participants resolved some of the dilemmas they face with the assessment processes and the delivery of negative assessment messages.

DAD behaviours: delay, avoid or distort the message

One assessor spoke about their practice of delaying delivery of their assessment message in the context of a WBA, in order that they process the event more fully and so have more valid and useful feedback for the trainee:

“more and more my feedback is delayed. I find that more helpful ... Sometimes I need it to be in my subconscious for a bit ... then I usually have something much more helpful to say” (A8)

They gave examples of how this delay permitted time for deep thought and reflection on a trainee’s performance and resulted in a more complete understanding of issues and so how to address them, with subsequent profoundly beneficial outcomes for the trainees. At other times, a delay in delivering a message occurred as the assessor recognised that the trainee was not in a position to take the feedback on board at that time:

“this is going nowhere; we’re not getting to the point in which I’d like to take this and I pull back ... people are not ready for information at various points in their life and then they become ready” (A8)

However, assessors also spoke about feeling intrinsically reluctant to deliver negative assessment messages to trainees:

“no-one likes delivering bad news” (A7)

and some recognised that at times, a trainee is passed when they had not met the standard:

“they are signing them off because people are too worried to mark people down because they’ll offend them” (A7)

Some interviewees described avoiding delivery of the message based on the idea that the “*exit exam will do the job*” (A6) or that the trainee will move on to another rotation and become “*someone else’s problem*” (A7). Others questioned the point of failing a trainee in their third year, seeing it as too late to have benefit — “*What are you going to do about it now?*” (A7). As mentioned, some assessors cited concerns about a trainee’s mental health as a reason for avoiding delivering their message in full, for example, by not giving the verbal information that they wished to.

Instead of delivering their assessment message in full, some assessors admitted to distorting their message to give the trainee the feedback that they thought they wanted:

“Oh, every single time” (A1)

“Yes, I think I’ve given a four instead of a three, yeah. Not lately but in the past I reckon I have.” (A2)

Sometimes they gave their reasons for this, such as their frustration with the system:

“Yeah, absolutely. Even I would be guilty of that. Because I just don’t care anymore. I’m just over the fight. I can’t fight anymore.” (A7)

or their view that doing so would ultimately achieve the best outcome:

“So give them positive feedback. I think it’s always has the best result...and doesn’t stress them additionally” (A6)

Assessors recognised that it can be difficult to assign low ratings when giving results face-to-face and when struggling with what to do with trainees on the border, sometimes erred to give them “*the benefit of the doubt*” (A4). Several assessors felt that grade inflation or giving a 4 or 5 instead of a 3 or 4 on the SITA was not too problematic, especially if the area was addressed in the comments:

“unless they’ve got their exam right around the corner there’s always room to mark up I think” (A4)

Give the message without the negative consequences

Several assessors described developing strategies to avoid marking a trainee as unsatisfactory, while at the same time still trying to communicate their true opinion of the trainee’s performance to them. For example, some did this by communicating their feedback verbally only, thus avoiding documentation of any negative ratings or comments:

“So, what I do is do it verbally and then I fill the form out” (A1)

Other methods included halting a WBA in order to avoid documenting it as unsatisfactory, instead explaining what was required of the trainee and arranging to repeat in the future.

“Sometimes I say to them, look you really didn’t do A, B and C so I don’t think we really – you’ll fail this if we put it through so I just – let’s just talk about you know what needs to happen and then we’ll do another one another time.” (A2)

Other assessors felt it sufficient to cover any concerns in the comments:

“I do think the sand [line] drawing thing is important but then not drawing it in my experience, is okay” (A8)

Some indicated that they had done this in the past but have changed their practice:

“I don’t do that anymore but I probably would have in the past. I probably would have said, look I’ll give you that, next time do blah blah you know.” (A2)

Methods and messages

Some assessors described their intention of using ratings in a SITA to provide a warning to trainees and so stimulate a change in their behaviour, though noted at times this could be unsuccessful:

“I thought, well look, this could be a signal that I’m giving by just giving one [rating of 2], but without necessarily them putting the registrar on the path of a [PIF]. Thinking back then I think it was probably a mistake.” (A4)

A further method employed was to inform a trainee verbally of the areas where they should be assigned a “3”, but to instead document a “4” and follow up by arranging further assessments thus giving the trainee a timeframe over which improved performance was expected:

“I find that helpful, to have that borderline there to be sort of like threatening, I’m not going to give you the threes but this is an area we’re focusing on.” (A8)

This technique also served to give the assessor permission to discuss poor performance should the borderline ratings fail to improve, as they had given warning of this in advance. At other times, assessors described using their written comments to suggest to the trainee that they were on the lower end of the green zone³.

As highlighted earlier, some assessors had beliefs about the best time in the training program to deliver negative assessment results; one outlining a method of giving bad feedback in the junior years with the aim of avoiding problems in later years:

“I’ve thought about the fact that when they’re very junior that is really the time to give them bad feedback. Because I think we’re all afraid to give the bad feedback and then they get to third year and we think, oh my god how have they

³ This refers to the colour coding used on the SITA form (Appendix 5). The green zone covers ratings of 4–5, the amber zone covers ratings of 3 and the red zone covers ratings of 1 or 2.

got to third year and they're still doing this kind of stuff which is... you think 'how do they not know that?'" (A2)

"You want to change their behaviour early and not confront them. Because then they think that they're being bullied or it's unfair or ..." (A2)

Others believed the opposite, thinking it better to mark a trainee approaching the end of their training more harshly, unless their exam was imminent. Further methods to help delivery of a message included taking the time to educate trainees about the ratings — in particular, that a low rating is not necessarily a fail, nor a negative reflection on the trainee as person, but an opportunity to address an area of need. This was seen as preferable to saying they are at standard when really they are not. One assessor rephrased criticism or the thinking around what was “wrong” with a trainee, to instead thinking about what was holding their progress and performance back.

Some assessors recognised their own personal difficulty with giving negative feedback or results to trainees and removed themselves from the situation by assigning others to complete the required assessments with the trainees.

What is needed?

This theme incorporated four subthemes:

- Practical support: *“I would like help”*
- The whole is greater than the sum of the parts
- Reframe a negative to a positive: *“spun in a positive way”*
- Emotional support

The interviewees both openly and by implication recognised they need help with the assessor role and, in particular, with the delivery of negative assessment messages, and some offered suggestions as a way forward.

Practical support: “I would like help”

As discussed, several assessors spoke of the lack of time for and training in assessment and for the formulation of valuable feedback:

“We don't have training and we don't have allocated time” (A1)

“None of us know what we’re doing” (A1)

“It takes a lot of effort” (A8)

and the impact that a lack of time then has on the assessment message that they deliver. The issue of funding was recognised by some and the fact that assessment work is done pro bono. Those who had experienced ACD training had found it valuable and suggestions for further sessions varied from small face-to-face sessions or a pack for supervisors of training covering assessment processes, standards and definitions of unsatisfactory performance, through to a more intensive option, recognising that skills beyond knowledge of processes are required:

“we probably all need to go away for a weekend and do endless roleplaying in awkward situations. And also we’re not very assertive, it’s very odd, I don’t think we’re very assertive.” (A5)

Some assessors also called for specific training on how to give feedback and how to remediate or help trainees:

“I’ve got to make someone better and I don’t know how to do that” (A8)

While opinions varied, some participants felt assessor training should be compulsory:

“at the end of the day, if you are a good dermatologist it doesn’t mean that you know how to assess a registrar” (A6)

One assessor mentioned practical support options, discussing technological advances used by some other colleges; for example, apps for recording and delivering feedback to trainees. Some interviewees mentioned the importance of recognising the fact that not all of those who are selected into the program will necessarily complete it. One mentioned the importance of trainees recognising this as a possibility and so developing a sense of themselves that is not completely tied to becoming a dermatologist. Another mentioned having employment and work options for those who cannot complete the assessment program, where such doctors could practice in dermatology without having reached the FACD qualification.

Having an external assessor — who is unbiased by an ongoing relationship with the trainee — deliver assessment information was seen as a further potential method to help. A suggestion of having trainees on a probationary period at the start of their training was also made. External support from the ACD or a hospital department (such as Medical Education or Human Resources) was seen to facilitate the delivery of negative assessment messages to trainees, and where such support was not obtained, assessors reflected that this could result in them not delivering their assessment message.

The whole is greater than the sum of the parts

Most assessors described that assessment should be a team event, especially in relation to the assessment of underperforming trainees. They recognised the need to involve and discuss with colleagues to ensure the accuracy of their own assessment:

“It’s partly just to reassure my own assessment of being accurate” (A4)

and to also know that they weren’t going out on a limb alone. Some talked about the fact that some amount of collaboration occurs already. For example, one assessor said that in the case of poor or polarised views of performance, their practice was to collect additional opinions from other CSs to ensure the documented reports were fair and consistent. Others expressed caution that discussions with others could potentially sway opinions, and their view was that an assessor should not be influenced away from a failing decision by a group discussion. Group discussions about trainees were felt to potentially help assessors create more knowledge about the particular problem areas facing trainees and so allow the development of better ways of helping them. This would also avoid them “flying under the radar”. Further, group discussions were also felt to promote the discovery of previously hidden things; for example, dishonest behaviour by trainees. In areas where discussions were already occurring, assessors felt the collaboration gave much more information about the trainee.

The benefit to the assessment message of broadening the pool of assessors was discussed by some. For example, some felt that senior nursing staff could make a valuable assessment of a trainee’s professional behaviours and potentially contribute different information, as they believed that at times, trainees show different behaviours to consultants compared with to other staff:

“She sees the people who will put in the half extra hour of work ... she sees if the persons are finishing the job and how they are communicating and how they are collegial” (A6)

Reframe a negative to a positive: “spun in a positive way”

The deep underlying fear of failure in these assessment contexts has been presented in the theme — Ideas around failure: *“fear of failure runs deeply”*. Some assessors felt that normalising the process of identifying unsatisfactory progress or areas where trainees were struggling and highlighting the benefits of remediation, would reassure assessors and help them deliver appropriate assessment results without them feeling that they were doing something harmful. Similarly, some felt they needed encouragement to stand by their assessment decisions.

Emotional support

Throughout the interviews, most assessors expressed or conveyed that the assessor role has a significant emotional impact on them. This did not just affect them, but also impacted on their families:

“it can be a bit stressful certainly my [partner] hears more about the stress of that than anything else in my working week” (A1)

Some talked about their wish for more support in their role, such as having a colleague to advise or guide them, and some highlighted the need for debriefing:

“I think having someone you know like psychologists do don’t they, they go off and they debrief at the end of every day. Like I think we probably don’t do that quite enough.” (A5)

“This was an extremely therapeutic opportunity for ... I really do need psychological support over all of this and this [interview] has been just therapy for me.” (A8)

This emotion and their need for support was conveyed through assessors’ talk about their concerns in regards to their role — the pressures they face and their feeling of responsibility for decisions that have potentially severe consequences such as being

removed from the program or leading to trainee suicides. Some assessors expressed feeling completely disempowered and frustrated by several aspects of the training program, not just limited to assessment:

“I’m so cross about all of this. Just thinking about all of this this morning has made me really irritated and upset.” (A7)

A need for acknowledgement of the role and recognition of the contribution made by assessors was highlighted by some, both from trainees and the ACD:

“I think they need to be thanked because they’re doing a thankless job” (A7)

“It’s been interesting to me some of the ones who I’ve put the most effort and time into, after hours, practice exams, given them practice essays, marked their essays on my weekends and then they’ll finish and they won’t say thank you. I have learned that you can’t expect that.” (A1)

Discussion

Summary and interpretation of findings

These results show that while assessors may have a deep commitment to the training program and the supervision and assessment of trainees, they face dilemmas in their role, which under some circumstances can lead to a reluctance to deliver their assessment message. This potentially places the outcome of the clinical performance assessments in jeopardy. This study shows how and why assessors experience reluctance and adds to the assessors questionnaire chapter to further illustrate the significant levels of stress that many assessors experience in their role. The wellbeing of clinical supervisors in general is an area gaining recognition (269), and these results contribute by highlighting the pressures and level of stress that many experience in regards to assessment duties. These issues need to be acknowledged and addressed, both for the clinical performance assessments to achieve their pedagogical aims and also, for the sake of the assessors experiencing these challenges.

Though many interviewees expressed problems and difficulties, the level of commitment the eight interview participants have for the trainees and the program was clearly evident. While many discussed aspects of the assessor role that they enjoy, all

highlighted the problems and dilemmas they face. Though several believed that they are obliged to contribute to the training and assessment of future generations of dermatologists, it seems that many are at their capacity and the pressures are impacting on themselves and the assessment outcomes. In addition, some questioned the purpose and importance of the assessments. As a consequence, and perhaps due to the significance place upon failure, the results show that all interviewees recognise that at times, assessors keep MUM by avoiding, delaying or distorting the assessment message:

“at the moment we probably don’t fill them out truthfully enough”

The situation risks approaching a downward spiralling one — where assessors, who for various reasons have certain doubts about the assessments, feel increasingly reluctant to deliver a negative assessment and end up engaging in MUM behaviours; but this then leads to added frustration, and further diminishes their view of and the value they place on these assessments and outcomes. While it is recognised that some of the MUM behaviours described by assessors may in fact promote the delivery of a negative assessment message; for example, by a limited amount of delay or distortion, others were seen to be detrimental and result in failure to fail and grade inflation.

The results from this study show that assessors are under enormous pressures. They face competing interests on a daily basis and this has been noted elsewhere [for example (269,270)]. Pressures come from the added duties of their role, such as the workload to help trainees arrange and complete their tasks, and the conflict and stress they feel when they need to formulate and then deliver an assessment message, especially if negative (271). Further to the results from the assessor questionnaire study, where some assessors self-reported being uncertain whether they had failed to fail or grade inflated a trainee before; in this study, some assessors expressed uncertainties about the standards required or the criteria for unsatisfactory performance. Yet in contrast, many indicated that they intrinsically “knew” which trainees are underperforming. This knowledge often occurred outside of the formal assessments and often within a very short period, similar to the “intuitive response” explained by Bearman and colleagues (270). However, whether they were confident or not of the trainee’s standard, the fact that they felt responsible for the delivery of the assessment message contributed to the dilemma faced by some:

“will I or won’t I give this person this negative mark here?”

Assessors recognised facing pressures from other stakeholders as well, staff who expect them to address trainee deficiencies; but in relatively few instances, voiced the collective obligation to the patient:

“there’s a real responsibility to do it for the patients”

This is reminiscent of the more extreme findings from Guerrasio and colleagues, who report that in their study across medical, physician’s assistant and nursing schools in the USA: “[s]adly, not one respondent to the survey commented on patient safety or the quality of health care provided by underperforming students that graduate and proceed into the health care workforce” (92). While assessors want to achieve the right outcome for all, the results show how they fear the potential consequences and in some cases have been affected by previous bad experiences:

“They don’t take it well”

The concerns with the delivery of negative assessment messages echo those reported in the literature (80), such as for assessors themselves (e.g. the additional paperwork, time involved, being hated or sued) and for the trainee (e.g. having to repeat a failed assessment, feeling sorry for their circumstances or where they are aware of their deficiencies or of losing their chosen career). Such concerns also highlight additional concerns assessors have about trainee’s mental health and the risk that a failing assessment may precipitate a suicide.

The level of distress experienced by some assessors in relation to these and other pressures was clear from the interviews, and together with results from the assessor questionnaire chapter contributes to the literature by documenting this in an empirical and original study. Here, for some, distress related to areas beyond the assessments, to include the selection process and political aspects, and some interviews provided what seemed like an opportunity for assessors to debrief about these. The clear view that failure is a very negative outcome feeds these difficulties that assessors face, especially as they feel ultimately responsible for putting the trainee “in the red or not”. These results show the importance of understanding the impact that assessment duties have on

assessors' level of wellbeing and the supports that they have for this voluntary role, given the clear impact that it has on them.

Implications

Assessors in this study clearly articulated some areas where they feel support for their role is lacking and how addressing such areas may help their situation. While several acknowledged and were grateful for the support of the ACD and their health service departments when faced with needing to deliver a negative message, there were indications that other supports are lacking: assessors have not been given enough resources, including time, to complete their role. In addition to making their role more difficult, this also risks sending the message that these assessments are not of great value. As highlighted in Chapter 4 in the systematic review, Shalhoub and colleagues report the feeling that WBAs were “dropped” (8) on assessors and the ongoing resentment this has created. It is important to recognise that over half of the assessors interviewed in this study completed their training prior to the introduction of the WBA and so do not have first-hand experience of their potential value from the perspective of a trainee. Commitment to a role, whether due to a sense of obligation or not, can only sustain an assessor so far, and when the pressures exceed the benefits or this sense of obligation for too long or by too great an amount and the benefits of the assessment to the trainee are not recognised, the outcome will surely suffer.

For assessments to generate the valuable outcomes they are intended to, these aspects need to be addressed. Assessors need dedicated time and space for assessment and this is seen not to be confined to just the study setting. Other disciplines are seen to face similar problems and some have generated solutions. For example, Jones and Nanda recognised the logistical difficulties of performing assessments on junior doctors in a busy (Emergency Department) work environment, and reported that the appointment of two dedicated Education Registrars in their department allowed for more assessments to be conducted, though they acknowledge that such a solution comes with a financial cost (272). While it is recognised that conventional training has not eliminated the issue of failure to fail to date (80), targeted assessor training, especially covering the consequences and outcomes of a failing result for trainees, appears to be required; though it is recognised that in this study, there was some variation in the training that those interviewed called for. Some assessors did not appear to have awareness or

understanding of the potential benefits and the role of feedback as an educational tool (53). The potential benefit of such training may be predicted from the difference seen between delivering bad clinical news to patients that participants recognised, with many feeling comfortable in that situation due to their training and expertise in that area. However, as Bing-You and colleagues discuss, not all clinicians will be able to deliver feedback (or “dance”) well (97) and even with training, not all assessors will develop the skills and experience of an expert in education, meaning external support will likely be needed in the long term. Ready access to such additional resources and supports, both locally and at ACD level is important. In addition, though this was mentioned with almost embarrassment by some, assessors should receive recognition and thanks for the work they do. Even a “thankless job” warrants some token of appreciation. Consideration of excusing those who are not able to “dance” well in an assessor role should also be an option.

In this study, many participants highlighted the benefits of a collaborative assessment process, both in order to calibrate their assessment, ensure fairness and gain further information about trainees, and reports from the literature concur. Castanelli and colleagues discuss the benefit of sharing the burden for decisions through discussions with others (even if one assessor was ultimately solely responsible for the outcome) (271,273). Other studies discuss the routine practice of obtaining input from other colleagues, though observe this may vary in amount (272). By design, tools such as the Colleague Questionnaire (CQ) and the Multi-source feedback (MSF) rely on the input of multiple assessors and studies highlight some benefits of this. For example, Richards and colleagues note in their study of the CQ, that non-medical respondents were more likely to provide a comment on a doctor’s performance than were medical colleagues or non-clinicians (248). Further, in a study of the MSF, Whitehouse and colleagues agree “that different health professionals do differ in their scoring of trainees, [with] senior nurses tending to identify more problems with trainees” (274). Having additional formal contributions to assessments would also potentially serve to reduce the pressure for the consequences of the assessment currently felt by individual assessors. It was noted in the interviews that discussions with other assessors was only talked about in the context of trainees who were performing poorly, but discussions about all trainees would be potentially valuable and so increase the breadth of feedback available to them and the support provided to assessors.

As described in the assessment background chapter, Dudek and colleagues stated “that [assessors] do not lack the ability to use the forms as they currently exist, rather they lack the willingness” (11). The results of this study show that assessors are under significant pressures and stresses that have not been satisfactorily addressed to date, and that this is perhaps a major contributor to difficulties with the delivery of negative assessment messages. Without addressing this, the problem may never be resolved. Assessors cannot be left to remain with such pressures and dilemmas, for the validity of the clinical assessment process, let alone for their own health and wellbeing.

Limitations

The small sample size might be regarded as a limitation of this study, though it is noted that there is no attempt to generalise results from these interviews, rather to highlight where issues exist currently for some assessors. All those who responded to the invitation to participate in interviews (and satisfied the selection criteria of recent experience in assessment) were interviewed and there was not provision in the ethics submission to promote further participation, for example with purposive sampling. As such, a broader range of experiences may have provided additional insights. However, the sample was felt to be appropriate as the interviews yielded sufficient data to contribute to the answer to the research question (as described by Varpio and colleagues (275).

The literature discusses potential limitations with the type of interview used. For example, while in this study, the use of telephone interviews permitted a wider range of interviewees to participate, differences in the outcomes of telephone compared with face-to-face interviews may exist (276). While there is some evidence that telephone interviews can yield a greater amount of honest data, this may arise from anonymity (277), which was not possible in this study. While the assessors interviewed in this study appeared mostly to be comfortable discussing the issues, at times, some reluctance was noted, especially around the area of the impact that assessment has on them personally. The insider status of the candidate did not appear to impede the interviews, in fact, it mostly seemed to facilitate discussions. However, insider status is known to limit research in several ways. Being an insider and having my own perspective, experiences and expectations, both conscious and subconscious, may make it more challenging to see “what is going on” (278) and this is important to

acknowledge. While the analysis and synthesis of the data were reviewed and discussed at all points with the candidate's supervisors, the findings of this work resulted largely from my interpretations and construction of meaning as a PhD candidate. While the process involved well-described methods (217) and was carefully documented, different researchers would likely have come to different interpretations of the data, and hence different conclusions.

Conclusion

This chapter reported on the results of a qualitative study with assessors of the ACD about their experiences with the clinical performance assessments. Interviewees held a deep commitment to assessment and the training of future dermatologists more generally, but many revealed that they experience significant distress and dilemmas as a part of their role. While a lack of practical support plays a role in this, assessors are conflicted by a desire to deliver an assessment message in order to help trainees, and the potential negative consequences of doing so. Some described how they managed this and some suggestions to resolve the situation were made. The perceptions around “failure” and their felt responsibility for this were seen to impact on assessors.

The discussion chapter which follows, will consider all of the background and research data obtained in the study to holistically address the research question of how and why reluctance to deliver negative assessment messages manifests in dermatology training in Australia.

CHAPTER 10

Discussion

Introduction

This chapter brings together the findings from the project — from the literature reviews and the four original studies — to discuss and answer the research question of how and why reluctance to deliver negative assessment messages manifests in dermatology training in Australia. It is separated into three sections. The first answers the research questions with a synthesis of the data collected across all studies and in light of the literature. The second part discusses the potential applications of this work, including the implications for practice and for future research. The final part of this chapter presents a discussion of the limitations of the findings of this study.

Synthesis of findings

The clinical performance assessments have been designed to promote trainee development by providing information on their performance, however they do not currently always achieve this goal, especially in relation to those trainees who are performing poorly. The MUM effect offers additional insight and understanding as to why this occurs. This research project has focused on how and why assessor reluctance to deliver negative assessment messages has manifested in dermatology training in Australia in order to better understand the problem and so provide potential new ways in which to address the issue.

The next section summarises the results of this project to show *how* reluctance to deliver negative assessment messages manifests and the following section discusses *why*.

How does reluctance to deliver negative assessment messages manifest?

MUM behaviours including failure to fail and grade inflation

This section summarises the types of behaviours seen from the studies. As the most commonly discussed in the literature, failure to fail and grade inflation are presented first followed by examples of other behaviours, now appreciated to be MUMing.

In this study, a reluctance to deliver a negative assessment message was seen at times to manifest in failure to fail or grade inflation which is consistent with reports from many

other settings, both within (10,11,21,67,68,80) and outside of medical education (13,228). Failure to fail and grade inflation was identified in the current study through both the quantitative and qualitative studies. In the assessor questionnaire study, some assessors self-reported engaging in such behaviours and these were also discussed in the interviews. While level of self-reports were low — in the questionnaire, 6 of 48 assessors self-reported failing to fail a trainee, being 2 in a WBA setting and 4 in a SITA; and 8 of 45 assessors self-reported to grade inflating a trainee in a SITA, and in the interviews almost all participants indicated that either they had done so or were aware of others doing so — a hesitation to divulge behaviours which have deviated from those expected, the social desirability bias, may predict the figures are in fact an under-representation of the true incidences. Of further importance, the results of this study show that reluctance to deliver negative assessment messages can occur across both summative and formative assessment formats, and in both the rating and comments components of the assessments. In addition, the interview study showed that reluctance to deliver can apply to a verbal assessment message. However, failure to fail and grade inflation were not the only manifestations of reluctance to deliver negative assessment messages seen in this context.

Through exploration of the MUM effect literature and consideration of its potential application to the medical education setting, the results from the current study add to the literature by highlighting other ways in which reluctance to deliver assessment messages can be seen to manifest in the medical specialist training context. These findings help broaden the thinking on this issue, and are discussed in the following section.

The behaviours of delay, avoidance and distortion of assessment messages have been seen to occur in the study context and are identified as examples of MUM. As presented in the assessor questionnaire study, use of the MUM effect scale showed that some assessors (4.5–81.8%) self-reported engaging in distortion (sugarcoating) behaviours, though less often (0–21.7%) in avoidance. Agreement with more moderate examples of behaviours (such as “I tried to make it sound nicer than it was”) were more common compared with more extreme examples (such as “I changed the feedback or rating so it was not so negative”). Similarly, in the interview study, assessors discussed delaying the delivery of negative information; for example, until the trainee was likely

to be more receptive to the information or until the assessor had had sufficient time to consider their thoughts about the performance. At other times, assessors described avoiding delivery of the assessment message that they had planned to give, partly or in its entirety. Some avoided or limited their participation in the assessor role, instead delegating it to others, as they recognised the difficulties they experienced with it. The results from the trainee questionnaire study support these findings by showing that trainees also strongly believe their assessors engage in MUM behaviours. In fact, they felt their assessors engaged in more MUM behaviours than the assessors believed they did. The systematic review of trainee perspectives in Chapter 4 showed that trainees have awareness of assessor behaviours such as avoidance of delivering negative assessment messages, but this study shows how such behaviours can now be considered and understood with reference to the MUM effect.

Taking this broader view of the issue by focusing on behaviours beyond failure to fail or grade inflation is important for several reasons. Most critically, it suggests that the developmental impact of the assessment message may be reduced. Introducing a common language or framework to aid the identification and description of these behaviours is an important step on the way to then addressing them. Thus an awareness and recognition of MUM behaviours as ways in which assessors may manifest reluctance to deliver negative messages can potentially help to identify issues with trainee performance at an earlier stage — before failure to fail or grade inflation have occurred. Further, while the assessor may not believe that engaging in such behaviours is harmful — they not may believe that MUMing would either reduce the effect of the message delivered or result in any negative result — the MUM effect literature argues differently. For if it is too delayed or distorted (sugarcoated) [for example (136)], the message that the assessor intends to deliver may be very different from the one the trainee receives. While Ginsburg and colleagues have reported that trainees can crack the code and interpret the underlying meaning of assessors' comments written about other trainees (38), the ideal situation would be one where assessors can deliver their intended message.

The occurrence of MUM behaviours is important to recognise as this may signal that an assessor is experiencing difficulty with their assessor role and highlight a need for support or assistance for them. However, while alerting assessors to the existence of the

MUM effect and how it manifests is an important step and will help them to identify and understand their own MUM behaviours, this alone is unlikely to be sufficient to address problem MUMing.

Delivery without consequences

In addition to the MUM behaviours of delay, avoidance or distortion of a message, this study has highlighted two other ways in which reluctance to deliver a negative assessment message can manifest. Both involve delivery of a message, but in a more indirect way which avoids any negative consequences for the trainee (such as a failing assessment). These have the benefit of ensuring the assessor does deliver their message, but may avoid a harsh judgement of them by the trainee, by minimising a negative outcome. The first method involves delivering the message in a (mostly verbal) comment rather than a rating, and the second involves using ratings to send a message of poor performance.

As presented in the assessor questionnaire study, some assessors reported that they had failed to fail or grade inflated a trainee's results, and did nothing further, that is, they did not deliver their private thoughts at all; however others did communicate their private thoughts either in a (more commonly) verbal or written format. The review of submitted SITAs study showed that negative written comments were more common than negative ratings which may support the proposal that some assessors choose to deliver their negative assessment message in the form of a comment, which has no consequences, rather than a rating. As previously discussed, in this study setting the comments section of the assessment forms does not contribute to the outcome (satisfactory or unsatisfactory) of the assessment, and so they are effectively immune from a consequence. While assessors who chose to deliver a negative message through a verbal comment rather than a rating may feel that they have still effectively delivered their assessment, the lack of an official label to the problem (28) and the lack of ability of written or verbal comments to lead to formal support for the trainee means that, though delivered, they may not in fact promote trainee development or remediation. Further, where comments are less favourable than ratings, the former run the risk of being completely ignored. While this "off-the-record" practice may in fact not be considered to be MUMing — as the assessor's private thoughts are delivered to the trainee and they may believe this to be sufficient — it may not achieve the aims of these

assessments to promote the development of trainees, and ultimately their ability to care for their patients, since it avoids a public outcome or result. Thus while assessors may feel comfortable that they delivered their message, unfortunately, it may not result in benefit for the trainee.

The second method identified in this study that assessors used to deliver a negative assessment message without consequences was by delivering a message in their mark. For example, in the interviews some assessors reported the practice of giving a borderline “3” or one rating of “2” in the SITA with the hope that this would send a message to the trainee that their performance was of a poor standard; but again without leading to an immediate consequence for either the assessor or the trainee. As with providing a comment only, this method risks the message becoming “coded” but not always “cracked”, and at least one assessor reflected this was a possible outcome in their experience. While the review of submitted assessments chapter showed that comments and ratings were overall aligned, isolated examples were seen where comments and ratings did appear to send different messages.

Table 10.1 summarises how reluctance to deliver negative assessment messages manifest in this study and gives explanations of how these may present aside from those reported and summarised in the literature (111) — being DAD behaviours: Delay in or Avoidance of delivery and Distortion of the message.

Table 10.1: Summary of *how* reluctance to deliver negative assessment messages manifests from this study

<i>How does reluctance to deliver negative assessment messages manifest?</i>	
Behaviour	Additions to DAD behaviours revealed through this work
Avoid delivering the message	Includes delegating assessment duties to others
Delay delivering message	Potential to result in an improved quality of the message or done with the aim of minimising harm to trainee, but harmful to communication if excessive
Distort the message	Grade inflation and failure to fail in the extreme
Deliver the message without consequences	Deliver via a verbal or written comment – “off-the-record” Send a message through the mark

Why does reluctance to deliver negative assessment messages manifest?

The next section presents the results of the current study, which answers *why* reluctance to deliver a negative message manifests. The project has identified many reasons why assessors may experience reluctance, and the results contribute to understandings in the clinical performance assessment context in two ways. Firstly, they add further reasons to the previously identified categories, to complement and expand on those reported in the literature. As an example, the importance of the consequence or outcome of the message is seen here as being a key factor contributing to reluctance to deliver a message. While the outcome of a message is presented as a modifying factor in the MUM literature, its consequence, for *both* the assessor and the trainee, are highlighted in this work. Examples of such outcomes ranged from that of failing an assessment to the long-term impact on relationships in the small college setting.

Secondly, the results of this study suggest that reasons identified as leading to reluctance have differing levels of importance. That is, they vary in their relative impact between assessors and are not of equal priority in terms of their need to address. However, some were recognised to be salient to most and these were seen as central to the problem of reluctance to deliver negative messages. These relate to the fundamental perceptions and beliefs around the concept of “failure”, and the responsibility that the assessor feels for the delivery of the assessment message. This work proposes that these concepts are key to overall reluctance to deliver negative messages, and that failing to address them would limit the potential outcomes of work in other areas. For example, if all tool, training and logistics issues were satisfactorily addressed, while those relating to the perception that failure in this context is destructive and that assessors can control whether or not such a result is made public were not addressed, reluctance to deliver negative assessment messages would most likely continue to persist. In sum, if assessors overwhelmingly feel that failure in this assessment context is “devastating” and that they are responsible for whether this occurs or not (“how tough will you be?”), MUMing will continue to occur, despite how well the other issues are addressed.

As discussed in the MUM effect literature review, Tesser and Rosen presented reasons why reluctance to deliver negative assessment messages may manifest (125) and more recently, the failure to fail systematic review (80) identified six barriers to delivering a

failing result to trainees. Table 10.2 below presents these two works side by side to show how the MUM effect literature can explain some of the failure to fail barriers. The third column of the table presents a combination of these two categories, which shows how the results of this study are presented. To summarise, Tesser and Rosen’s “Communicators self-concern” is combined with Yepes-Rios and colleagues “Assessor’s professional”; and “Assessor’s personal” considerations are combined to one category – “Assessor – related”; “Communicator’s concern with the recipient” is combined with “Trainee – related considerations”; and “Communicator’s concern with norms” is combined with “Institutional culture”. Further, “Unsatisfactory assessor development and evaluation tools” is renamed to “Tools, training and logistic issues” and “Consideration of available remediation for the trainee” is expanded to “The outcome of the assessment”.

Table 10.2: How the MUM effect literature explains barriers identified in the failure to fail literature

<i>Why does reluctance of deliver negative assessment messages manifest?</i>		
MUM effect literature (125)	Failure to fail systematic review (80)	Combination of categories used to present current study results
Communicator’s self-concern • Guilt hypothesis • Fear of negative evaluation hypothesis • Mood hypothesis	Assessor’s professional considerations	Assessor – related considerations
	Assessor’s personal considerations	
Communicator’s concern with the recipient • Recipient’s emotionality hypothesis • Recipient’s desire to hear hypothesis	Trainee – related considerations	Trainee – related considerations
Communicator’s concern with norms • Norms proscribe transmission • Norms are ambiguous	Institutional culture	Institutional culture
	Unsatisfactory assessor development and evaluation tools	Tools, training and logistic issues (also includes some themes from Assessor’s professional considerations)
	Consideration of available remediation for the trainee	The outcome of the assessment

These categories or themes are referred to as the outer level themes and are presented first as they largely expand on our current understandings from the literature.

Discussion of the central or core theme identified follows this section. This is presented after the outer themes in order to demonstrate the relevance across all others, and to highlight how this therefore contributes to our further understanding of the reluctance to deliver negative assessment messages. Table 10.3 shows all themes reported through this work and figure 10.1 represents their relationship in diagrammatic form.

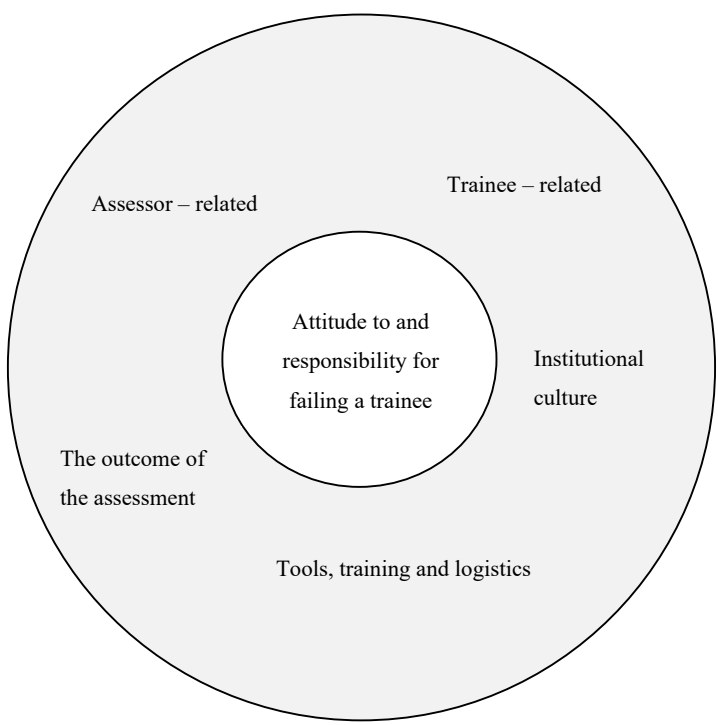


Figure 10.1: Diagrammatic representation of the relationship of identified themes. Central or core theme is fundamental with outer level themes varying in importance between assessors and contexts.

Table 10.3: Reasons *why* reluctance to deliver negative messages manifests in dermatology training in Australia (80,125)

Framework developed – <i>why</i> reluctance to deliver negative assessment messages manifests	
Outer level themes	
Assessor – related	
• Discomfort with delivery of assessment messages • Relationships • Impact of direct approaches	
Trainee – related	
• Assessors assume low trainee desire for and high sensitivity to feedback	
Institutional culture	
• Culture of giving negative information • Culture of lack of training	
Tools, training and logistics	
• Lack of time and opportunity for assessment duties • Insufficient assessor knowledge • Limits to assessment tools - Differences between assessment formats – WBA vs SITA - Number of assessments conducted by assessor with individual trainee	
The outcome of the assessment	
• Lack of other options for trainees unable to complete training	
Central theme	
• Attitude to and responsibility for failing a trainee - Fear of failure - Assessor responsibility for results	

Outer level themes

This section describes the outer level themes which answer the question of why reluctance to deliver negative assessment messages manifests. Five themes are presented: Assessor – related, Trainee – related, Institutional culture, Tools, training and logistics, and The outcome of the assessment.

As presented in the studies, demographics showed relatively few associations with behaviours. There were no statistically significant differences in the assessor self-report of failure to fail or grade inflation by years' experience as either a dermatologist or assessor, or by gender or SoT status. Similarly, trainees' views of the usefulness of their feedback or their sensitivity to it did not vary significantly with demographics. While this may suggest that results are equally applicable to all trainees and assessors, a

larger study may have revealed differences. At present however, both these results and the MUM literature indicate that, while people vary in the extent to which they MUM, the MUM effect is at least a widespread phenomenon and operated in the study setting.

Assessor – related

This theme describes factors related to the individual assessor which were seen to promote the avoidance of delivery of a negative assessment message. It incorporates the Communicator's self-concern reason as per Tesser and Rosen (125) and the themes Assessor's personal considerations and Assessor's professional considerations as presented by Yepes-Rios and colleagues (80). In this study, the impact of assessor discomfort, relationships and direct approaches by trainees are seen to be of particular relevance to the question of why reluctance can manifest.

Discomfort

Informed by the studies of discomfort published in the human resources and organisational psychology literature, this project explored the influence of assessor discomfort on reluctance to deliver negative assessment messages, finding a significant association. The assessor questionnaire study and the interviews showed that many assessors experience discomfort in their role and that this is associated with self-reports of failure to fail. These findings have important implications both for assessment outcomes and for assessors.

As presented in the assessor questionnaire study, the discomfort scores of those who admitted to failure to fail were statistically significantly higher than those who did not. Further, those assessors with discomfort scores above the median for the group of respondents to the questionnaire reported significantly more MUM behaviours than those below. As such, the impact of discomfort on the outcome of the clinical performance assessments in this context appears highly relevant and further exploration is important, not only for its apparent impact on assessment messages but for its impact on assessors as well.

As discussed, the impact that the assessor role has on individuals is particularly important, as the clinical performance assessments in this context are conducted by volunteer clinician assessors in addition to their usually demanding clinical workload.

As such, the impact of the role on the assessor's wellbeing is an important consideration as if it is too great, it may potentially jeopardise their health and wellbeing. If the impact of the role is seen as overall too negative, with the challenges of being an assessor outweighing the benefits, one sequela may be a decrease in the number of fellows who agree to participate.

Relationships

In addition to discomfort, other personal considerations related to the individual assessor were seen to be relevant. Consistent with earlier reports from this setting (12) assessors also reflected that at times, personal relationships influenced their assessments, whether consciously or not. Liking a trainee and willing them to do well were mentioned by some, as was the long-lasting impact that a “big fat fail” could have — that is the fear of negative evaluation. As mentioned earlier, this is particularly relevant for members of a relatively small college where Fellows interact, associate and meet, both virtually and face-to-face on a regular basis at meetings, conferences and college events throughout their whole career. As such, it may be that MUMing affects smaller colleges disproportionately more than larger ones.

Direct approaches

In this study, assessors were also seen to be impacted by a direct approach from a trainee, either to request a particular assessment outcome or to change a previously submitted one, and this was associated with the self-report of MUM avoidance behaviours in some assessors.

Many other previously reported reasons why assessors avoid identifying underperformance were identified in this study and seen to represent reluctance to deliver a negative message. Consistent with other studies (for example (279)), assessors here also described the profound emotional toll that failing a trainee has on them, often with great emphasis — “you feel terrible”, to describe the stress they feel after delivering a result. Similar to others also, some found that previous bad experiences or fears of bad experiences, would lead them to avoid doing so in the future (11).

Relatively few assessors discussed their fears of legal outcomes from their assessment but did describe that this increased their workload, for example through the increased work involved in documentation.

Further, assessors also described the significant burden in terms of time and effort that the assessor role places on them. Almost all in the interviews described needing to complete assessments in their own personal time and the impact this had on themselves and their families. Trainees also recognised the difficulties they face in arranging to complete assessments. These issues are further described in the section theme entitled Tools, training and logistics.

Assessors in this study did not directly discuss other self-concern reasons aside from the fear of negative evaluation, such as the cost to themselves, either through the need to adopt a negative mood in the setting of delivering a negative assessment message (the mood hypothesis); or feeling guilty due to inequity of their situations (the guilt hypothesis). Assessors' clinical experience and familiarity with the situation of giving bad news to patients and their families may offer a partial explanation for why reasons related to the mood hypothesis did not feature in this study. Though assessors recognised the clinical and educational situations were different, they may already have developed sufficient strategies to manage barriers, which they could then employ in the assessment context. While assessors did not directly discuss feeling guilty about failing a trainee, many did express feeling great empathy for the trainees — in particular their study load and other training tasks to complete, work and on-call responsibilities, and their personal life. While not overtly stated, it could be anticipated that this level of empathy for trainees may have contributed to avoidance behaviours by some.

Similarly, the failure to fail systematic review (80), presented a wider range of themes in this category than were seen in the current study. Aside from issues of time and limited exposure to trainees, minimal other reasons related to Assessor's professional considerations presented in the current study. In particular, as assessor roles in the study context are voluntary, issues relating to tenure and reappointment that are identified in the literature were not relevant and so not raised here.

Trainee – related

This theme describes factors related to the trainee which were seen to promote the avoidance of delivery of a negative assessment message. It incorporates the Communicator's concern with the recipient reason as per Tesser and Rosen (125) and the theme Trainee related considerations as presented by Yepes-Rios and colleagues

(80). Assessors' concerns for the trainee are highlighted in particular in the current study.

As outlined, many assessors reported awareness and understanding of the immense pressures faced by trainees in the training program and many seemed reluctant to contribute to these with the delivery of a negative assessment message. Several assessors directly discussed their concerns around the potential impacts of a failing assessment (including the Fellowship exam) on a trainee — including the risk that this could lead to suicide; unsurprisingly, this caused assessors considerable stress. Both assessors and trainees talked about failure as being embarrassing but there was minimal talk about trainees actually being exited from the program.

As described in the interview study, some assessors held views on the best time to deliver negative assessment messages to the trainee and this was a factor in their level of willingness to deliver the assessment message. However, opinions varied on when is: some felt that the junior years were the best time to deliver negative results, while others felt the senior years were.

Similar to other studies, further reasons for reluctance seen in this study included assessor fears of the potential trainee reactions to a negative assessment result — “they don't take it well”. Some described having faced negative reactions from trainees previously and expressed concerns about their possible reactions consistent with Tesser and Rosen's Recipient's emotionality hypothesis. However, the study results suggest that the trainee perspective contrasts with the Recipient's desire to hear hypothesis. The level of trainee agreement with items on the sensitivity subscale of their questionnaire was not overly high, suggesting the majority feel not greatly impacted by negative assessment information. Consistent with the MUM effect though, assessors seemed to largely assume that trainees did not wish to hear negative news which contributed to felt reluctance.

Institutional culture

This theme describes how the educational institution and culture impacts on the reluctance to deliver a negative message. In this study, the influence of the educational institution on assessor reluctance was not strong. For example, in contrast to other

literature (as summarised in (80) the undersupply of dermatologists was not offered as a reason for not delivering negative assessment messages (280). Nor were potential risks to the ACD, in terms of financial implications or reputation (10), or that decisions would be overturned by them (13,92). In fact, the opposite was seen and many assessors spoke of the support received from the ACD and/or their particular health service when they were in the situation of supervising a struggling trainee and needed to deliver a negative assessment message to them. Further, in the study setting, the expectation that trainees would not be marked competent on their first attempt at a given WBA is highlighted in the assessor training modules from the ACD, which perhaps contributes to the lack of institutional influence on reluctance to deliver negative messages seen in this work.

However, while the culture at the institutional level does not seem to have promoted MUMing, the culture of medicine in a wider sense was certainly observed and this echoes Tesser and Rosen's Norms proscribe bad news transmission explanation. In the assessor questionnaire and interview studies, some assessors recognised the failure of some of their colleagues to address trainee underperformance, citing the tendency of some assessors to avoid the issue as the rotating term program would ensure that it would become "someone else's problem". Some assessors were aware that previous assessors or rotations had not appropriately failed a trainee, thus broad concepts about "failure" are seen to be highly relevant and will be discussed further below.

In this study, not all assessors expressed a need or desire for further training. This perceived lack of need for training may stem from the common "see one, do one, teach one" culture that exists in medicine, which may foster the expectation that all clinicians will be teachers and assessors. Indeed, in the past, this occurred without any formal assessor training or support. Some assessors may be unaware of the accessibility of the ACD assessor modules that are currently available (and completion rates of these modules are noted to be low), nor the now extensive array of clinical teaching courses (281) available to support them. Though some assessors may feel they are sufficiently equipped for the role, others may believe the benefits of such training may not outweigh the costs of their time (and financial investment for external courses). This work highlights that a lack of assessor skills and knowledge can contribute to reluctance to deliver negative messages if the culture promotes assessing without training.

Therefore, in this study, the institution did not present as a barrier to reporting underperformance but some elements of the broader culture were relevant.

Tools, training and logistics

This theme describes the elements relating to the design and implementation of the clinical performance assessments which impact on the reluctance to deliver an assessment message. Many elements from the Unsatisfactory assessor development and/or evaluation tools theme highlighted in Yepes-Rios et al's review (80) were also reported by participants in this study and are of key relevance here, in particular, those relating to the practicalities and logistics of conducting the clinical performance assessments. For this reason, some themes in the category Assessor's professional considerations are included here also. The current results show they are relevant for both assessors and trainees.

Lack of time and opportunity for assessment duties

Assessors strongly reported feeling they lacked time for assessments, and trainees highlighted the difficulties they face in arranging to complete them, and both echo reports from the literature (8). In addition, the barrier that the "huge" amounts of time and effort involved in failing a trainee presents to assessors — many of whom are already overwhelmed by their clinical and administrative workloads — as well as multiple other issues such as changes in work practices and loss of autonomy that many face in their departments, were reported. While such themes may be well recognised, they remain unaddressed in many settings to date, including the present one. The results from this work further extend the literature, whereby several assessors communicated just how pressured and at times disillusioned that they feel, both due to this workload, and feeling a lack of support and appreciation in addition to pressures from other staff to resolve issues of trainee underperformance, when at times they did not have the resources, including knowledge, to do so. When faced with insufficient resources to cope, desired outcomes suffer.

The lack of allocated time for assessment, meant that assessment duties were often conducted or completed in personal time, and many assessors discussed the pressure this places on them, and in addition on their family and personal lives also.

Overworked and overwhelmed assessors may choose not to deliver a message which will have the effect of increasing their workload further.

Some trainees also reported feeling pressure in relation to assessment logistics in their questionnaire responses, and some assessors were aware that trainees could be strategic in the selection of cases presented for assessments. Trainees described reasons for such behaviours — such as wishing to avoid embarrassment or the wasting of the assessor's time with a poor performance. While this may not seem directly related to assessor reluctance to the delivery of negative assessment messages, it shows how assessment design and implementation can influence some trainees to behave in a way to try to avoid a negative assessment message due to the pressures they felt. As highlighted, many assessors showed great empathy for the trainees' situation and workload, and their awareness of such actions may have influenced their behaviours and contributed to reluctance to deliver a negative message. As some trainees reported approaching an assessor in relation to assessment outcomes, either to encourage them to give a certain rating or to change a previously submitted one; this too may indicate the pressures that trainees feel in relation to the completion of these assessments. The potential impact of such approaches on assessor behaviours was seen with those who reported approach, also reporting a greater engagement with MUM avoidance behaviours compared with assessors who were not approached.

Insufficient assessor knowledge

In this study, several assessors reported feeling that they lacked assessor development or training, and called for assistance in this area, though requests were not consistent across all assessors. Consistent with previous reports (10,75), some assessors expressed a lack of understanding of assessment standards, particularly with trainees at different levels of training. However, in contrast to this, there was also a strong feeling amongst many participants that they inherently “knew” the true standard of the trainee. Perhaps what assessors actually meant was that they don't know how to justify or defend their decision or deliver their message in a way that won't be questioned by the trainee. That is, they have a gestalt view of the trainee's level of performance but are less certain of how to translate this into a potentially defensible public judgement. Therefore assessor reports of feeling they lack training may more specifically relate to a lack of skills in delivering and justifying messages, and dealing with the situation where a trainee reacts

negatively, rather than a lack of knowledge of the standards. This is consistent with one interview participant's desire for simulation training (roleplaying) in challenging assessment scenarios.

While many assessors expressed that a common or shared knowledge of the trainee's standard of performance exists amongst them, in the interviews, several assessors however admitted to being uncertain of or not knowing the criteria for failure on the assessments, and the chapter which reviewed previously submitted assessment forms supports this. In this study, it was seen that a number of forms (18.4% or 29 of a total of 158) had been submitted to the ACD with the incorrect criteria for unsatisfactory performance documented on them. That is, they had been submitted without apparent question or recognition of the error. One explanation for the large difference between the number of borderline ratings given in each group is that assessors did read the (incorrect) definition for unsatisfactory criteria on these forms but did not have the knowledge to recognise this as erroneous, though did seem to change their rating behaviour because of it. The error was only detected when one assessor recognised that a trainee had reached the criteria for unsatisfactory performance which was printed on the form leading to the subsequent questioning of the definition recorded. This lends support to the argument that assessors, who at present volunteer for this role and are overwhelmingly committed to it, face too many pressures and insufficient supports in their role.

Limits to assessment tools

Results from this study show that assessors' views of the assessment formats represent further reasons for reluctance to deliver negative assessment messages. The importance of engagement with assessments has been highlighted previously (102) and the MUM effect predicts that feeling an obligation to deliver a message can increase the likelihood of doing so. In this study, while most interview participants felt a deep obligation to be involved in the assessment of trainees, which may be expected to include the delivery of assessment messages, this did not always translate into the practice of delivery of their private thoughts on the trainee's performance. In the interviews, assessors who voiced negative opinions about the assessments or doubted their ability to effect change at times showed reluctance to deliver difficult messages and this extends the literature by highlighting how such a lack of engagement can affect outcomes. The results further

showed that some assessors found difficulties with particular assessment formats over others. For example, some reported it was more difficult to deliver a negative message in a SITA as compared with a WBA. The number of assessments that either the trainee had completed in a period or the number they had completed with the individual assessor are further factors shown to make delivery of negative messages potentially difficult for assessors. Compared with the literature (11), lack of documentation was not commonly cited by participants as a reason for reluctance to deliver negative messages in this study, though there was occasional mention that improved methods of documentation (including technological advances) could help.

Importantly, some assessors expressed that the requirement to deliver the assessment message immediately after the assessment at times led to reluctance and avoidance of delivery. They noted a delay helped them to deliver a more honest and considered assessment message, and this is consistent with other studies that have shown that a delay in the delivery of an ITA result is associated with the delivery of lower ratings (59).

The outcome of the assessment

As described, not all themes previously identified in the literature as relevant to difficulties with delivering negative assessment messages were identified in the current study. For example, while the realisation of the potential benefits of identifying and reporting underperformance was discussed by some in other studies (75), it was little mentioned by assessors or trainees in this assessment context. Lack of remediation options were not discussed nor were the opportunities for the assessment message to lead to trainee remediation and growth. Further, there was minimal discussion of the ability of these assessments to prevent future issues including at the Fellowship exams or more importantly, with a future patient. Therefore this study highlights that perhaps assessor training needs to be more nuanced — not just in how to deliver bad news, but also in emphasising why it is important to do so, including an appreciation of the benefits of remediation or where this is not possible, the option of leaving the program before too great an investment has occurred or a severe outcome occurs.

Some assessors discussed that having other options for work for trainees who could not complete the program would be of value, with the possible implication that knowledge

that there were other options for them may help some to deliver a negative message when appropriate.

From the trainee perspective, the questionnaire study showed that while some trainees highlighted views of the importance of the assessments and their focus to do well, this seemed to be often motivated by a desire to avoid wasting the assessor's time, likely driven by the other practical and design difficulties and pressures outlined above, rather than for the benefits to their own performance. Indeed, there was no mention by trainees of the ability of these assessments to promote their development. It seems that in this context, the clinical performance assessments as they currently stand are often not viewed by either assessors or trainees as learning opportunities, rather largely as performance ones to have completed in order to move on with other tasks.

Central theme

Attitude to and responsibility for failing a trainee

The development of the clinical expression of disease is often multifactorial, and similarly here, reasons behind a reluctance to deliver negative assessment messages are seen to be multiple, with some more important than others and each likely affecting an individual assessor in a different way. However, this study has revealed that two common reasons lie entwined throughout most of those presented so far: firstly, that “failure” in this setting is viewed by assessors and trainees as a very negative event; and secondly, that the delivery of the assessment message depends on the assessor and is therefore malleable: trainees and assessors can both potentially influence this. Trainees, through strategic choice of case and assessor or through approaching assessors about their results; and assessors by choosing whether or not they deliver the message. Thus “Attitude to and responsibility for failing a trainee” is proposed as a core theme as these concepts are felt to be of prime central importance in why reluctance to deliver assessment negative messages manifests.

This theme recognises the profound and prevailing negative view of failure in this assessment context, and that it is the assessor who decides whether to deliver a negative assessment message or not. This contrasts with other assessments in medical education, such as a written exam, where a marking rubric often exists to guide assessors and

marks are documented and may be reviewed at a later time, or a long case exam where two examiners discuss a trainee's performance with reference to a marking guide. By their design, in the clinical performance assessment context, the assessor's private thoughts on the trainee's performance may never need be made public and if the consequences of doing so are felt to be too great — that is that the view of failure is too negative — MUMing may become the preferred option.

A profoundly negative view of “failure”, in combination with a lack of belief that any beneficial outcome may result from it, presents too large a barrier to the delivery of negative messages. Further, the belief that the exam will “do the job” anyway or that the trainee will move on and become “someone else's problem”, seems to reduce the impetus to act. Compared with the risk of a negative outcome, the possible benefits of delivery seem simply too small and were rarely mentioned in the study. Where this occurs in combination with an underlying lack of practical and logistical supports available for the clinical performance assessment process, this has the effect of communicating that they are not really important or valued, for if they were, there would be dedicated time and space in which to conduct assessments and assessors would be better supported and recognised for their time and efforts. Assessors and trainees overall seem to view these assessments from a strongly performance goal orientation as opposed to learning goal one, so any corrective or negative comments or ratings are seen as a negative event, rather than an opportunity to improve and develop into a better clinician. In sum, this promotes a lack of delivery of negative assessment messages.

Implications for practice

General considerations

The extensive body of work conducted in the area of failure to fail and grade inflation to date, coupled with their persistence, speaks to the complexity of the problem. This work contributes to furthering the understandings of these problems in two ways. Firstly, it broadens the perspective on the issues of failure to fail and grade inflation by emphasising that determining ratings and grades are only one component of the problem, and that focus on and improvements to the numerical aspects of an assessment will not lead to resolution of all issues. Such focus likely explains why much work in

the area to date has concerned attempts to improve ratings through interventions such as rater training and form modifications, but this study shows that the narrative comments are important also. They may also be affected by a reluctance to deliver negative messages, seen where some assessors chose not to give any; or they may signal the presence of reluctance to deliver accurate ratings, as seen on occasions when some assessors gave a verbal or written comment that reflected their true impression of the trainee's performance, while their ratings did not. Secondly, this work has shown further influencing factors that affect the assessor's level of motivation and willingness to deliver their assessment message — especially that of discomfort with the assessment process. This work has highlighted the importance of assessor discomfort and their overall views on “failure”, coupled with their sense of responsibility for delivering the result. In combination with the MUM effect literature, this has permitted a deeper appreciation and understanding of the phenomenon of failure to deliver private thoughts, being related to both the numerical ratings and narrative comments about a trainee's performance, on formative and summative assessments. This in turn allows for new insights on how to approach the problem to be developed.

As this work seeks to expand the thinking around the issues of failure to deliver assessment messages and consider it as a problem which may arise from a widespread human tendency, it perhaps risks making the issue seem even more complex and a resolution more challenging to reach. As with other currently recognised complex problems, such as wellbeing, it may be that multiple interventions, each contributing perhaps only a few percent of effect, will be necessary. It is further likely that assessors will show individual variation in the impact of each intervention upon them. Further, the recognition of the complexity of the issues around the delivery of negative information shows that they may not be possible to eliminate completely, and as such, some amount of acceptance and enhanced recognition of issues may be required.

As mentioned, many previous interventions have focused on addressing the barriers to the delivery of messages that have been identified in the failure to fail literature (80). However, enhancing enablers is also important and likely to be an effective way to modify assessor behaviour and so promote the delivery of appropriate assessment messages.

In the following section, the implications of this work and then opportunities for the future are presented. A discussion of general implications of this work is presented first and this is followed by specific implications for dermatology training in Australia.

Using the framework from the failure to fail systematic review (80), combined with relevant modifiers from the MUM effect literature [summarised in (111)], the first part of the section discusses two possible enablers which may help assessors to deliver their message. Firstly, those which relate to a sense of duty are discussed, followed by factors which support assessors. The impact of discomfort follows this along with a discussion of the implications of the barriers identified.

What helps assessors to deliver their assessment message

Duty to patients, to society and to the profession (80)

The implications of this work can be considered with reference to the study results and in light of the failure to fail and the MUM literature. In their systematic review, Yepes-Rios and colleagues (80) report that “[r]esponsibility towards patient safety, duty to the public and moral integrity were leading themes in supporting assessors in failing an underperforming trainee” (80). The MUM literature similarity highlights modifiers which may promote delivery of a challenging message and the combination of these has potential implications for dermatology training in Australia and beyond. As the current study focused beyond failure to fail to look more broadly at the overall concept of reluctance to deliver negative messages, it was seen that assessors expressed both reasons which related to themselves or the trainee as explanations for the times when they did not deliver their negative assessment message. These results showed less indication that enabling factors, such as the sense of duty to the profession or to patients, currently help assessors in their role. This is an important area to consider and several options are possible which may serve to increase a sense of duty as enabler; these are outlined below.

As previously described, the MUM literature reports that role-related factors, such as felt obligation and satisfaction in a role, can impact on the likelihood of a teller delivering a message. As such, raising the profile or status of the assessor role may go a significant way in promoting an assessor’s awareness of their duty to patients, society

and the profession and so assist them to deliver their assessment message. While most assessors interviewed in this study did show a deep level of obligation and commitment to their role, this did not necessarily result in their delivery of negative assessment messages when required. Working to increase the profile and identity of the assessor/supervisor role, beyond one of just obligation, and into a subspecialty role, is a potential area to consider. Options for this could include the formalising of assessor appointments and building on the information provided in role descriptions. For example, assessor role outlines could emphasise the role and duty of the assessor to provide appropriate teaching and assessment, and the implications that this has individually for the trainee and more broadly to society, the profession and importantly, patients. Assessors should be encouraged to carefully consider how the duties of the role align with their self-impression of what they can realistically achieve. For example, the literature discusses that some assessors identify a conflict between the “supporting versus assessing” role (10) and the problem with being a coach and a judge (24). Separating the assessing role may be a further outcome and may serve to increase commitment to the teaching of trainees without the stress of needing to also assess them. A subgroup of consultants who specialise in assessment and feedback (53) could also be developed, similar to Bing-You and colleagues’ suggestion to “[l]imit the number of well-trained faculty providers” (97).

Beyond a focus on duty, assessor commitment and engagement could be further enhanced through the development of assessor communities. Such communities could develop champions to foster the assessor role and encourage others. As assessors may be separated geographically or unable to meet physically on a regular basis due to other commitments, online options could be designed, supplemented with face-to-face meetings when possible. An online community could facilitate communication on assessment matters (for example, in regard to meetings or updates), store resources, and contribute to the education of assessors. This could also serve to mediate the many uncertainties and misunderstandings expressed by assessors in the current study, which may also apply to other settings.

Many assessors spoke of the time, effort and commitment they devote to the assessor role, and formal recognition and expressions of appreciation for this is important, and supported by Moffett and colleagues (269). Local solutions to address this could be

developed and would also contribute to role development and identification. As further recognition of the important contribution of the assessor role to the profession, a review of the Continuing Professional Development (CPD) points that are available for teaching and supervisory activities should be conducted to ensure that they accurately reflect the time and effort that supervisors and assessors contribute and so reward their efforts.

Emphasising and enhancing the duties and obligations of an assessor role alone will not however eliminate the difficulties associated with the delivery of negative messages, and the MUM effect predicts it will not be possible to entirely eliminate this. However, the results of the study and the literature highlight other factors which have the potential to help assessors to deliver their message. As identified by Yepes-Rios and colleagues (80), support of the institution, colleagues and training are further key factors and the results from the current study further suggest that practical and peer elements are important also.

Support from the institution, from colleagues and assessor development (80)

Increasing the sense of role identification that assessors feel, in addition to attention to appropriate training opportunities and ensuring they view the assessment system as strong, with clear criteria, policy and procedures is necessary, and the results of the current study provide some suggestions as to how to proceed.

This theme considers how supports can enable assessors to deliver this assessment message, even when negative. In regard to the first aspect of this theme — support from the institution — while respondents in this study did not express feeling a lack of support from either the ACD or their employing health service nor concerns that their assessment judgements would be changed or ignored, the literature notes this is a concern at other sites (for example (13,92)). Even so, while not identified as a problem in this study, enhancing access to assessor support may serve to normalise and encourage the seeking of assistance by them. As some assessors recognised that previous assessors had not appropriately delivered negative assessment messages to trainees and many spoke of the need to involve others, either directly or indirectly in the

assessment process, so enhancing the support provided from others, including colleagues, represents a further opportunity for intervention.

One of the broad themes to be seen from this work relates to support from colleagues, and in particular, that assessing should be considered as a team effort, involving more than one assessor. Placing responsibility for decisions onto individual consultants is unduly onerous and likely contributes to the discomfort that many feel, resulting in MUM behaviours. Many assessors would no doubt benefit from the support of others, not only in terms of the provision of assistance with difficult decisions, but also by calibrating their decisions to ensure that they are fair and defensible. Further, requiring the input of a team for assessment decisions may yield other benefits, such as identification of more information about trainee performance. This was noted in the interviews and is well-reported in the context of the MSF and in-training assessments, where the value of the rating information on trainees that peers and other members of the health care team can provide has been observed — for example, Gray describes how peers can provide useful rating information “in that they provide insights not necessarily identified by [assessors]” (229). Group decisions are currently used in many other educational settings including dermatology (and other medical specialist) training in the UK (282,283) and also in programmatic approaches to assessment [for example (44)].

Implementation of a programmatic assessment model is a promising area to consider and offers aspects which would address many of the issues identified in this study. Such models offer advantages in addition to those of group decisions and have been employed in many medical education settings to date (284). They recognise a longitudinal and combined approach to assessment, and reduce the focus and weighting on individual assessments by the recognition that multiple data points on trainee performance are required: “a single assessment or even several equivalent assessments will not provide information for a complete picture” (284). Through the use of multiple assessments and multiple assessors, the individual strengths and limitations of each are seen to be balanced. As such, the potential lack of standardisation of the clinical performance assessments (for example, the absence of structured marking rubrics or controlled circumstances) does not necessarily negate the value of the assessment data — rather validity lies in the multiple samples and perspectives of assessors. A key part of this approach is also the separation of assessment episodes from an outcome or

progress decision. As such, individual assessments do not simply numerically add up to a pass or fail, but rather provide cumulating evidence for decisions and guidance to trainees in terms of how to proceed on their learning path. Assessment decisions are therefore not needed after each event, rather are “deferred until there is sufficient evidence from a variety of sources” (284). This model could assist many assessors concerned about the possibly disproportionate impact of documenting occasional instances of poor clinical performance, and delaying the narrative comments may be of benefit to those who recognised the need for some time in order to help them better formulate their opinion on a trainee’s performance. For example, in the interviews some reported that delaying the delivery of their message to trainees was often beneficial: “then I usually have something much more helpful to say (A8)”. The separation of an assessment and decisions also encourages the concept of learning from an assessment — that is assessment for learning — to ultimately aid the development of trainees. As also seen from the interviews, some assessors reported that they would stop a WBA if they felt was going poorly and change from an assessment to a learning exercise.

However, all assessment methods have challenges and implementing a new program, with a change from an assessment focus to a learning one, requires a significant cultural change for an organisation (285). Other aspects are relevant also, including costs, and the use of group decisions have potential implications of their own (286) which introduces further aspects and issues that will need to be considered. The workings of groups have many areas that require further understanding and exploration — as one example “trust in group decisions has not been well studied within or outside of health professions education” (287). Further studies in how individual, group and environmental factors influence the decisions the group makes, so to ensure that they are in the best interests of the trainee and the patients they care for, would be required (287). Effective group decisions necessarily require that members are able to work with others, resolve differences in opinions as necessary, contribute evenly to the workload, and act fairly. Furthering our understanding of trust in relation to competency committees would help to ensure that the work produced and the decisions made by them are as effective as possible, and assurance that assessment data and individual opinion are appropriately translated into assessment decisions is necessary (287). But

ultimately, with all potential options, it is the delivery of the assessor's message, even when negative, that is key to consider and is the focus of the work here.

A further way in which assessors may enhance message delivery is through the support of their colleagues or peers, though this must be done with attention to both privacy concerns for the trainee and local policy about discussion of performance (e.g. feeding forward). As discussed earlier, where one does not exist, the development of a community of assessors would be an important initiative which could be both face-to-face and/or virtual. While regular meetings could serve to foster role-identification and a sense of belonging to a group, they could also provide assessors with support from other peers. In addition, regular discussions with other assessors in the local area could increase the identification of areas of concern with trainee performance.

In addition to providing shared or collective decision making, other assessors can support each other through other means. Peer review and support strategies represent such examples and could address two areas at once — peer support and assessor training — through a process of review and feedback on assessors' practices. Campbell and Wozniak described a process of observation and review of supervisors' practices as a method to provide peer review of a range of teaching and supervisory practices; and while this was developed as a teaching development initiative, in part due to the recognition that learners may be limited in the feedback that they may provide to their supervisor on their practice, the method could be further developed and focus on the delivery of assessment messages in particular (288). The process of peer review involves the observation of a peer's supervisory practices with a trainee, followed by the provision of feedback to the supervisor. The interaction involves the consent of all involved, in particular the trainee, and it should be emphasised to them that the process is a review of their assessor's performance and not their own. The process, which is designed to be purely formative, and not evaluative or for credentialing purposes, has the advantage that it may involve any health professions peer as the observer: they need not necessarily be an expert either in education or in the same field as the clinician being observed, which increases potential options for participants. In fact, choosing a peer from outside of the assessor's particular speciality would have additional benefits including the opportunities for cross-pollination of ideas and practices from other fields and the fostering of collegiality amongst different craft groups. These in turn have other

implications for areas as diverse as work satisfaction and wellbeing, improved relationships and even patient safety (289). While direct observation of the peer is ideal, dispersive options (such as video-link) are also possible (again only with the prior consent of all involved), which makes the process accessible to assessors who may be geographically isolated. If the assessor to be observed identifies particular issues with MUM behaviours as an area for their personal development, the evaluation form used and feedback session could be designed to focus on these areas in particular. As mentioned, such a process would require attention to details such as the permission of the trainee to whom the feedback is being delivered and issues regarding their and patient confidentiality would apply. It should also be clear to all involved, as highlighted by Wozniak and colleagues that the observation of the assessor's performance focuses on their educative behaviours and not their clinical ones. Further benefits of this process would be improvements to the assessment performance of the observer, through their reflection on performance, and to simultaneously highlight the importance and value placed on the clinical performance assessment process — to trainees, assessors and other stakeholders.

The final component of this theme — Support from the institution, from colleagues and assessor development — relates to assessor development and several, though not all, participants in this study called for further training. To explore this further, including the possible benefits, an appropriate initial step would be to conduct a needs analysis with assessors to explore their perceptions of the current training available and identify gaps between this and the training needs identified. Depending on the outcome, collaboration with other colleges and training or educational organisations may be of value to enhance and broaden the range of training available, including an expansion of the types of training beyond that of traditional workshop or webinar session formats. For example, simulation training was mentioned by one interview participant and this represents a potentially valuable option to explore. The role of simulation in learning and practising clinical skills in a safe environment is well described (290) and it is ideal for practising the delivery of negative feedback to a trainee. Simulation scenarios could be developed to allow assessors the opportunity to rehearse a range of difficult scenarios and would be of potential value to all participating — those acting in the simulated assessor role as well as those observing. Lyon-Maris and Scallan have described a technique they developed for GP trainers in the UK around giving feedback to a

simulated GP trainee (291); and scenarios specific to assessment situations could be developed, with multiple variations and levels of difficulty possible, depending on the needs of the participants. The principles of simulation training would apply, including the debriefing of participants after a session. The development of relationships with other educational institutions could identify further training opportunities, allow for sharing of resources and potentially collaborative research opportunities.

However, even with access to adequate assessor training, it should be recognised that not all assessors will develop the skills and experience of that of an expert in education, meaning that additional external support will likely be required in the long term. Assessors should be encouraged and expected to identify the need to engage such resources in a similar way to engaging another clinician for the optimal management of a patient where their needs extend beyond the practitioner's level of competence or experience.

While the areas of support outlined above — from the institution, from colleagues and from training — are important to address, those which relate to the practical and logistical aspects of assessment are key also. These aspects were highlighted in the current study and the literature, and represent further opportunities to enhance the assessment process and help assessors deliver their message.

Practical – logistical support

While logistical support was not initially presented as an enabler in the failure to fail systematic review (where it was presented as a barrier in the theme Assessor's professional considerations) (80), it is considered here as it logically follows on from considerations of other methods of providing support to assessors.

In this study, most assessors cited a lack of allocated time as a significant impediment to their assessment duties and this meant that most were performing such work outside of their paid work hours. This represents a huge imposition for the assessor and relies on their (and their families') goodwill, and risks being unsustainable in the longer term. Moffett and colleagues (269) echo the need for regular protected time for teaching activities from a wellbeing perspective and in addition, the lack of time dedicated for assessments within working hours implies that they are not strongly valued. It is

relevant to recall that many of the participant assessors in this study (especially in the interview study) did not experience the WBA as a part of their own training, so have no personal experience as to their value. Indeed all assessors will have their own views on the assessments, informed by their previous personal experiences as a trainee (285) and these can affect their assessment practices. One important way to improve the outcomes of the assessments, would be to demonstrate the value placed on the assessors through the provision of protected work time in which to complete them, and others have called for time for both the assessments and the development of assessor skills (292).

Further options to reduce the pressure on assessments could also be explored. As examples, dedicated CbD clinics have been described by Deepak and colleagues (213); and the employment of consultants dedicated to the role of conducting assessments described by Jones and Nanda (272), but these require financial input.

Similarly to the issues of time, providing space for assessment and feedback discussions is important both to improve the processes and elevate the standing of assessments. At present, in the study context, discussions often occur wherever possible — there is usually no dedicated space. Such corridor conversations are frequently rushed and lack privacy. While space may be found in an empty consulting room, these areas are prone to interruption and understandably, will be appropriated for patient needs when required. Further, they are limited in other sites such as on wards or in operating theatres. While this is not easily addressed, perhaps at least documenting a plan for where assessment discussions should occur at the start of a rotation would serve to further increase the profile of the task and so the importance with which these assessments are viewed. At present, the often ad-hoc scramble for a place in which to conduct an assessment discussion likely contributes to the impression that clinical performance assessments are not important, or are at least a lower priority than clinical roles, and so compounding the problems faced by them — in other words, a form of the hidden curriculum (293). A quiet, uninterrupted space needs to be found by each site, away from clinical area ideally.

The next section discusses the additional barrier of discomfort strongly highlighted in this study and the implications it may have for the delivery of assessment messages.

The impact of discomfort

In this study, assessor discomfort is seen to have two important implications. Firstly, it potentially affects the assessment message which is delivered; and secondly, it can affect the wellbeing of the assessor who experiences it. The importance of further exploring associations between discomfort and assessment outcomes will be discussed in the section on further research agenda. The impact that discomfort in the assessor role can have on the individual will be discussed below.

Assessor wellbeing

One of the clearest findings from this study has implications for assessor wellbeing, which is an area of current and increasing focus and work (269). The questionnaire and interviews clearly showed that some assessors experience high levels of discomfort and distress as part of their role. This is an important finding — both for the barrier to the delivery of assessment messages which results and for its impact on wellbeing.

Training organisations should be encouraged to explore the impact that the assessment role has on the wellbeing of assessors in their own settings, using methods such as those used here. Acting on and re-evaluating the results should then follow. Tools such as the PADS could be used to point to particular components of the assessment duties that assessors find challenging, and pooled results could be used to further understand how discomfort affects assessors. This could lead to the implementation of targeted training or additional supports in these areas, with their effect measured through periodic re-administration of the tool. Collaboration across training organisations could be of value in order to increase the effective resources, such as expertise available, in order to minimise the distress and discomfort assessors experience.

As some level of challenge in the assessor role will likely always be present, methods to increase supports to assessors are important to consider also. In this work, participants highlighted some options to address discomfort which may have relevance for other settings. For example, the opportunity for assessors to debrief and the availability of a senior to act as a mentor were mentioned and these are important areas to explore. Such supports would not necessarily have to be a clinician in the same area, either clinically or geographically. As with peer review outlined above, an assessor from another

speciality or health professional area could also be appropriate and potentially foster other benefits.

The literature highlights some recent works which consider the issue of assessor wellbeing and may offer potential benefits to the current and other settings. Moffett and colleagues recently explored issues around assessor wellbeing, highlighting the life as a clinical educator as a “double-edged sword” (269), due to conflicting demands and roles and the potential for this to drain the assessor’s resources. This aligns with comments from interviewees in this study, where several reported they enjoy many aspects of the role, but feel the immense pressures of others. Moffett and colleagues make many practical suggestions aimed at addressing and promoting assessor wellbeing. They stress the importance for individuals to attend to their own wellbeing needs [which they liken to the concept of “putting on your own oxygen mask first” (269)], though they recognise the challenge this may present to naturally altruistic people, who are used to suppressing, delaying or denying their own needs. Thus they highlight that responsibility for wellbeing needs to be implemented at higher levels with system level changes. They emphasise that addressing assessor needs must extend beyond that of attending to individuals only, such as with mindfulness courses, and argue that this in fact risks “implying fault [is] with individual staff members” (269). It is important to recognise that individual centre’s needs will vary and that not all recommendations of Moffett and colleagues may not be logistically possible or even necessary for all sites. For example, many unfortunately are not applicable to the study setting and the dermatology training set-up as most consultants are sessional only, so may not have time to make use of relaxation work spaces or exercise facilities at work. However, simple interventions, such as providing time for a tea break during an outpatient session, could be considered and implemented, and the resulting few minutes of relaxation and restoration as a team would no doubt serve to foster relationships and support for all, but also benefit patients, who would then be seen by doctors who have addressed some of their own needs (269). While research supports the addition of some wellbeing initiatives at the individual level [for example, high school educators who participated in mindfulness training were found to achieve significant benefits in a range of areas (294)], as an isolated event, the benefits to assessors would be limited if they are not provided with the practical and logistical supports available to assist them. Such logistical support to carry out the assessor role, as discussed earlier, is a

fundamental part of supporting assessors and reducing discomfort, and arguably the first area necessary to consider.

The benefits of connecting with others and the role of the communication of gratitude to educators has also been emphasised by Moffett and colleagues (269), and this was mentioned by some participants. Further Moffett and colleagues note the importance of other factors which affect wellbeing such the negativity bias and a lack of disconnect from work, and the possible benefits to wellbeing of activities such as journaling and engaging in and acknowledging simple pleasures (269). Other suggestions to address wellbeing could also be determined through a needs analysis conducted with assessors (combined with their educational needs as outlined earlier), again followed with an implementation and evaluation cycle. While there is currently a much-needed focus on trainee wellbeing, the wellbeing of assessors and those training trainees is also an important consideration and attending to this would no doubt enhance the wellbeing of trainees also.

The assessor role

Exploring why the assessor role should lead to feelings of discomfort in many is an important consideration also, especially since a goal of these assessments is to provide trainees with helpful guidance on their performance. Challenging and reframing concepts around the idea of “failure” in this assessment context are important areas to address in order to reduce the reluctance felt around the delivery of negative or corrective assessment messages. As a major theme from this study, the ideas around failure in many ways, interacted with many of the other reasons for reluctance. As presented, both assessors and trainees conveyed largely very negative perceptions around poor performance or failing — ranging from it being embarrassing, wasting time and potentially harming future prospects, through to it being devastating and a potential precipitant of severe outcomes such as suicide. Few spoke of the benefits of identifying poor performance and the opportunities for growth and improved performance it may lead to, including the potential to avoid failure at Fellowship exams. Similarly, the obligations and benefits of striving for excellence for the purpose of patient care were rarely discussed.

Open and honest discussion between trainees and assessors about the benefits of completing a further WBA or even the lack of progression may help to normalise the process. Especially where the focus of such discussion rests on the benefits for the trainee and their patient. Trainees need to be able to trust the assessment process and outcome, including the assessor's opinion. The use of group decisions for outcomes as discussed may help further, especially when such decisions are formed by a group of dedicated Fellows, who strongly identify themselves as assessors and have the logistical and educational expertise supports available to them. This may lead to an increase in the ability to deliver the private thoughts on performance. Similarly, open acknowledgement and discussion that, for a variety of reasons, not all who commence a training program will complete it is important, and that this outcome should not leave such trainees feeling a failure. Some assessors highlighted this and the development of other work options for those who do not complete was discussed. These may be important to develop further.

Trainees and assessors are aware of the subjective nature of assessment messages in the context of the clinical performance assessments and while as discussed, this can contribute a valuable richness and variation to the perspectives on their performance that they receive, awareness of variations in what was delivered was also seen, which is facilitated by the design of the assessments. Several assessors alluded to the malleability of their assessment messages — “will I or won't I give this person this negative mark here” — and many recalled at times, choosing not to deliver a negative message. Trainees were also aware that results could be pre-empted or manipulated. As compared with a formal assessment, where a rubric guides mark allocation and may be collaborated with another assessor, the clinical performance assessments are isolated and assessors have control over what they choose to deliver. Addressing and challenging the concepts around the fear of failure are also relevant, and options such as the use of group discussions or implementation of a programmatic assessment model and avoidance of marks or ratings may lessen reluctance to deliver the assessment message. However, individuals still need to believe that presenting their private thoughts, even if only a comment, will be beneficial rather than devastating in order for them to do so.

Assessment tools

While there have been many modifications proposed and made to assessment tools in the past (80), the outcomes from this study suggest some additional ones may also be important to consider, in the context of the wide range of other implications.

For example, it may be of value to reconsider some of the current practices in regard to completion of WBAs. As discussed, one assessor recognised that the comments given to the trainee may be more valuable if the assessor has had time to consider the assessment event in more detail, rather than provide immediate responses. This recognises the fact that as individuals, assessors may vary in the time they require to process information, and requiring all to give immediate feedback, even if separated from a mark or assessment decision, may potentially limit the quality of message delivered. Further, the current requirement for a single assessor to certify a trainee as competent at a given WBA places significant pressure on them. The opportunity for the assessor to reflect on the trainee's performance and consider their ratings, perhaps in combination with discussion with another assessor may be beneficial to the overall outcome. This is not to suggest that immediate comments could not be given to trainees, nor that the feedback should be excessively delayed, but allowing assessors the opportunity to reflect on their judgement prior to delivering it or reducing their individual responsibility for and the consequences of the message may reduce their reluctance to deliver an appropriately negative message. Of note, feedback is delayed in some settings by virtue of the use of electronic forms, which some assessors complete at a later stage (for example, as occurs in the Australian and New Zealand College of Anaesthetists); and in the ITA setting, some authors have reported that ratings recorded by assessors are lower the longer completion is delayed (59).

Modifications to assessment forms may be further areas for review, though as already noted, as assessor reluctance can be a key barrier to the delivery of assessment messages, such modifications alone would likely be insufficient to eliminate issues. As an example, the use of entrustability scales [discussed in Rekman and colleagues (89)] may be explored as an option to facilitate the delivery of an assessment message, in terms of making the implications of a trainee's current level of competence more explicit. However while these have advantages, they may not necessarily remove reluctance felt by an assessor, when faced with the need to deliver a message that a

senior trainee, approaching the end of their training, is not yet performing at a level where they can be entrusted with independent clinical practice. As we have seen, keeping MUM is about more than simply the perception of unclear criteria.

As an alternative and in line with programmatic assessment models at least one medical specialist training situation has changed the WBA to a purely formative assessment, (100) as was their initial intention, and the experiences would be important to follow. There is a long history of the use of grades in assessment, and modifications to assessments may have unintended consequences. For example, Lefroy and colleagues studied the use of grades in WBAs for medical students (295), where students were able to choose whether to be given grades in the final of a group of WBAs or not. They found that most chose to have grades assigned to their assessments, citing them as valuable and helpful in tracking their progress, though almost one quarter did prefer to avoid them (295). This study highlights the importance of considering the individual learner's preference in regard to assessment messages and how grades are used (or not) by them. The grading system used in the particular medical school (where the grades assigned reflect the end of course standard) may well have influenced the choice for some and differs from other settings, including the current one where trainees are rated according to their level in the training program. In addition, those in training programs may have a different perspective on grades compared with those in medical school; however, the value, use and impact of ratings for medical specialist trainees is a potential area for consideration.

While this study has recognised several potential areas where modifications to the conduct of the assessment program may help to promote the delivery of an assessor's message, it is important to emphasise they need to involve all stakeholders, including trainees and assessors in the design and implementation of changes. Some consequences of not doing so are highlighted by Harrison and colleagues, including feelings of frustration and disempowerment (285) which may jeopardise the effects of such changes. While ongoing resentment to their introduction has been noted by some authors (8), this may be reduced by reorientation to the underlying purpose of the assessments, and the close involvement of trainees and assessors with changes.

Implications for practice for Australian dermatology training

This section highlights how some of the general implications outlined above are applicable to and might be addressed in the Australian dermatology training setting.

Regarding the lack of time for assessments — which was emphasised as an area of concern for many of the participants — provisions already exist in many areas, but are currently not being accessed. For example in Victoria, medical consultants employed in public hospitals are covered by the Victorian Public Health Sector – Medical Specialists Enterprise Agreement 2018–2021, which entitles them to “devote 20% of their normal weekly hours to Clinical Support Duties”. This includes among other items “(d) Contribution to teaching and education of medical staff and other health professionals” and “(k) post-graduate and undergraduate teaching activities” (296). However, a random sample of Victorian dermatology consultants who work in public hospital settings showed that most health services currently do not permit access to this, and this is the case across most speciality training areas too.

Many trainees reported that it can be challenging to arrange the clinical performance assessments, and some assessors recognised that it can be difficult to deliver negative assessment messages when they had only assessed a trainee on a single occasion or where the trainee had not performed many WBAs previously. This could also be addressed at the commencement of a rotation with the supervisor suggesting or allocating assessors with whom to conduct WBA with the trainee. This would facilitate planning in addition to increasing the number of assessments performed with each assessor while preserving multiple assessors for each trainee. Depending on local circumstances, the SoT would be well placed to allocate assessors and suggest a timeframe for completing these in order to avoid both strategic selection and last-minute rush to complete, which places all under further pressure to achieve a satisfactory outcome. However, this should be a collaborative process between the SoT and trainee and not jeopardise the trainee-led focus of these assessments, but rather aim to reduce one of the logistical barriers identified.

As discussed, the WBA information presented by the College highlights that trainees are not expected to be marked as competent on their first attempt of an assessment, but

this does not seem to translate in practice. The logistical challenges outlined may contribute to a motivation for assessors to rate them as satisfactory even on the first assessment. So too may an awareness that most trainees are highly performing individuals who have likely never failed an assessment previously. While removing ratings from these assessments may seem an obvious solution and so avoid the issue of failure and therefore allow the focus to be on improvement, there may be unintended consequences of doing so. As highlighted by Lefroy and colleagues' study discussed earlier, grades may have a motivating influence for some learners (295). Similarly, changing the purpose (formative or summative) of an assessment may also not always have the desired or predicted impact, such as leading to an improvement in narrative comments. For example, Tavares and colleagues showed in a study that a larger number of evaluative comments (those that pass a judgement) as compared with constructive ones (those which are developmental) are recorded regardless of what participants believe the purpose of the assessment to be (297). The reluctance to deliver a message may underlie this.

Other modifications to assessment formats may be important to consider. Some authors have developed methods to review completed assessment forms; for example, with various checks built into them to ensure the message they deliver is accurate and consistent. For example, for the ITA, Dudek and colleagues have designed a validated scale — the Completed Clinical Evaluation Report Rating (CCERR) (298,299) — to measure the quality of the submitted assessment. This relies highly on the content of the comments section (including the word count), and a tool such as this may have application to the current setting where it was noted that more words were written in the strengths area than for development area, and the word count increased with a decreasing total mean rating; hence the word count could be a flag to indicate a lower rating ITA. Further study of the variables in the assessment form would be required to identify the relevant patterns.

A further modification to assessment processes could be to consider a review of the WBA forms (written comments) as a part of the SITA process, as this information could contribute further information to the assessment. As highlighted, in the study setting, comments have no impact upon the outcome of the assessments, but review of them may reveal valuable information to contribute to progress decisions. However, this

would lead to an increase in workload and the change to use comments in this way may also precipitate an increase in MUMing in this part of the form, which would need to be considered. A collective decision on the performance of the trainee by a group of supervisors using a framework and after discussion, similar to the process which occurs in the Annual Review of Competence Progression (ARCP) meetings in the UK (282,283) and in programmatic assessment models, could be proposed to enable such inclusion of all available information on assessment forms and decrease pressure on individual assessors as highlighted in the section on Support from the institution, from colleagues and assessor development, above.

An online community to support assessors as outlined in the section above may be of benefit in several ways and such communities already exist and are well used for the clinical support of Australian dermatologists. While assessor resources are currently readily available via the ACD website, the persisting uncertainties and misunderstandings seen through this study may suggest that assessors do not access the information or do not retain the information as for many, it would be infrequently used. Development of a strong assessor community who meet regularly in person where possible; for example, at the Annual Scientific Meeting and/or through regular webinars, and who are supported in between times with an online community, would promote development of an assessor identity. The ability for assessors to ask questions through such a community, with learning occurring for all as currently occurs with the clinically-based communities currently in existence would be valuable. However, as with the clinical communities in existence, attention to privacy issues would be of paramount importance and such discussions should be limited to those of a general and educational nature only. An enhancement to the assessor role could also come with encouragement and prompting to complete the available online assessor training and attend updates, mediated through such a group.

Recognition of expressing gratitude to assessors was mentioned earlier as important. In the study context, several options for this may be considered — such as the presentation of a certificate of appreciation at the end of each academic year, the institution of an organisation-wide award to recognise commitment to teaching, and/or a regular dedicated social function for supervisors and the training organisation — and these would serve to potentially foster both assessor role identification and wellbeing.

Future research agenda – areas for further study

The aim of this study was to increase the knowledge and understandings about how and why reluctance to deliver negative assessment messages manifests. The broader view of the problem that has been developed through it can inform the design of future projects to measure the impact of the MUM effect in wider contexts throughout medical specialist training, health professional training in general, and even in other education settings.

As such, future projects could explore additional features about the MUM effect and how it manifests. This research suggests that it is important to consider how the effect operates currently and how this varies from the current study. As the size of the study setting appears relevant to outcomes, similar studies in other settings including collaboration with other colleges and training organisations for further understandings of issues related to clinical performance assessments would also be valuable. Similarly, further study into the impact of demographics on MUM behaviours would be of value in a larger group. This research highlights the importance of exploring:

- How levels of MUMing vary between different medical specialist training settings. How factors such as rotation length impact on MUM behaviours — for example, whether programs where trainees rotate to a new site every year have different reports of MUM behaviours compared with those where trainees undertake all their training in one institution.
- Whether and how perceptions of MUMing vary across medical specialist training and other health professional education groups.

The development of objective methods to measure the impact of the MUM effect would be of value in assessing the outcome of interventions. Evaluation of any changes implemented through both qualitative and quantitative methodologies is regarded as important in order to gain a deeper understanding of outcomes than could be possible with a single one alone. Studies on the impact of MUM modifiers, such as role identification and whether enhancing assessor self-identity impacts on assessment behaviours, including their views of the assessments and the concept of failure in the clinical performance setting would be of value. Similarly, following the impact of interventions, such as the use of group decisions and simulation training on MUMing

both in dermatology training and beyond would be important, and collaboration with other organisations could lead to a greater expanse of study. As some limited amounts of MUM behaviours were seen to be beneficial to communication, further research into how much MUM is possible would also be important to undertake. Further relevant research areas include:

- How training and other interventions on MUM behaviours can alter assessor self-report and trainee perspectives. For example, whether and how simulation training or peer support interventions influence the perception of MUM behaviours. How interventions may influence assessor experiences of discomfort in clinical performance assessments.
- How the assessment culture may be modified to improve the delivery of corrective or negative assessment messages.
- How we can work with the MUM effect. Given that it cannot be eliminated entirely, exploration of how much MUMing can be present before the message is changed too much is of value. For example, the impact of delay in delivery on a message and whether this can help or hinder communication.
- How is it best to promote enablers, such as sense of duty and provision of supports, to encourage the delivery of negative assessment messages.

Future research should also expand on areas identified in this project, in particular, the issue of assessor discomfort, including the development of methods to monitor and ideally lessen its impact. Collaboration with other colleges and training organisations would be valuable to both understand discomfort in greater detail, including focusing on the particular areas of the assessor role which lead to discomfort and whether these differ across areas. The impact of direct approaches to assessors by trainees, including their motivations, would also be of value to research further. Finally, further exploration of assessor and trainee perspectives around the core theme — attitude and responsibility for failing a trainee — is required:

- Further qualitative study of the impact of the assessor role on assessors, including the impact on their workload, family, wellbeing.
- Further study into assessor and trainee perceptions of the purpose and potential benefits of clinical performance assessment.

- Further exploration of the meaning and implications of “failure” in clinical performance assessments — both from the assessor and the trainee perspectives.

This research also shows other important areas for study in relation to the issues of reluctance to deliver negative assessment messages. This should include research into how this affects others involved, aside from the assessors, and should consider:

- How trainees make sense of their clinical performance assessment results given they know that assessors keep MUM at times.
- How stakeholders (including patients, other healthcare professionals, broader society) react to the presence of MUMing in clinical performance assessments.

Ensuring that future research included a broad team, expanded beyond that of the candidate alone to include others with additional experience and “outside status” would be valuable.

Limitations

There are many limitations to this study and individual chapters focus specifically on limitations of each study. This section presents a summary of limitations of the study more broadly and conceptually.

While the intention of this work was to show potential application to other settings, it is important to recognise the impact of the setting on the findings. The results were obtained from one setting only, and while this context included the benefit of a national perspective, the relatively small size of the ACD compared with other medical specialist training colleges in Australia, has the potential to limit the applicability of the results. As one assessor highlighted, the outcome of an assessment result may have long lasting effects (“dermatologists hate dermatologists”) and while this is only one opinion, it suggests a culture which may or may not apply to other settings. Certainly, the small size of the College, compared with for example, the much larger colleges of surgeons or physicians, means that dermatologists are more likely to have ongoing interaction with their colleagues throughout their working life; for example, in the setting of continuing education sessions, both state-based and the Annual Scientific Meeting, where many of the country’s dermatologists come to meet annually, and on online clinical discussion communities where dermatologists from around the country interact to discuss clinical and other issues.

The low response rates to both questionnaires (particularly the trainee one) further limit the application of results both within the study setting and more generally. While attempts were made to increase the number responses through repeat advertising in line with Ethics Committee approval, they represent only a sample from the total pool.

Considering the potentially sensitive nature of the study and the data sought, the social desirability bias presents a further potential limitation. This bias predicts that people will be more likely to provide answers or responses to questions which they expect are in line with social norms. Hence the instances and examples of where behaviours deviated from what would be expected or desired presented in this work, may in fact be an under-representation both of true numbers of cases and in the details of the particular events.

A mixed methods study may present a conflict between research paradigms of the qualitative and quantitative components. The use of the pragmatic paradigm has attempted to reduce this conflict by focusing on the best methods for answering the research question and the aim of offering practical solutions to the problems identified (220).

Reflexivity

The development of this project came out of my own observations of a (particular) issue with the clinical performance assessments. While this was an important starting point, it was necessary that this did not restrict the interpretation of the data collected or unduly influence my own perspectives and the findings. Discussions with my supervisors, including the challenges to all aspects of this work and the interpretation of meaning in data was a necessary part of avoiding this, and occurred through regular meetings throughout the entire candidature. Awareness of my personal reactions to the data and the challenging of my own opinions also occurred at a personal level through journaling and reflection.

The choice of data collection tools — being the questionnaires, review of assessment practices and semi-structured interviews — have significantly influenced the data obtained. A mixture of methods was desired in order to obtain both a quantitative documentation of the issues and from the recognition of the limited ability of quantitative data to explore the reasons for the research problem in the depth desired. In

addition, quantitative methods are often favoured by clinicians, whom this work is designed to engage and be accessible and relevant for. Positivism may be the dominant and so the most familiar paradigm in medicine and some clinicians overtly reject ontologies other than realism, hence is an important component of the project. However, the ability to explore beyond this to create an increased understanding of the problem of delivery of negative assessment messages was a further important part of the work.

With the benefit of hindsight, there are several parts of this project that I would have done differently in retrospect.

Firstly, in regard to the Ethics application for the studies, I would have prepared more fully for the situation of low response rates and built in more options to promote participation in the first application. For example, having the scope to increase the number of advertisements or the advertising period depending on participation without needing to submit an amendment would have saved some time in the process. Similarly, having scope to use other sampling methods such as purposive sampling, may have made the interview recruitment process simpler and also ensured that participant demographics could be more diverse.

Considering the questionnaires, further training in IBM SPSS Statistics and access to graphics programs for the presentation of data would have reduced the amount of statistical support required. Using a research assistant to enter data for the review of the SITA would have improved efficiency. An audit of work could have been conducted to ensure accuracy of this. It would have been beneficial to use a program such as Excel when favourableness scores were assigned to the comments. This would have avoided the need for manual entry of the data and also permitted direct calculation of means, further increasing efficiency and reducing the chance of error.

Considering the collection of qualitative data, the use of focus groups in addition to semi-structured interviews may have potentially increased the richness of the data collected while also providing me with additional skills during this research apprenticeship. Formal training in the qualitative data software used (NVivo 12) would also no doubt have increased the efficiency of this.

Having considered areas for improvement, it is also important to acknowledge the aspects that did work well in the project. Practical things such as the referencing system (Mendeley) and systematic review software (Covidence) worked extremely well and made the relevant tasks much more manageable. The courses and sessions provided by the University (including Systematic Review sessions, library sessions on EndNote, Academic Skills courses — such as writing skills, avoiding plagiarism) were highly beneficial as was access to the Department staff and others such as the Statistical Consulting Centre. Finally, regular meetings with my supervisors and peers and journaling were also aspects that were crucial to the completion of this thesis.

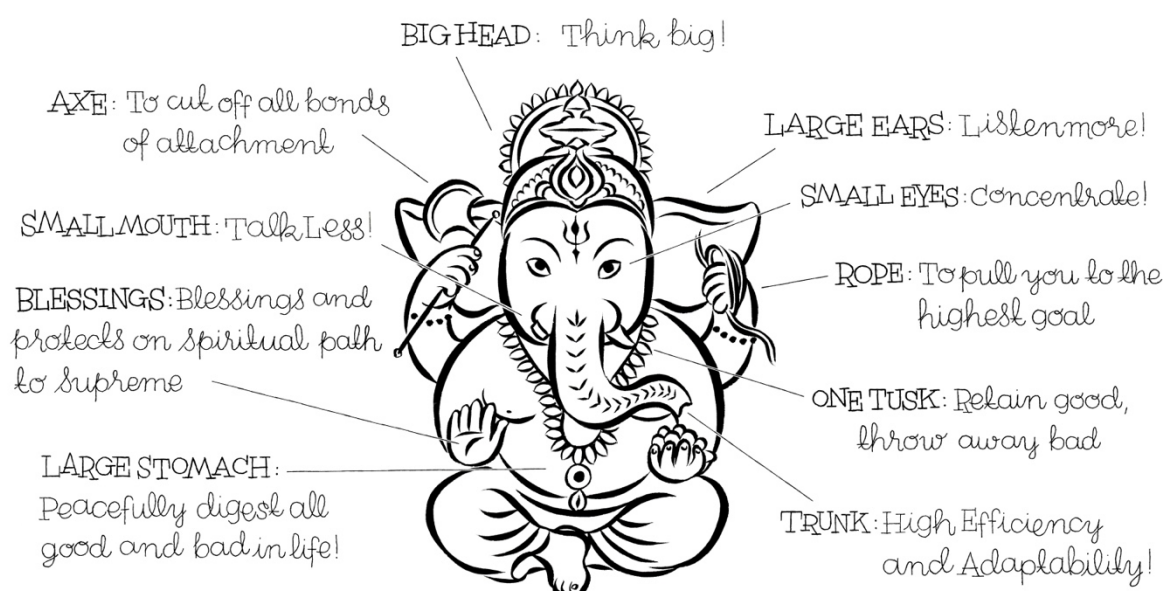


Illustration 10.1: Ganesha – the remover of obstacles (illustration by Jo Martin)

This work has shown that in the study setting, the aims of the clinical performance assessments are not currently universally achieved and much evidence, both anecdotal and from the literature suggests that this applies to other similar settings also. As Ganesha, remover of obstacles and one of the most distinctive Hindu gods represents, there are many aspects to the solution to a problem such as that with these assessment formats. There are currently many barriers preventing success at present and this illustration represents the need to implement multiple and imaginative strategies in order for a solution to be reached.

Summary

This chapter has brought together the findings of the literature reviews and the studies in order to answer the question of how and why does reluctance to deliver negative assessment messages manifest in dermatology training in Australia. Through the framework of the MUM effect, the problem of reluctance to deliver negative assessment messages is seen to extend beyond that of failure to fail and grade inflation, and so is viewed as a larger and arguably more complex one. The MUM effect permits an appreciation and understanding of other behaviours that reflect reluctance, the DAD and other behaviours, and so can enable the potential identification of issues with trainee performance, at an earlier stage — before grade inflation or failure to fail have occurred. The importance of attitudes towards, and the felt responsibility for, failing are seen as core reasons for reluctance to deliver negative messages and the chapter describes these as well as outer level themes in detail. In particular, the impact of discomfort on the assessment process is highlighted — both on the outcomes of the assessments and the wellbeing of assessors. The MUM effect is seen to offer guidance for strategies to address these issues so that assessors can deliver their assessment message and promote the development of the trainee to become the best clinician they can be.

CHAPTER 11

Conclusions

This research project began with an observation about the working of the clinical performance assessments used in the medical education context today — that they do not always achieve their intended aims. That at times, there is a gap between the view of the trainee’s performance that the assessor had formed and what they are prepared to deliver publicly. The literature reveals a great deal of work about the clinical performance assessments, showing that they do not always provide trainees with guiding information and at times, issues such as grade inflation and failure to fail result. While many interventions have been aimed at addressing these issues over a long period of time, none have resolved them entirely to date.

There are many reasons why an assessor may not deliver their assessment message to a trainee. Some may not understand the potential value of these assessments and their importance in guiding trainee improvement. This may be particularly relevant for assessors who did not experience the clinical performance assessments (particularly the WBA) themselves as trainees. Some assessors may not deliver their message if they disagree with the outcome or consequences that may arise from it. Challenging logistics and a lack of practical support with these assessments may make them too onerous to complete as designed, promoting the path of least resistance and their completion with a “tick-box” mentality. However, some assessors don’t deliver their assessment message as they are reluctant to do so. Especially when the message they have to deliver is negative and the problems that this may lead to are not insignificant. The MUM effect, which is observable across time and cultures, offers an explanation for this reluctance and shows how assessors may behave in this situation. Further, it shows that issues with the delivery of negative assessment messages are not confined to the numerical ratings component of an assessment, and so failure to fail and grade inflation are not the only ways in which such reluctance manifests. Knowledge of the MUM effect permits recognition of other more subtle features such as delay, avoidance or distortion of the message (DAD behaviours) which can highlight where reluctance exists.

This study has explored how and why reluctance to deliver a negative assessment message manifests in a medical specialist training context. It has shown that the MUM effect operates and manifests in the DAD behaviours and others such as delegation. Further, it has described other methods assessors may use to solve dilemmas with the

delivery of negative messages. These include delivering the message without consequences, through a written or more commonly verbal comment or via a message in a mark. As with other manifestations of the MUM effect though, these have the potential to impact on the delivery of the message by impairing communication.

While the MUM effect broadens our perspective on the issue of reluctance in the delivery of negative assessment messages, and in effect makes the issue seem larger and more complex, it also helps to increase our knowledge of the problem. Understanding why reluctance occurs and the factors which modify its expression can therefore help to address the issue.

This study has identified the importance of a central theme around assessment messages which is key to address and as a priority above others. Fundamentally negative views of failure as a devastating outcome were seen in this setting. Failure was viewed to be overwhelmingly negative, and its potential to lead to growth and development as a clinician were largely not identified. Further, the responsibility that the assessor holds for such a seemingly disastrous outcome was a further area identified. At present, the recognition that the assessor ultimately decides whether or not to deliver the assessment message may act to prevent delivery of something felt to be potentially harmful to both the trainee and the assessor. These concepts need to be challenged and addressed in order that the clinical performance assessments achieve their potential. Reframing the message — from a negative one to one considered to be helpful and promoting excellence in performance and patient care — would assist its delivery. While reframing this concept may be difficult in a group of high-achieving individuals, many of whom may never have failed anything before, this concept of failure is seen as a barrier necessary to remove in order to permit the delivery of the assessment message.

This study has increased knowledge about other factors that impact on delivery of a negative assessment message which are seen to entwine with the above theme — that being that failure is a very negative outcome and the assessor controls whether such an outcome is made public. A significant association between assessor discomfort with assessment tasks and self-report of MUM behaviours was seen in this work. This is crucial to explore further, both for its impact on assessment outcomes and its impact on assessor wellbeing. Relationships between the assessor and trainee, and direct approaches of assessors by trainees in relation to assessment results were also seen to

impact upon the delivery of a message. Assessors expressed great empathy for trainees and their circumstances and were reluctant to make the situation worse for them, again with the underlying view that failure in this context is a negative outcome. While trainees self-reported relatively low levels of sensitivity to assessment information, the MUM effect predicts that recipients will not wish to hear bad news, which presents a further reason for lack of delivery of a message. The challenging logistics and practical aspects around assessments, which are well reported, were also highlighted as potential barriers to delivery of a message as the lack of practical support provided for them (for example, dedicated time for assessments to occur) communicates the message that the assessments must not be of real importance. In contrast to some other settings, the ACD and health services did not present as reasons for reluctance to deliver a negative message in this context, but the underlying culture and norms around negative information are seen to be relevant. The importance of supporting assessors is recognised and the use of a group for decision making is seen to be an important option.

The outcomes of this study can be used to guide and inform methods to reduce reluctance to delivering a negative assessment message. As outlined, the most important being to consider the concept that currently, negative messages are largely considered devastating or embarrassing, rather than helpful. Promoting the assessor role and understandings of the benefits of delivering a negative message are further important areas to consider. Changing a longstanding culture however is not a simple task.

Monitoring of assessor discomfort and further research leading to nuanced training are further options to decrease its impact and negative effects on both assessment messages and assessor welfare. Fostering assessor supports, including those of colleagues and the training institution in particular, are additional areas of importance.

Delivering a message that a trainee's level of performance requires improvement or that they have not achieved the standard required in a clinical performance assessment may never be an easy task for many assessors. However, the MUM effect can help to anticipate when problems may arise and identify when a gap between private thoughts and public judgements is occurring. As a pervasive phenomenon, it may not be possible to eliminate entirely, and indeed some amount of MUMing can enhance communication of a message. Further research into the right amount of MUM and the use of modifiers

to help promote delivery may help assessors deliver their message, and so ultimately help trainees to develop into the best clinician they can be. For their patients are depending on it.

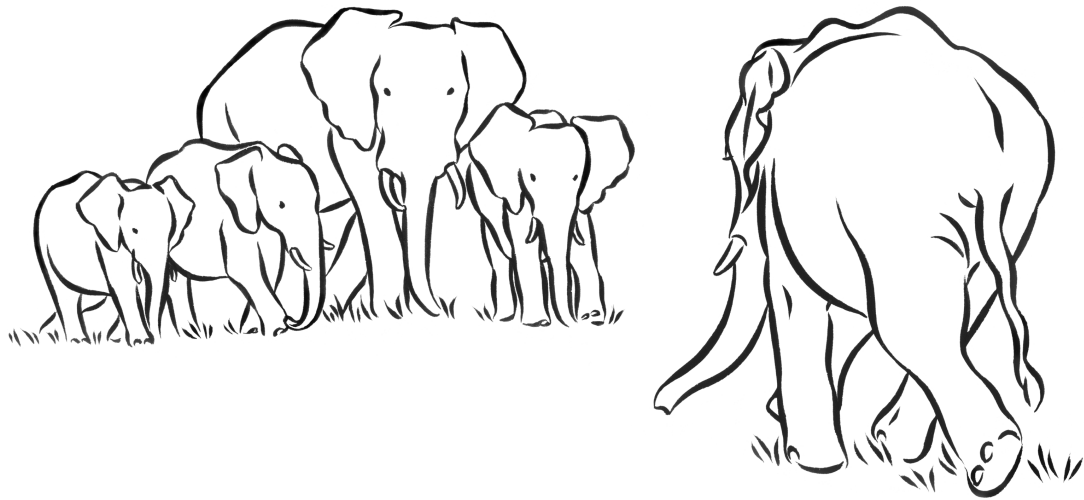


Illustration 11.1: Returning (illustration by Jo Martin)

REFERENCES

1. Epstein RM, Hundert EM. Defining and Assessing Professional Competence. *JAMA*. 2002;287(2):226–35.
2. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Ann Intern Med*. 1995;123(10):795–9.
3. Ende J. Feedback in clinical medical education. *JAMA*. 1983;250(6):777–81.
4. Miller GE. The Assessment of Clinical Skills/Competence/Performance. *Acad Med*. 1990;65(9):S63–7.
5. Ringsted C, Henriksen AH, Skaarup AM, Van der Vleuten CPM. Educational impact of in-training assessment (ITA) in postgraduate medical education: A qualitative study of an ITA programme in actual practice. *Med Educ*. 2004 Jul;38(7):767–77.
6. Barrett A, Galvin R, Scherpbier AJJA, Teunissen PW, O’Shaughnessy A, Horgan M. Is the learning value of workplace-based assessment being realised? A qualitative study of trainer and trainee perceptions and experiences. *Postgrad Med J*. 2016;postgradmedj-2015-133917.
7. Alofs L, Huiskes J, Heineman MJ, Buis C, Horsman M, van der Plank L, et al. User reception of a simple online multisource feedback tool for residents. *Perspect Med Educ*. 2015;4(2):57–65.
8. Shalhoub J, Marshall DC, Ippolito K. Perspectives on procedure-based assessments: A thematic analysis of semistructured interviews with 10 UK surgical trainees. *BMJ Open*. 2017;7(3):1–8.
9. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med*. 2003;15(4):270–92.
10. Cleland JA, Knight L V, Rees CE, Tracey S, Bond CM. Is it me or is it them?

- Factors that influence the passing of underperforming students. *Med Educ*. 2008 Aug;42(8):800–9.
11. Dudek NL, Marks MB, Regehr G. Failure to fail: The perspectives of clinical supervisors. *Acad Med*. 2005 Oct;80(10):S84–7.
 12. Scarff CE, Corderoy RM, Bearman M. In-training assessments: ‘The difficulty is trying to balance reality and really tell the truth’. *Australas J Dermatol*. 2018 Feb;59(1):e15–22.
 13. Larocque S, Luhanga FL. Exploring the issue of failure to fail in a nursing program. *Int J Nurs Educ Scholarsh*. 2013;10(1):115–22.
 14. Pratt M. Evaluation of unsatisfactory student performance. *Clin Teach*. 2019;16:1–6.
 15. Scarff CE. Supervisor perspectives of influences upon their assessments of Australian Dermatology Trainees. Monash University; 2014.
 16. Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An Exploration of Faculty Perspectives on the In-Training Evaluation of Residents. *Acad Med*. 2010;85(7):1157–62.
 17. Wass V, Van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;357:945–9.
 18. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Heal Sci Educ*. 1996;1:41–67.
 19. Tabish SA. Assessment methods in medical education. *Int J Heal Sci*. 2008;2(2):3–7.
 20. Ferris H, O’Flynn D. Assessment in Medical Education; What Are We Trying to Achieve? *Int J High Educ*. 2015 Jan 1;4(2):139–44.
 21. Speer AJ, Solomon DJ, Fincher RME. Grade Inflation in Internal Medicine Clerkships: Results of a National Survey. *Teach Learn Med*. 2000;12(3):112–6.

22. Bingham CM, Crampton R. A review of prevocational medical trainee assessment in New South Wales. *Med J Aust*. 2011 Oct 3;195(7):410–2.
23. Bodgener S, Tavabie A. Is there a value to case-based discussion? *Educ Prim Care*. 2011;22(4):223–8.
24. Cavalcanti RB, Detsky AS. The education and training of future physicians: Why coaches can't be judges. *Jama*. 2011;306(9):993–4.
25. Long G. Documentation of in-training assessment for radiology trainees. *Clin Radiol*. 2001;56(4):310–5.
26. Aram N, Brazil V, Davin L, Greenslade J. Intern underperformance is detected more frequently in emergency medicine rotations. *EMA - Emerg Med Australas*. 2013;25(1):68–74.
27. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874–81.
28. Cohen GS, Blumberg P, Ryan NC, Sullivan PL. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med*. 1993;5(1):10–5.
29. Boud D, Falchikov N. Aligning assessment with long-term learning. *Assess Eval High Educ*. 2006;31(4):399–413.
30. Eva KW, Regehr G. 'I'll Never Play Professional Football' and Other Fallacies of Self-Assessment. *J Contin Educ Health Prof*. 2008 Dec 1;28(1):14–9.
31. Cleland JA, Arnold R, Chesser A. Failing finals is often a surprise for the student but not the teacher: identifying difficulties and supporting students with academic difficulties. *Med Teach*. 2005 Sep;27(6):504–8.
32. Langendyk V. Not knowing that they do not know: self-assessment accuracy of third-year medical students. *Med Educ*. 2006 Feb;40(2):173–9.
33. Eva KW, Regehr G. Self-Assessment in the Health Professions: A Reformulation

- and Research Agenda. *Acad Med*. 2005;80(Supplement):S46–54.
34. Bing-you RG, Trowbridge RL. Why medical educators may be failing at feedback. *JAMA*. 2009;302:1330–1.
 35. Probyn L, Lang C, Tomlinson G, Bandiera G. Multisource Feedback and Self-Assessment of the Communicator, Collaborator, and Professional CanMEDS Roles for Diagnostic Radiology Residents. *Can Assoc Radiol J*. 2014;65(4):379–84.
 36. Jones R, Panda M, Desbiens N. Internal medicine residents do not accurately assess their medical knowledge. *Adv Heal Sci Educ*. 2008;13(4):463–8.
 37. Kruger J, Dunning D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol*. 1999;77(6):1121–34.
 38. Ginsburg S, van der Vleuten CPM, Eva KW, Lingard L. Cracking the code: Residents' interpretations of written assessment comments. *Med Educ*. 2017;51(4):401–10.
 39. Boehler ML, Rogers DA, Schwind CJ, Mayforth R, Quin J, Williams RG, et al. An investigation of medical student reactions to feedback: a randomised controlled trial. *Med Educ*. 2006;40(8):746–9.
 40. Norcini JJ, McKinley DW. Assessment methods in medical education. *Teach Teach Educ*. 2007 Apr 1;23(3):239–50.
 41. Govaerts MJB, van de Wiel MWJ, van der Vleuten CPM. Quality of Feedback Following Performance Assessments: Does Assessor Expertise Matter? *Eur J Train Dev*. 2013 Jan 1;37(1):105–25.
 42. Humphrey-Murto S, Khalidi N, Smith CD, Kaminska E, Touchie C, Keely E, et al. Resident evaluations: The use of daily evaluation forms in rheumatology ambulatory care. *J Rheumatol*. 2009;36(6):1298–303.
 43. Govaerts MJB, Van Der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. The Use of Observational Diaries in In-Training Evaluation: Student Perceptions.

Adv Heal Sci Educ. 2005 Aug 1;10(3):171–88.

44. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve tips for programmatic assessment. *Med Teach*. 2015;37(7):641–6.
45. Weller JM, Misur M, Nicolson S, Morris J, Ure S, Crossley J, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth*. 2014;112(6):1083–91.
46. Holmboe ES, Yepes M, Williams F, Huot SJ. Feedback and the mini clinical evaluation exercise. *J Gen Intern Med*. 2004;19(5 Pt 2):558–61.
47. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. *Med Teach*. 2013;35(6):e1197-210.
48. Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. *Br J Anaesth*. 2017;118(2):207–14.
49. Brown JM, Lowe K, Fillingham J, Murphy PN, Bamforth M, Shaw NJ (Ben). An investigation into the use of multi-source feedback (MSF) as a work-based assessment tool. *Med Teach*. 2014;36(11):997–1004.
50. Johnson G, Booth J, Crossley J, Wade W. Assessing trainees in the workplace: Results of a pilot study. *Clin Med J R Coll Physicians London*. 2011;11(1):48–53.
51. Wilkinson TJ, Wade WB. Problems with using a supervisor's report as a form of summative assessment. *Postgrad Med J*. 2007 Jul 1;83(981):504–6.
52. Liao K-C, Pu S-J, Liu M-S, Yang C-W, Kuo H-P. Development and implementation of a mini-Clinical Evaluation Exercise (mini-CEX) program to assess the clinical competencies of internal medicine residents: from faculty development to curriculum evaluation. *BMC Med Educ*. 2013;13(1):31.
53. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teach*. 2007;29(9):855–71.

54. Findlay N. General practioner registrars' experiences of multisource feedback: a qualitative study. *Educ Prim Care*. 2012;23(5):323–9.
55. Malling B, Bested KM, Skjelsager K, Ostergaard HT, Ringsted C. Long-term effect of a course on in-training assessment in postgraduate specialist education. *Med Teach*. 2007;29(9):966–71.
56. Gingerich A, Regehr G, Eva KW. Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. *Acad Med*. 2011;86(10):S1–7.
57. Newble DI, Haore J, Sheldrake PF. The selection and training of examiners for clinical examinations. *Med Educ*. 1980;14(5):345–9.
58. Jason H. Future medical education: Preparing, priorities, possibilities. *Med Teach*. 2018;40(10):996–1003.
59. Paget M, Wu C, McIlwrick J, Woloschuk W, Wright B, McLaughlin K. Rater variables associated with ITER ratings. *Adv Heal Sci Educ*. 2013 Oct 1;18(4):551–7.
60. Yeates P, O'Neill P, Mann K, Eva KW. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA*. 2012;308(21):2226–32.
61. Wilkinson TJ, Tweed MJ, Egan TG, Ali AN, McKenzie JM, Moore M, et al. Joining the dots: Conditional pass and programmatic assessment enhances recognition of problems with professionalism and factors hampering student progress. *BMC Med Educ*. 2011;11(1):29.
62. Quiller-Couch AT. *On the art of reading*. Pocket add. Cambridge at the University Press; 1939.
63. Rojstaczer S, Healy C. Where A Is Ordinary : The Evolution of American College and University Grading , Sign-in Purchase this Article. *Teach Coll Rec*. 2012;114(July 2012):1–2.
64. White HB. Commentary: Problem-based learning and grade inflation. *Biochem*

- Mol Biol Educ. 2005;33(2):133–4.
65. Twenge JM, Campbell WK, Gentile B. Generational Increases in Agentic Self-evaluations among American College Students, 1966–2009. *Self Identity*. 2012;11(4):409–27.
 66. Pear J. Grade inflation. *Science* (80-). 1976;191:340.
 67. Fazio SB, Papp KK, Torre DM, Defer TM. Grade inflation in the internal medicine clerkship: a national survey. *Teach Learn Med*. 2013;25(1):71–6.
 68. Cacamese SM, Elnicki M, Speer AJ. Grade inflation and the internal medicine subinternship: a national survey of clerkship directors. *Teach Learn Med*. 2007;19(4):343–6.
 69. Daelmans HEM, Overmeer RM, Van Der Hem-Stokroos HH, Scherpbier AJJA, Stehouwer CDA, Van Der Vleuten CPM. In-training assessment: Qualitative study of effects on supervision and feedback in an undergraduate clinical rotation. *Med Educ*. 2006;40(1):51–8.
 70. Colletti LM. Difficulty with Negative Feedback : Face-to-Face Evaluation of Junior Medical Student Clinical Performance Results in Grade Inflation. *J Surg Res*. 2000;90(1):82–7.
 71. Tunbridge M, Dickinson D, Swan P. Outcomes of assessments of registrars in the medical specialties. *Clin Med (Northfield Il)*. 2004;4(1):66–8.
 72. Patel M, Agius S, Wilkinson J, Patel L, Baker P. Value of supervised learning events in predicting doctors in difficulty. *Med Educ*. 2016;50(7):746–56.
 73. Kendal WS, MacRae R, Dagg P. Problems with subjective in-training evaluations [4]. *South Med J*. 2004;97(10):1024.
 74. Hesketh EA, Laidlaw JM. Developing the teaching instinct, 1: feedback. *Med Teach*. 2002 May;24(3):245–8.
 75. Bush HM, Schreiber RS, Oliver SJ. Failing to fail: Clinicians’ experience of assessing underperforming dental students. *Eur J Dent Educ*. 2013;17(4):198–

76. Roberts NK, Williams RG. The Hidden Costs of Failing to Fail Residents. *J Grad Med Educ*. 2011;3(2):127–9.
77. Challis M, Flett A, Batstone G. An accident waiting to happen? A case for medical education. *Med Teach*. 1999;21(6):582–5.
78. Paice E. Identification and management of the underperforming surgical trainee. *ANZ J Surg*. 2009;79(3):180–4.
79. Cleland J, Leggett H, Sandars J, Costa MJ, Patel R, Moffat M. The remediation challenge: Theoretical and methodological insights from a systematic review. *Med Educ*. 2013;47(3):242–51.
80. Yepes-Rios M, Dudek N, Duboyce R, Curtis J, Allard RJ, Varpio L. The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. *Med Teach*. 2016;38(11):1092–9.
81. Kogan JR, Conforti LN, Bernabeo E, Iobst W, Holmboe E. How faculty members experience workplace-based assessment rater training: a qualitative study. *Med Educ*. 2015;49(7):692–708.
82. Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. *Acad Med*. 1999;74:1203–7.
83. Holmboe ES, Denham S, Peznick R, Katsufakis P, Leslie K, Patel V, et al. Faculty development in assessment: the missing link in competency-based medical education [Internet]. *Academic Medicine*. 2011 [cited 2015 Jul 9]. p. 460–7. Available from: <https://eds-b-ebshost-com.ezp.lib.unimelb.edu.au/eds/detail/detail?vid=6&sid=9fcbce59-33a8-484d-b5a8-0f729c41079c%40sessionmgr111&hid=113&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3D%3D#db=cmedm&AN=21346509>
84. Zedeck S, Baker HHT. Nursing performance as measured by behavioural expectation scales: a multitrait multirater analysis. *Organ Behav Hum Perform*.

1972;7(3):457–66.

85. Singh T, Modi JN. Workplace based assessment: a step to promote competency based postgraduate training. *Indian Pediatr.* 2013;50(6):553–9.
86. van der Vleuten C, Verhoeven B. In-training assessment developments in postgraduate education in Europe. *ANZ J Surg.* 2013;83(6):454–9.
87. Govaerts M, van der Vleuten CP. Validity in work-based assessment: Expanding our horizons. *Med Educ.* 2013;47:1164–74.
88. Weller JM, Jones A, Merry AF, Jolly B, Saunders D, J.M. W, et al. Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: Implications for implementation. *Br J Anaesth.* 2009;103(4):524–30.
89. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment. *Acad Med.* 2016;91(2):186–90.
90. Mellsop G, Berah E. The relationship between in-service assessment and post-graduate examination performance in psychiatry. *Aust N Z J Psychiatry.* 1982 Jun;16(2):84–6.
91. Walling A, Merando A. The fourth year of medical education: a literature review. *Acad Med.* 2010;85(11):1698–704.
92. Guerrasio J, Furfari KA, Rosenthal LD, Nogar CL, Wray KW, Aagaard EM. Failure to fail: The institutional perspective. *Med Teach.* 2014;1–5.
93. Ryan J, Mandel F, Sama A, Ward M. Reliability of Faculty Clinical Evaluations of Non-Emergency Medicine Residents during Emergency Department Rotations. *Acad Emerg Med.* 1996;3(12):1124–30.
94. Bok HGJ, Jaarsma DADC, Spruijt A, Van Beukelen P, Van Der Vleuten CPM, Teunissen PW. Feedback-giving behaviour in performance evaluations during clinical clerkships. *Med Teach.* 2016;38(1):88–95.
95. Whitla DK, Tirrell JE. The Validity of Ratings of Several Levels of Supervisors.

Organ Behav Hum Perform. 1970;5:366–77.

96. Noel GL. A system for evaluating and counseling marginal students during clinical clerkships. *J Med Educ*. 1987;62(4):353–5.
97. Bing-You R, Varaklis K, Hayes V, Trowbridge R, Kemp H, McKelvy D. The feedback tango: An integrative review and analysis of the content of the teacher-learner feedback exchange. *Acad Med*. 2018;93(4):657–63.
98. Klein RH, Babineau R. Evaluating the competence of trainees: It's nothing personal. *Am J Psychiatry*. 1974;131(7):788–91.
99. Barone JE. Problems with the fourth-year curriculum of students entering surgical residencies. *Am J Surg*. 1995;169:334–7.
100. Devlin AP, Fazackerley J, Penfold N. Improving workplace-based assessments. *Br J Anaesth*. 2016;116(5):718–9.
101. Fokkema JPI, Teunissen PW, Westerman M, van der Lee N, van der Vleuten CPM, Scherpbier AJJ a, et al. Exploration of perceived effects of innovations in postgraduate medical education. *Med Educ*. 2013;47(3):271–81.
102. Watling CJ, Kenyon CF, Zibrowski EM, Schulz V, Goldszmidt MA, Singh I, et al. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med*. 2008 Oct;83(10 Suppl):S97-100.
103. Gingerich A, Kogan JR, Yeates P, Govaerts MJB, Holmboe ES. Seeing the 'Black Box' differently: Assessor cognition from three research perspectives. *Med Educ*. 2014;1055–68.
104. Sim F. Record of in-training assessment (RITA): a look at ethical issues in assessment. *Hosp Med*. 1999;60(9):676–8.
105. Qadan L, Al-Ozairi E, Ayed A, Huang G. Avoiding honest feedback: Discordance between formal evaluations and candid assessments of Kuwaiti PBL students. *Med Teach*. 2013;35(6):459–64.
106. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box

- of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011 Oct;45(10):1048–60.
107. Tsui AS, Barry B. Interpersonal affect and rating errors. *Acad Manag J*. 1986;29(3):586–99.
 108. Borges NJ, Manuel S, Elam CL, Jones BJ. Comparing millennial and generation X medical students at one medical school. *Acad Med*. 2006;81(6):571–6.
 109. Gaunt A, Patel A, Rusius V, Royle TJ, Markham DH, Pawlikowska T. ‘Playing the game’: How do surgical trainees seek feedback using workplace-based assessment? *Med Educ*. 2017;51(9):953–62.
 110. Cambridge Dictionary [Internet]. [cited 2019 Dec 7]. Available from: <https://dictionary.cambridge.org/dictionary/english/an-elephant-in-the-room>
 111. Scarff CE, Bearman M, Chiavaroli N, Trumble S. Keeping mum in clinical supervision: private thoughts and public judgements. *Med Educ*. 2019 Feb;53(2):133–42.
 112. Mohrman AM, Lawler EE. Motivation and performance appraisal behaviour. In: *Performance Measurement: Directions for the Future*. 1981. p. 33.
 113. Rosen S, Tesser A. On reluctance to communicate undesirable information: The MUM effect. *Sociometry*. 1970;33(3):253–63.
 114. Kivlighan DM, Luiza JW. Examining the Credibility Gap and the Mum Effect: Rex Stockton’s Contributions to Research on Feedback in Counseling Groups. *J Spec Gr Work*. 2005;30(3):253–69.
 115. Tesser A, Rosen S, Batchelor T. Some Message Variables and the MUM Effect. *J Commun*. 1972;22(3):239–56.
 116. Dibble JL, Levine TR. Sharing Good and Bad News with Friends and Strangers: Reasons for and Communication Behaviors Associated with the MUM Effect. *Commun Stud*. 2013;64(4):431–52.
 117. Park C, Im G, Keil M. Overcoming the mum effect in IT project reporting:

- Impacts of fault responsibility and time urgency. *J Assoc Inf Syst.* 2008;9(7):409–31.
118. Ramingwong S, Sitthichok S. A Survey on Mum Effect and Its Influencing Factors. *Procedia Technol.* 2013;9:618–26.
 119. Cox SS, Credo KR. Do I trust you enough to tell you what you don't want to hear ? An Examination of the Moderating Effect of Trust on Job Satisfaction and Mum. *Int J Acad Bus World.* 2014;8(1):11–6.
 120. Merker BM, Hanson WE, Poston JM. National survey of psychologists' training and practice in breaking bad news: a mixed methods study of the MUM effect. *J Clin Psychol Med Settings.* 2010;17(3):211–9.
 121. Yariv E. 'Mum effect': Principals' reluctance to submit negative feedback. *J Manag Psychol.* 2006;21(6):533–46.
 122. Ramingwong S, Ramingwong L. A tale behind mum effect. *Int J Inf Syst Proj Manag.* 2013;1(3):47–58.
 123. Dibble JL. Breaking Good and Bad News: Face-Implicating Concerns as Mediating the Relationship Between News Valence and Hesitation to Share the News. *Commun Stud.* 2014;65(3):223–43.
 124. Dibble JL, Levine TR. Breaking Good and Bad News: Direction of the MUM Effect and Senders' Cognitive Representations of News Valence. *Communic Res.* 2010;37(5):703–22.
 125. Tesser A, Rosen S. The reluctance to transmit bad news. In: Berkowitz L, editor. *Advances in experimental social psychology.* Vol 8. New York: Academic Press; 1975. p. 193–232.
 126. Sawyer TP, Hollis-Sawyer LA, Pokryfke A. Personality and social-evaluative anxieties associated with rating discomfort in anticipated positive and negative feedback conditions. *Soc Behav Personal an Int J.* 2002;30(4):359–371.
 127. Rosen S, Tesser A. Fear of negative evaluation and the reluctance to transmit bad news. *J Commun.* 1972;22:124–41.

128. Tesser A, Rosen S, Waranch E. Communicator Mood and the Reluctance to Transmit Undesirable Messages (The Mum Effect). *J Commun.* 1973;23(3):266–83.
129. Tesser A, Rosen S, Tesser M. On the reluctance to communicate undesirable messages (the MUM effect): A field study. *Psychol Rep.* 1971;29:651–4.
130. Benedict ME, Levine EL. Delay and Distortion : Tacit Influences on Performance Appraisal Effectiveness. *J Appl Psychol.* 1988;73(3):507–14.
131. Fisher CD. Transmission of positive and negative feedback to subordinates: A laboratory investigation. *J Appl Psychol.* 1979;64(5):533–40.
132. Marler LE, McKee L, Cox S, Allen D. Don't Make Me the Bad Guy: Organizational Norms, Self-monitoring, and the Mum Effect. *J Manag Issues.* 2012;XXIV(1):97–116.
133. Bisel RS, Messersmith a. S, Kelley KM. Supervisor-Subordinate Communication: Hierarchical Mum Effect Meets Organizational Learning. *J Bus Commun.* 2012;49:128–47.
134. Larson JR. Supervisors' performance feedback to subordinates: the impact of subordinate performance valence and outcome dependence. *Organ Behav Hum Decis Process.* 1986;37:391–408.
135. Ploeger N a., Kelley KM, Bisel RS. Hierarchical Mum Effect: A New Investigation of Organizational Ethics. *South Commun J.* 2011;76(5):465–81.
136. Fang C, Kim J-H, Milliken FJ. When bad news is sugarcoated: Information distortion, organizational search and the behavioral theory of the firm. *Strateg Manag J.* 2014;35:1186–201.
137. Tesser A, Rosen S. Similarity of objective fate as a determinant of the reluctance to transmit unpleasant information: The MUM effect. *J Pers Soc Psychol.* 1972;23(1):46–53.
138. Weenig MW, Groenenboom a C, Wilke H a. Bad news transmission as a function of the definitiveness of consequences and the relationship between

- communicator and recipient. *J Pers Soc Psychol.* 2001;80(3):449–61.
139. Weenig MWH, Wilke H a. M, Mors ET. Personal Outcomes and Moral Responsibility as Motives for News Transmission: The Impact of Fate Similarity, Fate Uncertainty, and Relationship Closeness. *Communic Res.* 2014;41(3):404–29.
 140. Rutkowski K. Failure to fail: assessing nursing students' competence during practice placements. *Nurs Stand.* 2007;22(13):35–40.
 141. Uysal A, Öner-Özkan B. A Self-Presentational Approach To Transmission of Good and Bad News. *Soc Behav Personal an Int J.* 2007;35(1):63–78.
 142. Bond CF, Anderson EL. The Reluctance to Transmit Bad News: Private Discomfort or Public Display? *J Exp Soc Psychol.* 1987;23(2):176–87.
 143. Dibble JL, Wisner AM, Dobbins L aure., Cacal M, Taniguchi E, Peyton A, et al. Hesitation to Share Bad News: By-Product of Verbal Message Planning or Functional Communication Behavior? *Communic Res.* 2015;42(2):213–36.
 144. Shaw JM, Brown RF, Dunn SM. A qualitative study of stress and coping responses in doctors breaking bad news. *Patient Educ Couns.* 2013;91(2):243–8.
 145. Buckman R. Breaking bad news: why is it still so difficult? *Br Med J (Clin Res Ed).* 1984;288(6430):1597–9.
 146. Arnold SJ, Koczwara B. Breaking bad news: Learning through experience. *J Clin Oncol.* 2006;24(31):5098–100.
 147. Ptacek JT, Eberhardt TL. Breaking bad news: A Review of the Literature. *JAMA.* 1996;276(6):496–502.
 148. Meunier J, Merckaert I, Libert Y, Delvaux N, Etienne AM, Liénard A, et al. The effect of communication skills training on residents' physiological arousal in a breaking bad news simulated task. *Patient Educ Couns.* 2013;93(1):40–7.
 149. Tobler K, Grant E, Marczynski C. Evaluation of the Impact of a Simulation-enhanced Breaking Bad News Workshop in Pediatrics. *Simul Healthc.*

2014;9(4):213–9.

150. Conlee MC, Tesser A. The Effects of Recipient Desire to Hear on News Transmission. *Sociometry*. 1973;36(4):588–99.
151. Sender Liberman A, Liberman M, Steinert Y, McLeod P, Meterissian S. Surgery residents and attending surgeons have different perceptions of feedback. *Med Teach*. 2005;27(5):470–2.
152. Park C, Keil M. Organizational silence and whistle- blowing on IT projects: An integrated model. *Decis Sci*. 2009;40(4):901–918.
153. Keil M, Smith HJ, Pawlowski S, Jin L. ‘Why didn’t somebody tell me’: Climate, information asymmetry and bad news about troubled projects. *SIGMIS Database*. 2004;35(2):65–84.
154. Tesser A, Rosen S, Batchelor TR. On the reluctance to communicate bad news (the MUM effect): A role play extension. *J Pers*. 1972;40(1):88–103.
155. Speer A, Solomon D, Ainsworth M. An innovative evaluation method in an internal medicine clerkship. *Acad Med*. 1996;71:S76–8.
156. Smith WJ, Harrington KV, Houghton JD. Predictors of performance appraisal discomfort: a preliminary examination. *Public Pers Manag*. 2000;29(1):21–31.
157. Cox SS, Marler LE, Simmering MJ, Totten JW. Giving feedback: Development of scales of the Mum effect, discomfort giving feedback and feedback medium preference. *Perform Improv Q*. 2011;23(4):49–69.
158. Scarff CE, Bearman M, Corderoy RM. Supervisor perspectives on the summative in-training assessment. *Australas J Dermatol*. 2016;57(2):128–34.
159. O’Neal EC, Levine DW, Frank JF. Reluctance to transmit bad news when the recipient is unknown: experiments in five nations. *Soc Behav Pers*. 1979;7(1):39–47.
160. Sussman SW, Sproull L. Straight Talk: Delivering Bad News through Electronic Communication. *Inf Syst Res*. 1999;10(2):150–66.

161. Ye C, Fulton J, Tourangeau R. More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opin Q.* 2011;75(2):349–65.
162. Waung M, Highhouse S. Fear of conflict and empathic buffering: Two explanations for the inflation of performance feedback. *Organ Behav Hum Decis Process.* 1997;71(1):37–54.
163. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med.* 1992 Nov 1;117(9):757–65.
164. Witt PL, Kerssen-Griep J. Instructional Feedback I: The Interaction of Facework and Immediacy on Students' Perceptions of Instructor Credibility. *Commun Educ.* 2011;60(1):75–94.
165. Kerssen-Griep J, Witt PL. Instructional Feedback II: How Do Instructor Immediacy Cues and Facework Tactics Interact to Predict Student Motivation and Fairness Perceptions? *Commun Stud.* 2012;63(4):498–517.
166. Moss SE, Martinko MJ. The effects of performance attributions and outcome dependence on leader feedback behavior following poor subordinate performance. *J Organ Behav.* 1998;19(3):259–74.
167. Sajeev ASM, Ramingwong S. Mum effect as an offshore outsourcing risk: A study of differences in perceptions. *Comput J.* 2010;53(1):120–6.
168. Philipson A. Translation table explaining the truth behind British politeness becomes internet hit. *The Telegraph.* 2013;
169. King PE, Schrodtt P, Weisel JJ. The instructional feedback orientation scale: Conceptualizing and validating a new measure for assessing perceptions of instructional feedback. *Commun Educ.* 2009;58(2):235–61.
170. Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP. SPIKES--A Six-Step Protocol for Delivering Bad News: Application to the Patient with Cancer. *Oncologist.* 2000;5(4):302–11.
171. Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM.

- Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Heal Sci Educ*. 2007 May;12(2):239–60.
172. Scarff CE, Bearman M, Chiavaroli N, Trumble S. Trainees' perspectives of assessment messages: a narrative systematic review. *Med Educ*. 2019 Mar;53(3):221–33.
 173. Whitehouse A, Walzman M, Wall D. Pilot study of 360 degree assessment of personal skills to inform record of in training assessments for senior house officers. *Hosp Med*. 2002;63(3):172–5.
 174. Saedon H, Salleh S, Balakrishnan A, Imray CH, Saedon M. The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. *BMC Med Educ*. 2012;12(1):25.
 175. Ajjawi R, Regehr G. When I say ... feedback. *Med Educ*. 2019;53(7):652–4.
 176. McGill DA, van der Vleuten CPM, Clarke MJ. A critical evaluation of the validity and the reliability of global competency constructs for supervisor assessment of junior medical trainees. *Adv Heal Sci Educ*. 2013 Oct 1;18(4):701–25.
 177. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010;341:c5064.
 178. Weller JM, Jolly B, Misur MP, Merry AF, Jones A, Crossley JGM, et al. Mini-clinical evaluation exercise in anaesthesia training. *Br J Anaesth*. 2009;102(5):633–41.
 179. Bing-You R, Hayes V, Varaklis K, Trowbridge R, Kemp H, McKelvy D. Feedback for Learners in Medical Education: What is known? A scoping review. *Acad Med*. 2017;92(9):1.
 180. Massie J, Ali JM. Workplace-Based Assessment: A Review of User Perceptions and Strategies to Address the Identified Shortcomings. *Adv Heal Sci Educ*. 2016 May 1;21(2):455–73.
 181. Collins English Dictionary [Internet]. Collins English Dictionary. [cited 2018

Aug 16]. Available from:

<https://www.collinsdictionary.com/dictionary/english/perspective>

182. Beller EM, Glasziou PP, Altman DG, Hopewell S, Bastian H, Chalmers I, et al. PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts. *PLoS Med*. 2013;10(4).
183. Reed DA, Cook DA, Beckman TTJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. *JAMA*. 2007;298(9):1002–9.
184. Walsh D, Downe S. Appraising the quality of qualitative research. *Midwifery*. 2006;22(2):108–19.
185. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77–101.
186. Bindal T, Wall D, Goodyear HM. Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise? *Med Teach*. 2011;33(January):919–27.
187. Castanelli DJ, Jowsey T, Chen Y, Weller JM. Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training. *Can J Anesth*. 2016;63(12):1345–56.
188. Patel R, Drover A, Chafe R. Pediatric faculty and residents perspectives on in-training evaluation reports (ITERs). *Can Med Educ J*. 2015;6(2):41–53.
189. Sabey A, Harris M. Training in hospitals: What do GP specialist trainees think of workplacebased assessments? *Educ Prim Care*. 2011;22(2):90–9.
190. Bearman M, Dawson P. Qualitative synthesis and systematic review in health professions education. *Med Educ*. 2013;47(3):252–60.
191. Mehta F, Brown J, Shaw NJ. Do trainees value feedback in case-based discussion assessments? *Med Teach*. 2013;35:e1166-72.
192. Goel A, Singh T. The usefulness of Mini Clinical Evaluation Exercise as a

- learning tool in different pediatric clinical settings. *Int J Appl Basic Med Res*. 2015;5:S32-4.
193. Grennan S, Crowley S, Quidwai S, Barrett O, Kooblall M. Is the current BST ePortfolio fulfilling its role in the training of clinical medicine SHO's? *Ir Med J*. 2016;109(1):343–6.
 194. Mohanaruban A, Flanders L, Rees H. Case-based discussion: perceptions of feedback. *Clin Teach*. 2018;15(2):126–31.
 195. Babu KS, Htike MM, Cleak VE. Workplace-based assessments in Wessex: the first 6 months. *Psychiatr Bull*. 2009;33(12):474–8.
 196. Malhotra S, Hatala R, Courneya C-A. Internal medicine residents' perceptions of the Mini-Clinical Evaluation Exercise. *Med Teach*. 2008;30(4):414–9.
 197. Alves de Lima A, Henquin R, Thierier J, Paulin J, Lamari S, Belcastro F, et al. A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. *Med Teach*. 2005;27(1):46–52.
 198. Hunter AR, Baird EJ, Reed MR. Procedure-based assessments in trauma and orthopaedic training-The trainees' perspective. *Med Teach*. 2015;37(5):444–9.
 199. Cohen SN, Farrant PBJ, Taibjee SM. Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools. *Br J Dermatol*. 2009;161(1):34–9.
 200. Simmons M. How are workplace-based assessments viewed by UK psychiatry trainees? *Psychiatr Danub*. 2013;25(SUPPL.2):182–4.
 201. Tailor A, Dubrey S, Das S. Opinions of the ePortfolio and workplace-based assessments: a survey of core medical trainees and their supervisors. *Clin Med*. 2014;14(5):510–6.
 202. Phillips A, Lim J, Madhavan A, Macafee D. Case-based discussions: UK surgical trainee perceptions. *Clin Teach*. 2016;13(3):207–12.
 203. Marriott J, Purdie H, Crossley J, Beard JD. Evaluation of procedure-based

- assessment for assessing trainees skills in the operating theatre. *Br J Surg*. 2011;98(3):450–7.
204. Tsagakataki M, Choudhary A. Mersey deanery ophthalmology trainees' views of the objective assessment of surgical and technical skills (OSATS) workplace-based assessment tool. *Perspect Med Educ*. 2013;2(1):21–7.
 205. Menon S, Winston M, Sullivan G. Workplace-based assessment: survey of psychiatric trainees in Wales. *Psychiatr Bull*. 2009;33(12):468–74.
 206. Pearce I, Royle J, O'Flynn K, Payne S. The record of in-training assessments (RITAs) in urology: An evaluation of trainee perceptions. *Ann R Coll Surg Engl*. 2003;85(5):351–4.
 207. Mukhtar O, Griffin J, Naheed S, Bryant C. The Multiple Consultant Report: new assessment; same old problems. *Clin Teach*. 2018;15(1):34–7.
 208. Dijksterhuis MGK, Schuwirth LWT, Braat DDM, Teunissen PW, Scheele F. A qualitative study on trainees' and supervisors' perceptions of assessment for learning in postgraduate medical education. *Med Teach*. 2013;35(8):e1396-402.
 209. Powell H, DiMarco A, Saeed S. Trainee and trainer opinion on increasing numbers of workplace-based assessments. *Bull R Coll Surg Engl*. 2014;96(5):160–2.
 210. Jackson D, Wall D. An evaluation of the use of the mini-CEX in the foundation programme. *Br J Hosp Med*. 2010;71(10):584–8.
 211. Plymale MA, Donnelly MB, Lawton J, Pulito AR, Mentzer RM. Faculty evaluation of surgery clerkship students: important components of written comments. *Acad Med*. 2002;77(10 Suppl):S45-7.
 212. Sokol-Hessner L, Shea JA, Kogan JR. The Open-Ended Comment Space for Action Plans on Core Clerkship Students' Encounter Cards: What Gets Written? *Acad Med*. 2010;85(10):S110–4.
 213. Deepak S, Ponnampalam A, Brown C. An innovative and effective model of completing WPBAs - CbD (case based discussion) clinics. *Arch Dis Child*.

- 2017;102(5):A83.
214. Mackenzie N, Knipe S. Research dilemmas: Paradigms, methods and methodology. *Issues Educ Res.* 2006;16(2):193–205.
 215. Hammersley M. *Methodological Paradigms in Educational Research: British Educational Research Association On-line Resource.* 2012.
 216. Bunniss S, Kelly DR. Research paradigms in medical education research. *Med Educ.* 2010;44(4):358–66.
 217. Braun V, Clarke V. *Successful Qualitative Research.* First. Carmichael M, editor. SAGE Publications; 2013. 1–382 p.
 218. Morgan DL. Pragmatism as a Paradigm for Social Research. *Qual Inq.* 2014;20(8):1045–53.
 219. Rees CE, Monrouxe L V. Theory in medical education research: how do we get there? *Med Educ.* 2010;44(4):334–9.
 220. Mertens D. *Research and Evaluation in Education and Psychology: integrating diversity with quantitative, qualitative, and mixed methods.* 3rd ed. California: SAGE Publications; 2010.
 221. Bernardin HJ, Beatty RW. *Performance appraisal: Assessing human behaviour at work.* McHugh JB, editor. California, United States of America: Kent Publishing Company; 1984. 273–274 p.
 222. The Australasian College of Dermatologists. *Training Program Handbook Dermatology 2019.* 2019. p. 1–56.
 223. Pope C, Mays N. Qualitative Research: Reaching the Parts Other Methods Cannot Reach: An Introduction to Qualitative Methods in Health and Health Services Research. *Br Me.* 1995;311(July):42–5.
 224. Dillman DA, Smyth JD, Christian LM. *Internet, phone, mail and mixed-mode surveys.* Fourth. New Jersey: Wiley; 2014. 1–509 p.
 225. Polgreen PM, Septimus EJ, Parry MF, Beekmann SE, Cavanaugh JE, Srinivasan

- A, et al. Relationship of Influenza Vaccination Declination Statements and Influenza Vaccination Rates for Healthcare Workers in 22 US Hospitals. *Infect Control Hosp Epidemiol*. 2008;29(7):675–7.
226. DiCicco-Bloom B, Crabtree BF. The qualitative research interview. *Med Educ*. 2006;40(4):314–21.
 227. Cook DA, Beckman TJ. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Adv Heal Sci Educ*. 2009 Dec 1;14(5):655–64.
 228. Hughes LJ, Mitchell ML, Johnston ANB. Just how bad does it have to be? Industry and academic assessors' experiences of failing to fail – A descriptive study. *Nurse Educ Today*. 2019;76(December 2018):206–15.
 229. Gray JD. Global rating scales in residency education. *Acad Med*. 1996;71(1):S55–63.
 230. Timmerman PD. The Discretionary Use of Electronic Media: Four Considerations for Bad News Bearers. *J Bus Commun*. 2005;42(4):379–89.
 231. Molinsky A. 5 key mistakes when delivering bad news [Internet]. *Leadership and Management*. 2016 [cited 2016 Feb 25]. Available from: [file:///Users/catherinescarff/Desktop/PhD articles/5 Key Mistakes When Delivering Bad News %7C Andy Molinsky %7C Pulse %7C LinkedIn.html](file:///Users/catherinescarff/Desktop/PhD%20articles/5%20Key%20Mistakes%20When%20Delivering%20Bad%20News%20-%20Andy%20Molinsky%20-%20Pulse%20-%20LinkedIn.html)
 232. Gbadamosi G, Ross C. Perceived stress and performance appraisal discomfort: The moderating effects of core self-evaluations and gender. *Public Pers Manage*. 2012;41(4):637–59.
 233. Roman BJB, Trevino J. An approach to address grade inflation in a psychiatry clerkship. *Acad Psychiatry*. 2006;30(2):110–5.
 234. Weaver CS, Humbert AJ, Besinger BR, Graber JA, Brizendine EJ. A More Explicit Grading Scale Decreases Grade Inflation in a Clinical Clerkship. *Acad Emerg Med*. 2007;14(3):283–6.
 235. Dudas RA, Barone MA. Setting standards to determine core clerkship grades in

- pediatrics. *Acad Pediatr*. 2014;14(3):294–300.
236. Schmahmann JD, Neal M, MacMore J. Evaluation of the assessment and grading of medical students on a neurology clerkship. *Neurology*. 2008;70(9):706–12.
 237. Grover S, Swisher-McClure S, Sosnowicz S, Li J, Mitra N, Berman AT, et al. Grade Inflation in Medical Student Radiation Oncology Clerkships: Missed Opportunities for Feedback? *Int J Radiat Oncol*. 2015;92(4):740–4.
 238. Varney A, Todd C, Hingle S, Clark M. Description of a Developmental Criterion-Referenced Assessment for Promoting Competence in Internal Medicine Residents. *J Grad Med Educ*. 2009;1(1):73–81.
 239. Nulty DD. The adequacy of response rates to online and paper surveys: what can be done? *Assess Eval High Educ*. 2008;33(3):301–14.
 240. Sheehan KB. E-mail survey response rates: a review. *J Comput Commun*. 2001;6(2):1–19.
 241. Mulliner E, Tucker M. Feedback on feedback practice: perceptions of students and academics. *Assess Eval High Educ*. 2017;42(2):266–88.
 242. Evans AW, Leeson RMA, Petrie A. Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. *Med Educ*. 2007;41(9):866–72.
 243. Violato C, Lockyer J. Self and Peer Assessment of Pediatricians, Psychiatrists and Medicine Specialists: Implications for Self-Directed Learning. *Adv Heal Sci Educ*. 2006 Aug 1;11(3):235–44.
 244. Menon S, Winston M, Sullivan G. Workplace-based assessment: attitudes and perceptions among consultant trainers and comparison with those of trainees. *Psychiatrist*. 2012;36(1):16–24.
 245. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology*. 1990;1(1):43–6.
 246. Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H.

- Competencies “Plus”: The Nature of Written Comments on Internal Medicine Residents’ Evaluation Forms. *Acad Med*. 2011;86(10):S30–4.
247. Ginsburg S, Eva K, Regehr G. Do In-Training Evaluation Reports Deserve Their Bad Reputations? A Study of the Reliability and Predictive Ability of ITER Scores and Narrative Comments. *Acad Med*. 2013;88(10):1539–44.
 248. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK general medical council colleague questionnaires. *Med Educ*. 2009;43(8):757–66.
 249. Durning SJ, Hanson J, Gilliland W, McManigle JM, Waechter D, Pangaro LN. Using Qualitative Data From a Program Director’s Evaluation Form as an Outcome Measurement for Medical School. *Mil Med*. 2010;175(6):448–52.
 250. Battistone MJ, Pendleton B, Milne C, Battistone ML, Sande MA, Hemmer PA, et al. Global descriptive evaluations are more responsive than global numeric ratings in detecting students’ progress during the inpatient portion of an internal medicine clerkship. *Acad Med*. 2001;76(10 Suppl):S105–7.
 251. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: Faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296–306.
 252. Smither JW, Walker AG. Are the characteristics of narrative comments related to improvement in multirater feedback ratings over time? *J Appl Psychol*. 2004;89(3):575–81.
 253. Crano WD, Brewer MB, Lac A. Principles and methods of social research. Third. Routledge; 2015.
 254. Frohna A, Stern D. The nature of qualitative comments in evaluating professionalism. *Med Educ*. 2005;39(8):763–8.
 255. Shulruf B, Adelstein BA, Damodaran A, Harris P, Kennedy S, O’Sullivan A, et al. Borderline grades in high stakes clinical examinations: Resolving examiner uncertainty. *BMC Med Educ*. 2018;18(1):1–9.

256. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Acad Med*. 2010;85(5):780–6.
257. Brutus S. Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. *Hum Resour Manag Rev*. 2010;20(2):144–57.
258. Fishbein M. The role of theory in HIV prevention. *AIDS Care*. 2000;12(3):273–8.
259. Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM. Mixed Messages or Miscommunication? Investigating the Relationship Between Assessors' Workplace-Based Assessment Scores and Written Comments. *Acad Med*. 2017;92(12):1774–9.
260. Goodyear HM, Lakshminarayana I, Wall D, Bindal T. A multisource feedback tool to assess ward round leadership skills of senior paediatric trainees: (2) Testing reliability and practicability. *Postgr Med J*. 2015;91(1075):268–73.
261. Luhanga FL, Larocque S, MacEwan L, Gwekwerere YN, Danyluk P. Exploring the Issue of Failure to Fail in Professional Education Programs: A Multidisciplinary Study. *J Univ Teach Learn Pract*. 2014;11(2).
262. Rees CE, Knight L V., Cleland J a. Medical educators' metaphoric talk about their assessment relationships with students : ' you don ' t want to sort of be the one who sticks the knife in them ' . *Assess Eval* 2009;34(4):455–67.
263. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: A synthesis of recommendations. *Acad Med*. 2014;89(9):1245–51.
264. Leech BL. Asking Questions : Techniques for Semistructured Interviews. *PS Polit Sci Polit*. 2002;35(I):665–8.
265. Liamputtong P. Qualitative data analysis: conceptual and practical considerations. *Heal Promot J Aust*. 2009;20(2):133–9.

266. Sandelowski M. Whatever Happened to Qualitative Description? *Res Nurs Health*. 2000;23(4):334–40.
267. Sandelowski M. What's in a name? Qualitative description revisited. *Res Nurs Health*. 2010 Feb;33(1):77–84.
268. Torff B. Folk Belief Theory, the Rigor Gap, and the Achievement Gap. *Educ Forum*. 2014;78(2):174–89.
269. Moffett J, Crawford R, Pawlikowska T. Enhancing clinical educator well-being. *Clin Teach*. 2019;16(4):306–11.
270. Bearman M, Molloy E, Ajjawi R, Keating J. 'Is there a Plan B?': clinical educators supporting underperforming students in practice settings. *Teach High Educ*. 2013 Jul;18(5):531–44.
271. Castanelli DJ, Weller JM, Chander AR, Molloy EK, Bearman ML. A balancing act: The Supervisor of Training role in anaesthesia education. *Anaesth Intensive Care*. 2019;0310057X1985359.
272. Jones CL, Nanda R. Assessing a doctor you've rarely worked with: The use of workplace-based assessments in a busy inner city emergency department. *EMA - Emerg Med Australas*. 2016;28(4):439–43.
273. Castanelli DJ, Weller JM, Molloy E, Bearman M. Shadow systems in assessment : how supervisors make progress decisions in practice. *Adv Heal Sci Educ*. 2019;(Sep 2019).
274. Whitehouse A, Hassell A, Wood L, Wall D, Walzman M, Campbell IAN. Development and reliability testing of TAB a form for 360 assessment of Senior House Officers ' professional behaviour , as specified by the General Medical Council Assessing doctors ' professional behaviour. *Med Teach*. 2005;27(3):252–8.
275. Varpio L, Ajjawi R, Monrouxe L V., O'Brien BC, Rees CE. Shedding the cobra effect: Problematising thematic emergence, triangulation, saturation and member checking. *Med Educ*. 2017;51(1):40–50.

276. Irvine A, Drew P, Sainsbury RD. Am I not answering your questions properly? : Clarification, adequacy and responsiveness in semi-structured telephone and face-to-face interviews. *Qual Res.* 2012;13(1):87–106.
277. Trier-Bieniek a. Framing the telephone interview as a participant-centred tool for qualitative research: a methodological discussion. *Qual Res.* 2012;12(6):630–44.
278. Morse JM. How different is qualitative health research from qualitative research? Do we have a subdiscipline? *Qual Health Res.* 2010 Nov;20(11):1459–64.
279. Duffy K. Failing students : a qualitative study of factors that influence the decisions regarding assessment of students ' competence in practice. English. 2003. 1–88 p.
280. Luhanga F, Yonge OJ, Myrick F. 'Failure to assign failing grades': issues with grading the unsafe student. *Int J Nurs Educ Scholarsh.* 2008;5(1).
281. Williams SE, Dewey CM. Identification of training opportunities in medical education for academic faculty. *Med Teach.* 2019;41(8):912–6.
282. Joint Royal Colleges of Physicians Training Board. Specialty Training Curriculum for Dermatology. Vol. 2010. London; 2012. 1–93 p.
283. Dermatology ARCP Decision Aid. 2014.
284. Wilkinson TJ, Tweed MJ. Deconstructing programmatic assessment. *Adv Med Educ Pract.* 2018;Volume 9:191–7.
285. Harrison CJ, Könings KD, Schuwirth LWT, Wass V, Van der Vleuten CPM. Changing the culture of assessment: the dominance of the summative assessment paradigm. *BMC Med Educ.* 2017;17(1):1–15.
286. Tweed M, Wilkinson T. Student progress decision-making in programmatic assessment: Can we extrapolate from clinical decision-making and jury decision-making? *BMC Med Educ.* 2019;19(1).
287. Sapp JE, Torre DM, Larsen KL, Holmboe ES, Durning SJ. Trust in Group Decisions: a scoping review. *BMC Med Educ.* 2019;19(309):1–13.

288. Wozniak H, Campbell N. Work-based peer review of clinical supervision practice: A guide to creating a culture of quality supervision. A Greater Northern Australia Regional Training Network (GNARTN) Project with Flinders University, Darwin, Australia. 2014.
289. Brewster DJ, Waxman BP. Adding kindness at handover to improve our collegiality: the K-ISBAR tool (Letter in reply). *Med J Aust.* 2019;211(2):93–4.
290. Kneebone R, Nestel D. Learning clinical skills – the place of simulation and feedback. *Clin Teach.* 2005;2(2):86–90.
291. Lyon-Maris J, Scallan S. High fidelity trainee simulation to improve trainer performance. *Clin Teach.* 2012;9(1):18–21.
292. Farrant P, Cohen SN, Burge SM. Attributes of an effective trainer: Implications of the views of U.K. dermatology trainees. *Br J Dermatol.* 2008;158(3):544–8.
293. Benbassat J. Undesirable features of the medical learning environment: A narrative review of the literature. *Adv Heal Sci Educ.* 2013;18(3):527–36.
294. Frank JL, Reibel D, Broderick P, Cantrell T, Metz S. The effectiveness of mindfulness-based stress reduction on educator stress and well-being: Results from a pilot study. *Mindfulness (N Y).* 2015;6(2):208–16.
295. Lefroy J, Hawarden A, Gay SP, McKinley RK, Cleland J. Grades in formative workplace-based assessment: a study of what works for whom and why. *Med Educ.* 2015;49(3):307–20.
296. Australian Medical Association. AMA Victoria - Victorian Public Health Sector - Medical Specialists Enterprise Agreement 2018-2021. 2018. p. 1–115.
297. Tavares W, Young M, Gauthier G, St-Onge C. The effect of foregrounding intended use on observers' ratings and comments in the assessment of clinical competence. *Acad Med.* 2019;
298. Dudek NL, Marks MB, Wood TJ, Lee a. C. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ.* 2008;42(8):816–22.

299. Bismil R, Dudek NL, Wood TJ. In-training evaluations: Developing an automated screening tool to measure report quality. *Med Educ.* 2014;48(7):724–32.
300. Barrett A, Galvin R, Steinert Y, Scherpbier A, O’Shaughnessy A, Horgan M, et al. A BEME (Best Evidence in Medical Education) systematic review of the use of workplace-based assessment in identifying and remediating poor performance among postgraduate medical trainees. *Syst Rev.* 2015;4(1):1–6.
301. Cook DA, Reed DA. Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. *Acad Med.* 2015;90(8):1067–76.
302. Sullivan GM. Deconstructing Quality in Education Research. *J Grad Med Educ.* 2011;3(2):121–4.
303. Yarris LM, Simpson D, Sullivan M. How Do You Define High-Quality Education Research? *J Grad Med Educ.* 2013;(June):180–1.
304. Lin H, Lin E, Auditore S, Fanning J. A Narrative Review of High-Quality Literature on the Effects of Resident Duty Hours Reforms. *Acad Med.* 2016;91(1):140–50.
305. Bache J, Brown J, Graham D. In-training assessment for specialist registrars: views of trainees and trainers in the Mersey Deanery. *J R Soc Med.* 2002;95(12):612–3.
306. Goodyear H, Wall D, Bindal T. Annual review of competence: Trainees’ perspective. *Clin Teach.* 2013;10(6):394–8.
307. Kolli VB, de Waal H, Shivashankar S, Rubinsztein J. Two years after introduction, trainees remain unconvinced about WPBAs and ARCPs. *Psychiatrist.* 2010;34(3):114.
308. McGraw M. The views of paediatric specialist registrars on the annual record of in-training assessments. *Med Educ.* 2003;37(6):571.

APPENDICES

Appendix A Quality level determination for systematic review articles

Quality assessment was performed using the MERSQI tool (183) for quantitative studies and the Walsh and Downe tool (184) for qualitative. The BEME quality assessment tool (300) was also examined but was not used for this review as a separate tool was desired for each type of study (qualitative and quantitative), which could provide a greater depth of information as compared with a tool that covered both in one. Cook and Reed have shown that the MERSQI and Newcastle-Ottawa Scale – Education (NOS-E) are reasonably well correlated (301) and the former was chosen given its increasing use (302). It is acknowledged that quality assessment of studies is a dynamic and evolving area (303). While definitions of scores which correlate with high-quality studies have been defined by some authors [for example (304)], Cook and Reed advocate the use of scores for “relative rather than absolute judgements” as the scores generated are highly dependent on the area of study and the inclusion criteria (301). Walsh and Downe echo this recommendation for qualitative study assessment, suggesting that while they would expect their overall criteria to be addressed sufficiently in studies they would include in a primary meta-synthesis, they offer their tool to be used “imaginatively rather than prescriptively” (184).

For qualitative study assessment, the original article (184) states that it would be expected that higher quality studies would have “yes” to the criteria in column 2 and some purely qualitative studies demonstrate this (187). Many other qualitative studies however, showed weaknesses in the audit trail and reflexivity categories, so in the current context, a study has been considered higher quality if it met 10 or more criteria.

Many of the other studies included in this study are primarily quantitative or mixed, and the qualitative criteria have been applied to the qualitative aspects only. In this group, other aspects of the studies were also noted to be weak for example “Demonstration of clear audit trail” (explaining how they obtained the information – free-text comments etc.) and “Data used to support interpretation”, as there was often not extensive use of the quotes in discussion of findings. Rather, these data were used to support the quantitative data.

In regards to quantitative studies, a higher quality study in this context was considered as one with a MERSQI score of ≥ 7 . This score was determined because the research question for the review itself precluded obtaining a high score. In explanation, given the research question asked for trainee views and perspectives, the studies included were therefore mainly observational ones, with assessment by study participant and with an outcome of perception/attitude etc. only. This limited the potential score that could be achieved. As such the maximum score for most studies was considered to be 12.

Appendix B Articles included in the systematic review

Appendix B: Data extracted from included articles

Author, year	Assessment type, Country	Research aim/objective	Data type, data collection method Location	Participants – respondents Trainee type – year, specialty	Key findings relevant to the current research question	Quality assessment ¹
Alofs, Huiskes, Jan Heineman, Buis, Horsman, Van der Plank & ten Cate, 2015 (7)	MSF Netherlands	An evaluation to collect opinions on the use of the MSF and its effect (on Kirkpatrick levels) to enable improvements if necessary	Quantitative Survey – online (anonymous)	108/404 residents in anaesthesiology, OBGYN, internal medicine, paediatrics, ENT, etc. up to 6 years (22/59 programme directors)	MSF is a useful feedback tool, especially in relation to the breadth of assessors, the anonymous nature and the written comments. Risk of biases exist and it is not applicable to all specialties.	MERSQI higher – 7
Alves de Lima, Henquin, Thierer, Paulin, Lamari, Belcastro & van der Vleuten, 2005 (197)	Mini-CEX Argentina	To focus on how residents perceive an assessment tool and its influences on their approach to learning	Qualitative Open and semi-structured interview Buenos Aires province, Argentina	16/16 cardiology residents, all 4 years of training	Mini-CEX is a valuable real-life learning exercise, but anxiety and biases can interfere with the results.	W&D lower – 2
Babu, Htike, & Cleak, 2009 (195)	WBA (8 types) UK	To determine the experience of supervisors and trainees with the first 6 months of using WBAs	Quantitative Survey (anonymous) Wessex	39/62 psychiatry trainees, all training years (1–4) (40/65 supervisors)	WBAs found very/useful by only 39% of trainees and barriers to their completion identified. Useful to identify strengths and weaknesses though tools vary in their ability to do so.	MERSQI higher – 7.5

Bache, Brown, & Graham, 2002 (305)	RITA UK	To ascertain whether the broadened review process in the Mersey Deanery was thought beneficial	Quantitative Survey – paper Mersey Deanery	605/744 specialty trainees (309/349 trainers)	The RITA provides a useful review of a training period and trainee performance and is able to provide career advice and determine the educational value of the post.	MERSQI higher – 7
Barrett, Galvin, Scherpbier, Teunissen, O'Shaughnessy, Horgan 2016 (6)	WBA (Mini-CEX, DOPS, CbD) Ireland	What are trainer's and trainee's perceptions of the learning value of WBAs? How have trainer and trainee WBA experiences shaped these perceptions?	Qualitative Semi-structured interviews Royal College of Physicians, Ireland	8 medicine trainees, all in first three years of higher specialist training (9 trainers)	Value of WBA in ability to approach or force trainer for feedback. Potential learning exercise, but barriers noted, especially engagement in process. WBA are not being reified as the formative assessments they were intended to be.	W&D higher – 11
Bindal, Wall, & Goodyear, 2011 (186)	WBA (Mini-CEX, DOPS, CbD) UK	To explore whether WBAs had been successfully integrated into run-through paediatric specialty training programmes	Quantitative 20-item questionnaire West Midlands Deanery	130/150 regional paediatric trainees	Trainees are ambivalent about the ability of the WBA to reflect their capabilities and opinions on their usefulness vary widely. Issues of variability in scoring and poor practices by assessors noted. Concerns about tick-box mentality during completion.	MERSQI higher – 10 W&D higher – 10
Bodgener & Tavabie, 2011 (23)	CbD UK	To explore the attitudes of trainees and trainers towards the CbD, its educational value and its ability to assess trainee performance	Mixed Questionnaire and semi-structured interviews (different participants) Kent, Surrey and Sussex GP Deanery	Questionnaire – 45/45 final year GP trainees (30 or 31/40 GP trainers) Interviews – 5 trainees (5 trainers)	Feedback is overwhelmingly useful for development and to show trainee progress, though not all agree that CbD assessed performance and this may relate to the view of the assessment as task-centred. The trainer's feedback giving skills are noted to be key.	MERSQI higher – 8 W&D lower – 7

Castanelli, Jowsey, Chen, & Weller, 2016 (187)	Mini-CEX Australia, New Zealand	To explore trainees' and supervisors' understanding of the mini-CEX, their experience with the assessment and their perceptions of its influence on learning and supervision.	Qualitative Semi-structured telephone interviews Australia and New Zealand	17 trainees across all years of training and diverse practice locations anaesthesia (18 supervisors)	Scepticism expressed by a number in relation to leniency and variability and poor practice by trainees. The breadth of feedback is appreciated and narrative feedback valued over ratings. Views conflict as to purpose of assessment – formative or summative and this directs assessment behaviours.	W&D higher – 12
Cohen, Farrant, & Taibjee, 2009 (199)	WBA (Mini-CEX, DOPS, MSF) UK	To gather the views of trainees on the WBA to assess whether or not the newly introduced assessments were working as planned and to identify any problems with them.	Quantitative Questionnaire – paper and electronic All UK dermatology trainees	138/269 dermatology trainees all in UK	Feedback is appreciated, especially when timely, though not all trainees report receiving useful feedback. Reassurance is valued, but frustrations with conflict in results (in DOPS and MSF) and lack of training noted. Some report the process to be stressful and artificial.	MERSQI higher – 7.5
Findlay, 2012 (54)	MSF UK	To explore the experiences of GP specialty training registrars, of the ways in which MSF impacts upon their self-perceptions and professional practice	Qualitative Semi-structured interviews Northumbria GP training programme	5 GP trainees	The educational supervisor plays a key role in helping the trainee to interpret and develop from feedback in the MSF. MSF is useful but open to misuse, including by trainee manipulation and a reluctance of raters to provide constructive feedback. Logistical issues hamper use.	W&D lower – 6

Gaunt, Patel, Rusius, Royle, Markham and Pawlikowska, 2017 (109)	WBA UK	To explore how WBAs influence the ways surgical trainees seek feedback and feedback interactions	Qualitative Focus groups West Midlands, London, Wales and Yorkshire	42 surgical trainees All levels of training	Trainees' perceptions of the purpose of WBA and their relationship with their assessors impacts on how they use WBA and what they can achieve from it. Two distinct trainee types were seen – those who view assessment as “of” learning play the game; those who view if “for” learning seek to develop through negative feedback. The implications of the WBA affected this too – that is informing the annual review of performance. Timing of feedback is important.	W&D higher – 10
Goel & Singh, 2015 (192)	Mini-CEX India	To assess the acceptability and feasibility of mini-CEX in different paediatric settings and to know if this exercise helped the residents by improved learning	Quantitative Questionnaire Ludhiana	23 paediatric residents (11 teachers)	Mini-CEX provides useful feedback to help learning. Some noted the assessor level is of relevance and that anxiety may play a role.	MERSQI lower – 4.5
Goodyear, Wall, & Bindal, 2013 (306)	Annual review of competence progression (ARCP) UK	To explore specialty trainees' views on ARCPs including the usefulness of the annual review and how well prepared they felt for it.	Mixed Questionnaire, telephone semi-structured interviews West midlands deanery	Questionnaire: 63/70 (years 1–4) Interview: 21 (years 2–6) Paediatric trainees	The majority agreed that the ARCP is helpful, though did not feel prepared for it and were dissatisfied with the process. Changes in the process and training are relevant. Some feedback is not a true reflection and misinterpretation of assessment messages is possible.	MERSQI higher – 8 W&D lower – 4
Grennan, Crowley, Quidwai, Barrett, & Kooblall, 2016 (193)	ePortfolio and WPBAs Ireland	To determine if trainees and supervisors see the ePortfolio as an important learning exercise or as a time-consuming obligation	Quantitative Online survey 2 hospitals in Dublin	27/149 trainees, SHO's completing BST in clinical medicine (24/307 consultants)	There is significant dissatisfaction among trainees in relation to the ePortfolio (of which WBA are part). Engagement is key. Less than half agreed the feedback is useful and feedback is only given promptly in 1/3 cases (never in a significant minority). Some trainees manipulate case selection to avoid negative feedback.	MERSQI higher – 7

Hunter, Baird, & Reed, 2015 (198)	PBA UK	To gain an understanding of the attitudes of trauma and orthopaedic trainees regarding procedure-based assessments and identify factors influencing any perceived educational benefit.	Quantitative Internet-based questionnaire All UK trainees	616/1144 Trauma & Orthopaedic trainees Year 3 to post-CCT trainees	Formative role valued over a summative one. Over half the trainees felt the verbal feedback a useful learning tool, but only a minority felt they improved their practice due to this. Timeliness of completion of the PBA (reflecting the enthusiasm of the assessor) had the most significant effect on perceptions of satisfaction and outcomes of the tool. Half perceived the exercise to be a form-filling one and differing standards of assessors noted.	MERSQI higher – 10.5
Kolli, de Waal, Shivashankar, & Rubinsztein, 2010 (307)	WBA UK	Nil stated	Quantitative Questionnaire	Psychiatry trainees (number not known)	Significant majority of trainees rate their satisfaction with quality of feedback received as at least 3/5, but almost half report that the ARCP is not meaningful.	MERSQI lower – 5
Malhotra, Hatala, & Courneya, 2008 (196)	Mini-CEX Canada	To assess residents' perceptions of the mini-CEX using qualitative methods	Qualitative Semi-structured focus groups (2) UBC Vancouver	12 internal medicine residents – year 1 and 2	The use of the mini-CEX as an assessment tool placed limitations on its perceived value as an educational one. Anxiety influences performance but decreases with experience.	W&D lower – 9
Marriott, Purdie, Crossley, & Beard, 2011 (203)	PBA UK	To evaluate the validity, reliability and acceptability of procedure-based assessments in surgical training programmes.	Quantitative Questionnaire at end of study period Three hospitals in Sheffield	81 trainees Surgery – all types and all levels of trainee (focus on years 3–7)	Trainees report comfort with the dual purposes of the tool for both learning and assessment. Most are satisfied with the feedback received, though some received no feedback. Trainees report the feedback received to be moderately to very useful.	MERSQI higher – 12.5
McGraw, 2003 (308)	RITA UK	Review of the trainees' perception of the annual record of in-training assessments.	Quantitative Questionnaire Southwest region	34/50 trainees Paediatrics	The majority of trainees felt they had received positive and constructive comments on their portfolios and detailed discussion on future training needs.	MERSQI higher – 7.5

Mehta, Brown, & Shaw, 2013 (191)	CbD UK	To explore trainee's perceptions of the educational value of case-based discussions specifically focusing on feedback.	Mixed Questionnaire (online), interviews Mersey Deanery	Questionnaire: 26/32 Interviews: 9 interviewed from 12 volunteers Paediatric trainees, years 1 and 2	Opinions on usefulness vary significantly but many found the tool has potential for learning, is fair and provides a reasonable picture of level of competence but can be subjective. Trainees value narrative feedback over ratings. Trainees note limitations to the tool such as poor practices by trainees and assessors and environmental constraints. Limitations on generation of learning points and action plans were noted. Conflict between an assessment and learning tool noted.	MERSQI higher – 10 W&D lower – 8
Menon, Winston, & Sullivan, 2009 (205)	WBA UK, Wales	To explore attitudes and perceptions regarding workplace-based assessment (WPBA) among psychiatric trainees in Wales, identify current problems and recommend measures to successfully implement WPBA.	Quantitative Questionnaire (anonymous, cross-sectional, online) Wales	81/88 psychiatric speciality trainees (minus one who was author) All levels of training (years 1–6 plus “other”)	There were widespread negative attitudes to the WBAs. Most trainees felt they have no real educational benefit and are of dubious validity. Many expressed concerns that assessor inconsistencies, including background, training and the impact of the personal preferences, may result in an inaccurate assessment outcome. Concerns raised about the potential for the tools to become tick-box exercises. Both trainers and trainees need training in education to avoid unintended outcomes of the process.	MERSQI higher – 8
Mohanaruban, Flanders, Rees, 2018 (194)	CbD UK	To explore how medical trainees perceive the feedback they receive during CbDs and how these assessments are being put into practice in the workplace.	Mixed Questionnaire (paper-based, distributed to whole trainee population) Two focus groups District general hospital, London, UK	78/91 medical trainees completed questionnaire, 12 volunteered to participate in focus groups All levels of training	Feedback occurs in less than 5 minutes in most cases, but one third report not receiving feedback and less than half report it is useful. Most trainees do address learning points arising from the feedback though think feedback is not specific enough and assessors are too lenient and afraid of failing trainees. Assessor engagement in process is of key importance.	W&D lower – 5 MERSQI higher – 7

Mukhtar, Griffin, Naheed, Bryant, 2017 (207)	MCR UK	To explore opinions of medical trainees on the MCR	Qualitative Questionnaire (open questions and free text) and focus group A London teaching hospital	22/42 medical trainees completed questionnaire, 6 volunteered to participate in focus group Junior trainees	A lack of understanding of clear purpose of the assessment lead it to be treated like a tick-box exercise. Ineffective feedback limited the potential impact of the assessment. The assessor completing the form may not be the most appropriate one to assess the trainee	W&D lower – 3
Patel, Drover, & Chafe, 2015 (188)	ITER Canada	To explore the perspectives of faculty and residents on structure and process of ITER as a means to improving the use of the ITER as an evaluation tool.	Qualitative Two focus groups – trainees and assessors Memorial University of Newfoundland paediatric training program	8/23 invited paediatric resident attended All training years	Trainees unanimously agreed that there is lack of understanding about scale use. Written comments were felt to be the most useful aspect of the ITER. Residents greatly appreciate the effort that assessors go to in their assessments. The ITER is useful but not being used as designed and attention to factors discussed could improve the quality of feedback.	W&D higher – 10
Pearce, Royle, O'Flynn, & Payne, 2003 (206)	RITA UK	To evaluate the effect of the Calman reforms on trainees in urology with respect to their educational goals, their experience of the RITA process and its value in preparing them for consultant careers.	Quantitative Questionnaire (anonymous, paper) All UK urology trainees holding national training numbers who had completed at least one RITA review	100 replies Urology trainees All years	Trainees described fair satisfaction with the RITA process. Less than half set educational goals and a large minority felt the process would not identify a failing trainee. The majority felt that a failing assessment results from a lack of the training establishment fulfilling its educational requirements.	MERSQI lower – 6.5

Phillips, Lim, Madhavan, & Macafee, 2016 (202)	CbD UK	To review the use of CbDs by surgical trainees to determine if and when they are useful, and whether they are perceived as being used correctly.	Qualitative Interviews Surgical trainees in two deaneries	42 surgical trainees – 21 higher surgical trainees, 21 core surgical trainees, in all subspecialties	Not all CbDs are equal. Potentially useful to improve performance, but many factors impact, including stage of training (being most useful at lower levels) and quality of feedback. Opinions vary from CbD being a tick-box task to complete only, through to a valuable learning experience. Poor practice and lack of engagement in the process by trainees and assessors reported.	W&D lower – 7
Ringsted, Henriksen, Skaarup, & Van Der Vleuten, 2004 (5)	ITA Netherlands	To investigate the experiences and opinions of programme directors, assessors and trainees about a recently introduced ITA programme	Qualitative Semi-structured interviews 3 departments of anaesthesiology at different hospitals	14 first year trainees in anaesthesiology (3 programme directors, 9 supervisors)	Viewing ITA as an educational process rather than just documentation of performance dictated perceived value. Trainees reported valuing the learning experience even when they did not pass an assessment. Successful completion of assessments viewed by some to increase confidence. The importance of immediate feedback was noted, though some assessors were known to be lenient.	W&D lower – 8
Sabey & Harris, 2011 (189)	WBA UK	To establish how the new system of WBA is working in day-to-day practice in hospital posts.	Mixed Questionnaire – Focus groups and semi-structured interviews 2 training locations in Severn Deanery	Questionnaire – 52/78 GP trainees, Focus groups and semi-structured interviews – 20/31 trainees	Trainees cited narrative feedback as useful more than scores, though noted that the quality of feedback varies. Poor practice by assessors and trainees noted, including bias and lack of honesty in judgements seen to limit the scope for learning, validity and usefulness and risks degeneration to a tick-box exercise. Views about the usefulness of the various assessments varied.	MERSQI higher – 8 W&D higher – 10

Shalhoub, Marshall, Ippolito, 2017 (8)	PBA UK	To explore the opinions of trainees on their PBA experiences – how they perceive and make sense of them.	Qualitative Semi-structured interviews Surgical trainees across UK who had used the Intercollegiate Surgical Curriculum Programme for >1 year	10 surgical trainees all levels of training	Lack of clarity of the purpose of the tool leads to problems. Tool is not used as intended and a quota to complete encourages tick-box behaviours. Poor practices are noted and the feedback is seen as more important than the grade assigned, though feedback is hard to obtain. The tool standardises assessors.	W&D higher – 10
Simmons, 2013 (200)	WBA UK	To explore trainees' perceptions of opportunities, barriers and suggestions for development of WBAs as formative learning tools	Qualitative Two focus group interviews Cambridge	6 psychiatry trainees first and fourth year trainees	Trainees note the opportunity for personalised feedback and to detect trainees in difficulty. Poor practice of trainees noted (e.g. choosing assessor based on their leniency). Concerns that WBA are not appropriate for psychiatry, and that there is little room for formative feedback on forms. The need for training of assessors and trainees to enhance outcomes was noted.	W&D lower – 4
Taylor, Dubrey, & Das, 2014 (201)	WBA portfolio UK	To assess core medical trainee and trainer opinions of the ePortfolio and WPBAs.	Quantitative Surveys (paper-based) Five NHS trusts in north-west London	60/158 core medical trainees (30/62 trainers)	The vast majority viewed WBA as “means to passing the year”, with only one third viewing them as a learning exercise. Poor practices by trainees and assessors reported. Delays in feedback reported only 20% felt it had improved their practice. Direct verbal feedback felt to be essential.	MERSQI higher – 8

Tsakataki & Choudhary, 2013 (204)	Objective assessment of surgical and technical skills (OSATS) UK	To gather the views of a sample of UK ophthalmology trainees on the OSATS WPBA for cataract surgery for its acceptability and to identify potential problems	Quantitative Questionnaire – paper and online Mersey Deanery	28/34 ophthalmology trainees	Strong opinions for and against the tool. Some trainees valued the opportunity to be observed and the resulting structured formative feedback allowing for reflective practice. Some felt feedback was unhelpful. Many trainees report feeling stressed and senior trainees reported insufficient time for feedback. Concerns noted that assessments can strain relationships and encourage tick-box behaviours.	MERSQI higher – 8
Watling, Kenyon, Zibrowski, Schulz, Goldszmidt, Singh, Maddocks, Lingard, 2008 (102)	ITER Canada	To describe residents' experiences with and perceptions of the ITER process to gain insight into why the process succeeds or fails.	Qualitative Semi-structured interviews London, Ontario	20 residents – medical, pediatric and surgical Year 1–5	Engagement of both the assessor and trainee is key to the success of ITER and whether it is perceived to be meaningful. Features which promote engagement by each discussed.	W&D higher – 10

¹ Quality assessment was performed on quantitative aspects of the studies with the MERSQI tool (183), where “higher” was defined as a score of 7 or more, and on qualitative aspects of the studies with the Walsh and Downe tool (W&D) (184), where “higher” was defined as 10 or more “yes” responses to questions.

Appendix C Questionnaire – assessors

Assessor questionnaire <i>Assessor grades and feedback: private and public judgements</i>
Your experience with trainee assessments

* 1. Have you completed performance assessments, either work-based assessments (WBA) and/or summative in training assessments (SITA) for a trainee in the last 12 months and would you like to commence the questionnaire?

- ☐ Yes
☐ No

Assessor questionnaire <i>Assessor grades and feedback: private and public judgements</i>
Information about you

* 2. Approximately how many of the following assessments have you completed in the past 12 months?

	1-2	3-4	5-6	7-10	11-15	16-20	more than 20	none
Derm-CEX (assessment where you observe a trainee conducting an initial consultation with a patient)	☐	☐	☐	☐	☐	☐	☐	☐
Pro-DA (assessment where you observe a trainee's core technical skills while they perform a clinical procedure)	☐	☐	☐	☐	☐	☐	☐	☐
CbD (assessment where you discuss a case with a trainee in order to evaluate their clinical practice, interpretation, decision making, judgement and clinical management)	☐	☐	☐	☐	☐	☐	☐	☐
SITA (assessment of a trainee at a point in time against a number of standards and criteria, performed 4 - 6 monthly depending on rotation length)	☐	☐	☐	☐	☐	☐	☐	☐

3. How long have you been practising as a dermatologist?

- ☐ <5 years
☐ 6 - 10 years
☐ 11 - 15 years
☐ 16 - 20 years
☐ 21 - 25 years
☐ 26 - 30 years
☐ >31 years

4. How many years have you been involved in conducting performance assessments on trainees?

- ☐ 0-<5 years
- ☐ 5-<10 years
- ☐ 10-<15 years
- ☐ 15 years or more

5. In which state or territory do you primarily practise?

- ☐ ACT
- ☐ New South Wales
- ☐ Northern Territory
- ☐ Queensland
- ☐ South Australia
- ☐ Tasmania
- ☐ Victoria
- ☐ Western Australia
- ☐ Prefer not to say

6. Are you:

- ☐ Female
- ☐ Male
- ☐ Prefer not to say

7. Are you a supervisor of training (SOT)?

- ☐ Yes
- ☐ No

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Experiences with performance appraisals

The first part of the questionnaire asks about your experiences with completing WBA and/or SITA forms for trainees. Work has shown that some assessors may feel some discomfort when completing performance appraisals, while others do not. This scale describes several situations that may occur in performance appraisal situations and asks how you would generally feel when in them from 1 = no discomfort, through to 5 = high discomfort.

8. Please rate how you would feel in the following situations.

	1 No discomfort	2	3	4	5 High discomfort
Telling a trainee that he or she must stop coming to work late	☐	☐	☐	☐	☐
Telling a trainee that his or her work is only satisfactory, when you know that he or she expects an above satisfactory rating	☐	☐	☐	☐	☐
Talking to a trainee about his or her performance on the job	☐	☐	☐	☐	☐
Conducting a feedback session with a poorly performing trainee	☐	☐	☐	☐	☐
Asking a trainee if he or she has any comments about your ratings of his or her performance	☐	☐	☐	☐	☐
Telling a trainee who has problems in dealing with others (e.g. patients, other trainees, nursing staff, administrative staff etc) that he or she should do something about it	☐	☐	☐	☐	☐
Telling a male trainee that his performance needs to improve	☐	☐	☐	☐	☐
Responding to a trainee who is upset over your rating of his or her performance	☐	☐	☐	☐	☐
Conducting a feedback session with a trainee who is performing well	☐	☐	☐	☐	☐
Letting a trainee give his or her point of view regarding a problem with performance	☐	☐	☐	☐	☐

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Experiences with performance appraisals

9. Please rate how you would feel in the following situations.

	1 No discomfort	2	3	4	5 High discomfort
Giving a satisfactory rating to a trainee who has done a satisfactory (but not exceptional) job	☐	☐	☐	☐	☐
Letting a trainee talk during a feedback session	☐	☐	☐	☐	☐
A trainee's challenging you to justify your evaluation in the middle of a feedback session	☐	☐	☐	☐	☐
A trainee's accusing you of playing favourites in the rating of trainees	☐	☐	☐	☐	☐
Supporting a recommendation that a trainee be removed from the training program	☐	☐	☐	☐	☐
Telling a trainee that his or her performance can be improved	☐	☐	☐	☐	☐
Warning a poorly performing trainee that unless performance improves, you will support a recommendation for his or her removal from the training program	☐	☐	☐	☐	☐
Telling a female trainee that her performance must improve	☐	☐	☐	☐	☐
Encouraging a trainee to evaluate his or her own performance	☐	☐	☐	☐	☐
Telling a trainee that he or she must stop taking long breaks	☐	☐	☐	☐	☐

This survey instrument (questions 8 to 9) is adapted from the Performance Appraisal Discomfort Scale, Bernardin HJ and Beatty RW. 1984. Performance appraisal: assessing human behaviour at work. Boston, MS: Kent.

Assessor questionnaire

Assessor grades and feedback: private and public judgements

Making and delivering your assessment

This part of the questionnaire asks about situations where you have had some negative information to share with a trainee.

* 10. Have you had to share negative information in relation to a WBA or SITA (such as a low rating or negative feedback) with a trainee within the past twelve months?

- ☐ Yes
- ☐ No

Assessor questionnaire

Assessor grades and feedback: private and public judgements

Making and delivering your assessment

11. The second part of the questionnaire contains items which reflect some of the ways that assessors may act when faced with having to share some negative information (that is a low rating or some negative feedback) with a trainee. Please respond to each statement in a way that best describes what you generally do or have done.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I told them only part of the problem	☐	☐	☐	☐	☐
I didn't go into much detail about the negatives	☐	☐	☐	☐	☐
I left out some on the negative information	☐	☐	☐	☐	☐
I did not point out all of the weaknesses	☐	☐	☐	☐	☐
I avoided the trainee at work, so I would not have to talk to him/her	☐	☐	☐	☐	☐
I asked another consultant to give the trainee any negative information	☐	☐	☐	☐	☐
I avoided it altogether	☐	☐	☐	☐	☐
I left giving the feedback for as long as I could	☐	☐	☐	☐	☐
I steered clear of circumstances in which I'll have to give negative information	☐	☐	☐	☐	☐
I found ways to get out of sharing negative information	☐	☐	☐	☐	☐

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

12. These items focus on how you might have acted when faced with having to share some negative information (that is a low rating or some negative feedback) with a trainee. Please respond to each statement in a way that best describes what you generally do or have done.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I put the negative feedback between two positive comments	☐	☐	☐	☐	☐
I changed the feedback or rating to make it more positive	☐	☐	☐	☐	☐
I tried to make it sound nicer than it was	☐	☐	☐	☐	☐
I changed the feedback or rating so it was not so negative	☐	☐	☐	☐	☐
I tried to make the feedback more positive than it really was	☐	☐	☐	☐	☐
I tried to make the trainee feel better by emphasising positive things	☐	☐	☐	☐	☐

This survey instrument (questions 11 to 12) is adapted from Marler LE, McKee D, Cox S, Simmering M and Allen D. 2012 "Don't make me the bad guy: Organisational norms, self-monitoring and the MUM effect." *Journal of Managerial Issues* XXIV: 97-116.

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

There may have been other ways that you handled the situation.

13. Indicate your agreement with the following descriptions of how you might have acted when faced with having to share negative information with a trainee. Choose the responses that best describe what you generally do or have done.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I changed the feedback during the feedback session due to the trainee's response to it	☐	☐	☐	☐	☐
I gave the feedback as quickly as I could and before the trainee could react to it	☐	☐	☐	☐	☐
I adopted a very formal or clinical manner when I gave the feedback so the trainee would be less likely to question me	☐	☐	☐	☐	☐

Are there any other ways that you use to manage your discomfort in delivering 'negative' feedback to trainees (please specify)?

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

We know that most assessors can think of a time when it was challenging for them to deliver their assessment judgement to a trainee.

14. In the past 12 months, have you marked a trainee as competent in a WBA (Derm -CEX, Pro-DA or CbD) when you did not think they were?

- ☐ Yes
- ☐ No
- ☐ Not sure
- ☐ I haven't completed a work-based assessment in the past 12 months

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

15. The main reason I marked the trainee as competent in a WBA when I did not think they were is because

- ☐ of the impact that marking the trainee as not competent would have had on me - e.g. I'd feel guilty, the trainee may think negatively of me, I'd end up in a negative mood, I wanted to avoid any conflict etc.
- ☐ of the impact that marking the trainee as not competent would have had on the trainee - e.g. the result would have upset the trainee, the result would have effected the trainee's confidence, the trainees prefer to hear good results etc.
- ☐ I **disagreed** with the outcome that would have resulted - e.g. trainee progress being interrupted, trainee being put on performance improvement plan, getting a black mark against his or her name etc.
- ☐ Other (please specify)

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

16. If there has been a time when you marked a trainee as competent when you didn't think they were, did you do anything else also?

- ☐ No, I just marked them as competent
- ☐ Yes, I marked them as competent, but gave accurate verbal feedback directly to the trainee
- ☐ Yes, I marked them as competent, but gave accurate written feedback directly to the trainee
- ☐ Other (please specify)

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

17. Which assessments has this occurred with? Check all that apply.

- ☐ Derm-CEX
- ☐ Pro-DA
- ☐ CbD
- ☐ Other (please specify)

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

18. In the past 12 months, have you found that you marked a trainee as satisfactory in a SITA (that is, **not** given a score of 2 or less on 2 or more areas) when you did not think they were?

- ☐ Yes
- ☐ No
- ☐ No sure
- ☐ I haven't completed a SITA in the past 12 months

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

19. The main reason I marked the trainee as satisfactory in a SITA when I did not think they were is because

- ☐ of the impact that giving an unsatisfactory rating would have had on **me** - e.g. I'd feel guilty, the trainee may think negatively of me, I'd end up in a negative mood, I wanted to avoid any conflict etc.
- ☐ of the impact that giving an unsatisfactory rating would have had on the **trainee** - e.g. the result would have upset the trainee, the result would have effected the trainee's confidence, the trainees prefer to hear good results etc.
- ☐ I **disagreed** with the outcome that would have resulted - e.g. trainee progress being interrupted, trainee being put on performance improvement plan, getting a black mark against his or her name etc.
- ☐ Other (please specify)

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

20. If there has been a time when you didn't give a trainee an unsatisfactory for a SITA, did you do anything else also?

- ☐ No, I just didn't give 2 or more scores of 2 or less
- ☐ Yes, I gave accurate verbal feedback to the trainee directly
- ☐ Yes, I gave accurate written feedback to the trainee
- ☐ Other (please specify)

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

21. At times, assessors can feel pressure to give a trainee a higher rating than the one they feel they have earned. In the past 12 months, have you given a trainee a higher rating on a SITA than you thought was justified?

- ☐ Yes
- ☐ No
- ☐ Not sure

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

22. The main reason I gave a higher rating on a SITA than I thought was justified is because

- ☐ of the impact that giving my real rating would have had on me - e.g. I'd feel guilty, the trainee may think negatively of me, I'd end up in a negative mood, I wanted to avoid any conflict etc.
- ☐ of the impact that giving my real rating would have had on the trainee - e.g. the result would have upset the trainee, the result would have effected the trainee's confidence, the trainees prefer to hear good results etc.
- ☐ I disagreed with the outcome that would have resulted - e.g. trainee progress being interrupted, trainee being put on performance improvement plan, getting a black mark against his or her name etc.
- ☐ Other (please specify)

Assessor questionnaire
Assessor grades and feedback: private and public judgements

Making and delivering your assessment

23. If you gave the trainee a higher score in a SITA than what you thought was justified, did you do anything else also?

- ☐ No, I just gave them the higher score
- ☐ Yes, I gave verbal feedback directly to the trainee
- ☐ Yes, I gave written feedback to the trainee
- ☐ Other (please specify)

Assessor questionnaire Assessor grades and feedback: private and public judgements

Making and delivering your assessment

We know that sometimes trainees approach assessors directly about their assessment results.

24. In the past twelve months, has a trainee:

	Yes	No	N/A
Encouraged or asked you to mark them as competent on a WBA?	☐	☐	☐
Encouraged or asked you to give them a certain rating on a SITA?	☐	☐	☐
Encouraged or asked you to change a score that you have given them on a WBA or a SITA?	☐	☐	☐

Assessor questionnaire Assessor grades and feedback: private and public judgements

Other comments

25. Do you have any comments you would like to make about performance assessments?

Assessor questionnaire Assessor grades and feedback: private and public judgements

Thank you for your participation

Thank you very much for taking the time to complete this survey. I would be grateful if you would avoid discussing your responses with others during the period of data collection. If this questionnaire has raised any issues for you which you wish to discuss further, please contact Prof Steve Trumble, 03 8344 8035, s.trumble@unimelb.edu.au or Dr Cate Scarff, scarffc@student.unimelb.edu.au. If you would like to participate in the next phase of the study, which will involve individual interviews on this topic, further information regarding this will be sent in the coming months. With kind regards,
Cate

Appendix D Questionnaire – trainees

Trainee questionnaire <i>Assessor grades and feedback: private and public judgements</i>
Participant information

* 1. Have you completed at least one performance assessment, that is, either a work-based assessment (WBA) and/or summative in training assessment (SITA) in the last 12 months and would you like to commence the questionnaire?

☐ Yes

☐ No

Trainee questionnaire <i>Assessor grades and feedback: private and public judgements</i>
Information about you

2. Which year of training are you currently in (if training part-time, please convert this to the full-time equivalent year that you are in)?

☐ 1

☐ 2

☐ 3

☐ 4

☐ IMG

3. In which state or territory are you a trainee?

☐ ACT

☐ New South Wales

☐ Northern Territory

☐ Queensland

☐ South Australia

☐ Tasmania

☐ Victoria

☐ Western Australia

☐ Prefer not to say

4. Are you:

- ☐ Female
- ☐ Male
- ☐ Prefer not to say

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Trainee feedback orientation scale

This section asks for your opinions about the ways that you may receive feedback on your performance as a dermatology trainee. It asks for your opinions on feedback processes, including how useful they are to you and your future development.

Thinking about the feedback that you receive in a work-based assessment (WBA, that is the Derm-CEX, Pro-DA or CbD) situation or in a SITA feedback situation, please indicate your level of agreement with the following statements.

5. Usefulness of feedback

Thinking about a typical WBA or SITA feedback situation, please indicate your level of agreement with the following statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I think feedback from my assessors is vitally important in improving my performance	☐	☐	☐	☐	☐
I will usually reflect on an assessor's feedback	☐	☐	☐	☐	☐
I listen carefully when an assessor provides feedback	☐	☐	☐	☐	☐
I am extremely encouraged by positive feedback from assessors	☐	☐	☐	☐	☐
I think that feedback provides clear direction on how to improve my performance	☐	☐	☐	☐	☐
Feedback from my assessors can be a valuable form of praise	☐	☐	☐	☐	☐
I pay careful attention to instructional feedback	☐	☐	☐	☐	☐
Feedback from my assessors motivates me to improve my performance	☐	☐	☐	☐	☐
Feedback from assessors is a waste of time	☐	☐	☐	☐	☐
I feel relieved when I receive positive feedback	☐	☐	☐	☐	☐

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Trainee feedback orientation scale

6. My response to feedback

Thinking about a typical WBA or SITA feedback situation, please indicate your level of agreement with the following statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Negative feedback hurts my feelings	☐	☐	☐	☐	☐
I feel threatened by negative feedback	☐	☐	☐	☐	☐
Negative feedback is intimidating	☐	☐	☐	☐	☐
My feelings are not easily hurt by negative feedback from an assessor	☐	☐	☐	☐	☐
It is difficult to "get over" negative feedback	☐	☐	☐	☐	☐
Negative feedback is embarrassing	☐	☐	☐	☐	☐
I tend to dwell on the negative feelings that result from negative feedback	☐	☐	☐	☐	☐
Negative feedback from an assessor increases the stress I feel about future performance	☐	☐	☐	☐	☐

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Trainee feedback orientation scale

7. Confidentiality of feedback

Thinking about a typical WBA or SITA feedback situation, please indicate your level of agreement with the following statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I don't like to receive negative feedback in front of other people	☐	☐	☐	☐	☐
I don't like others to hear the feedback that I am receiving	☐	☐	☐	☐	☐
I don't mind being singled out by feedback from an assessor	☐	☐	☐	☐	☐
I prefer to receive feedback from an assessor in private	☐	☐	☐	☐	☐
I like others to hear the feedback I am receiving from my assessor	☐	☐	☐	☐	☐

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Trainee feedback orientation scale

8. Retention of the feedback I receive

Thinking about a typical WBA or SITA feedback situation, please indicate your level of agreement with the following statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I can't remember what assessors want me to do when they provide feedback	☐	☐	☐	☐	☐
I tend to miss out on the details of what assessors want when they provide me with feedback	☐	☐	☐	☐	☐
I typically do not make note of the assessor's corrective comments	☐	☐	☐	☐	☐

The survey instrument (questions 5 to 8) is adapted from the Student Feedback Orientation Scale, King PE, Schrod P and Weisel JJ. 2009. "The instructional feedback orientation scale: conceptualizing and validating a new measures for assessing perceptions of instructional feedback." *Communication Education* 58: 235-261.

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Difficulties giving feedback

This section asks about negative feedback in a training situation.

Most people find that giving bad or negative news to someone can be a challenging task and this can effect the way that they deliver the news. This can apply to an assessor giving a trainee some negative feedback about their performance in a task.

Please think about a typical feedback session for a WBA or a SITA when you answer the following questions.

9. In your experience, do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Tell trainees only part of the problem?	☐	☐	☐	☐	☐
Don't go into much detail about the negatives?	☐	☐	☐	☐	☐
Leave out some of the negative information?	☐	☐	☐	☐	☐
Do not point out all the weaknesses?	☐	☐	☐	☐	☐
Avoid the trainee at work so they would not have to talk to him/her?	☐	☐	☐	☐	☐
Ask others to give any negative information to the trainee?	☐	☐	☐	☐	☐
Avoid it altogether?	☐	☐	☐	☐	☐
Put it at the bottom of their "to do" list?	☐	☐	☐	☐	☐
Steer clear of circumstances in which they'll have to give a trainee negative information	☐	☐	☐	☐	☐
Find ways to get out of sharing negative information with a trainee?	☐	☐	☐	☐	☐

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Difficulties giving feedback

10. In your experience, do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Put the negative news between two positives?	☐	☐	☐	☐	☐
Change the message they have to give a trainee to make it more positive?	☐	☐	☐	☐	☐
Try to make it sound nicer than it is?	☐	☐	☐	☐	☐
Change the message they have to give a trainee so it is not so negative?	☐	☐	☐	☐	☐
Try to make the information more positive than it really is?	☐	☐	☐	☐	☐
Try to make the trainee feel better by emphasising positive things?	☐	☐	☐	☐	☐

This survey instrument (questions 9 to 10) is adapted from Marler LE, McKee D, Cox S, Simmering M and Allen D. 2012 "Don't make me the bad guy: Organisational norms, self-monitoring and the MUM effect." *Journal of Managerial Issues* XXIV: 97-116.

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Difficulties giving feedback

11. In your experience, do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Change the feedback they are giving during a feedback session	☐	☐	☐	☐	☐
"Unload" the feedback quickly before the trainee has a chance to say anything	☐	☐	☐	☐	☐
Adopt a very formal manner that doesn't encourage dialogue	☐	☐	☐	☐	☐

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Your experiences with performance assessments

This part of the questionnaire asks about your experiences with performance assessments (WBA or in a SITA).

Please choose the response which most reflects your feelings about each statement.

12. Some trainees find that arranging to complete a work-based assessment can be challenging. Do you think that arranging to do a WBA (Derm-CEX, Pro-DA, CbD) is a challenge?

- ☐ No - I can easily get the ones I need completed
- ☐ Yes - it can be difficult to find an assessor to complete a WBA with
- ☐ Yes - it can be difficult to find an appropriate patient
- ☐ Yes - it can be difficult to find the time to complete these
- ☐ Other issues (please specify)

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Your experiences with performance assessments

13. When you complete a WBA, do you feel pressure at times to make sure that you are marked as competent on it?

☐ Yes

☐ No

14. If yes, please describe

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Your experiences with performance assessments

15. Does this mean that you may look for an assessor who is more likely to mark you as competent?

☐ Yes

☐ No

16. If yes, please describe

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Your experiences with performance assessments

Sometimes trainees directly communicate with their assessors about their assessment results.

In the past 12 months, have you ever:

17. Encouraged or asked an assessor to mark you as competent on a WBA?

☐ Yes

☐ No

18. If yes, please describe

19. Encouraged or asked an assessor to give you a certain rating on a SITA?

☐ Yes

☐ No

20. If yes, please describe

21. Encouraged or asked an assessor to change a score that you have been given on a WBA or a SITA?

☐ Yes

☐ No

22. If yes, please describe

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Your experiences with performance assessments

23. Are there any comments that you would like to make about the WBA and/or SITAs as they currently are?

Trainee questionnaire
Assessor grades and feedback: private and public judgements

Thank you for your participation

Thank you very much for taking the time to complete this survey. I would be grateful if you would avoid discussing your responses with others during the period of data collection. If this questionnaire has raised any issues for you which you wish to discuss further, please contact Prof Steve Trumble, 03 8344 8035, s.trumble@unimelb.edu.au or Dr Cate Scarff, scarffc@student.unimelb.edu.au.
If you would like to participate in the next phase of the study, which will involve individual

interviews on this topic, further information regarding this will be sent in the coming months.

With kind regards,

Cate

Appendix E Clinical performance assessment forms

SITA form



SUMMATIVE IN-TRAINING ASSESSMENT (SITA)

This form needs to be returned to the Supervisor of Training via the e-portfolio. The trainee does not see this form.

Date:

Trainee Name:								
Training Year:		Rotation:						
Training Hospital:								
Name of Supervisor:								
How many sessions/clinics do you supervise the trainee per fortnight?								
Have you completed any Derm-CEX's or Pro-DA's with this trainee?								
RATE THE FOLLOWING AREAS								
Red zone		Amber zone	Green zone		N/A			
1 Unsatisfactory; significantly below expectation for stage of training	2 Below expected standard for stage of training	3 Borderline; requiring some development	4 Meets expectation for stage of training	5 Beyond expected standard for stage	Insufficient opportunity to observe			
CLINICAL EXPERTISE			1	2	3	4	5	n/a
Possesses sound knowledge of current dermatological literature and evidence-based medicine								
Patient History and Exam – obtains pertinent information, thorough								
Clinical Decisions and Diagnoses – develops a provisional clinical diagnosis, selects appropriate tests and/or investigations, can interpret results.								
Treatment Management Plans – makes sound decisions about clinical management and checks risks if uncertain about intervention.								
Procedural Skills – creates safe procedural environment, obtains appropriate informed consent, conducts procedure in a technically competent manner, organises appropriate follow-up and after-care.								
PROFESSIONAL QUALITIES								
Communication			1	2	3	4	5	n/a
Communicates with Patients and their Families – communicates effectively with patients and their families, shows appropriate rapport, provides adequate information, recognizes patient's rights and respects patient confidentiality.								
Communicates with Colleagues – communicates clinical reasoning and intended clinical action, facilitates referral and clinical handover, co-operates and generally works well in a team.								
Written Communication – maintains legible records with adequate detail, accurate drug charting and writes timely and effective referral letters								

Quality and Safety	1	2	3	4	5	n/a
Optimises safe work practice - follows verification procedures, implements measures to prevent error, reports hazards and risks in the workplace.						
Prescribes and administers medication safety – accurately calculates drug doses and educates patients about medications.						
Promotes safe continuity of care for patients – ensures health care team are adequately informed of management plan.						
Cultural Competency	1	2	3	4	5	n/a
Uses resources available to communicate effectively with people from culturally and linguistically diverse backgrounds (interpreters, Aboriginal health workers etc)						
Treats patients as individuals and uses knowledge of patients' cultural and religious backgrounds in the management, treatment and care of the patient.						
Leadership and Management	1	2	3	4	5	n/a
Balances personal and professional priorities – shows coping skills, asks for help						
Reliability and dependability – carries out instruction, fulfils obligations, is punctual to work and tutorials, keeps up to date with work, notifies staff if expecting to be absent from duty						
Initiative and Enthusiasm – self-motivated, able to identify the needs of the job, follows up without being prompted, thinks and plans ahead, asks questions of supervisors						
Aware of limitations of expertise.						
Orders case specific investigations, uses healthcare resources appropriately.						
Health Advocacy	1	2	3	4	5	n/a
Encourages patients and communities to improve and maintain health, and prevent disease.						
Teaching and Learning	1	2	3	4	5	n/a
Takes responsibility for own learning – evidence of reading up on cases, attends hospital clinical meetings and tutorials, keen to discover new knowledge and integrate it into practice.						
Delivers meaningful education to patients, the wider community and colleagues.						
Meets teaching responsibility.						
Accepts feedback and guidance and shows commitment to improvement.						
Ethics	1	2	3	4	5	n/a
Shows understanding and judgment of ethical issues.						
Complies with hospital policies and legal legislation.						

In your opinion, what are the main strengths and areas most in need of further development of this trainee's performance (based on criteria you rated as 4)?
In your opinion, what are the main strengths and areas most in need of further development of this trainee's performance (based on criteria you rated as 1 – 3)?

Note: A trainee is deemed unsatisfactory if they:

- Receive '2' or below from two or more CSs on two or more criteria
- Have not completed the required number of Derm-CEXs or ProDAs commensurate with their year of training.

If a trainee is deemed unsatisfactory, a Performance Improvement Plan is required.

☐

By checking this box I acknowledge that I,

(Supervisor) have completed this SITA based on my observations of the Trainee. I also understand that it is my responsibility to upload this SITA into the e-Portfolio and that I may be contacted to confirm the content of this SITA

Please return this form to the SoT by uploading via the e-Portfolio.

The trainee does not see this form.

SITA summary form



THE AUSTRALASIAN COLLEGE
OF DERMATOLOGISTS

SUMMATIVE IN-TRAINING ASSESSMENT (SITA)

SUMMARY FORM

Date:

Trainee Name:			
Training Year:		Rotation:	
Training Hospital:			
Supervisor of Training:			

Red zone		Amber zone	Green zone	N/A				
1	2	3	4					
Unsatisfactory; significantly below expectation for stage of training	Below expected standard for stage of training	Borderline; requiring some development	Meets expectation for stage of training	Insufficient opportunity to observe				
Using the SITA forms please collate two CSs and your assessment by indicating the number of responses for each rating:				2	3	4	5	n/a
<i>Example of a collated response</i>						2	1	
CLINICAL EXPERTISE			1	2	3	4	5	n/a
Possesses sound knowledge of current dermatological literature and evidence-based medicine								
Patient History and Exam – obtains pertinent information, thorough								
Clinical Decisions and Diagnoses – develops a provisional clinical diagnosis, selects appropriate tests and/or investigations, can interpret results.								
Treatment Management Plans – makes sound decisions about clinical management and checks risks if uncertain about intervention.								
Procedural Skills – creates safe procedural environment, obtains appropriate informed consent, conducts procedure in a technically competent manner, organises appropriate follow-up and after-care.								
PROFESSIONAL QUALITIES								
Communication			1	2	3	4	5	n/a
Communicates with Patients and their Families – communicates effectively with patients and their families, shows appropriate rapport, provides adequate information, recognizes patient's rights and respects patient confidentiality.								
Communicates with Colleagues – communicates clinical reasoning and intended clinical action, facilitates referral and clinical handover, co-operates and generally works well in a team.								
Written Communication – maintains legible records with adequate detail, accurate drug charting and writes timely and effective referral letters								

Quality and Safety	1	2	3	4	5	n/a
Optimises safe work practice - follows verification procedures, implements measures to prevent error, reports hazards and risks in the workplace.						
Prescribes and administers medication safety – accurately calculates drug doses and educates patients about medications.						
Promotes safe continuity of care for patients – ensures health care team are adequately informed of management plan.						
Cultural Competency	1	2	3	4	5	n/a
Uses resources available to communicate effectively with people from culturally and linguistically diverse backgrounds (interpreters, Aboriginal health workers etc)						
Treats patients as individuals and uses knowledge of patients' cultural and religious backgrounds in the management, treatment and care of the patient.						
Leadership and Management	1	2	3	4	5	n/a
Balances personal and professional priorities – shows coping skills, asks for help when needed.						
Reliability and dependability – carries out instruction, fulfils obligations, is punctual to work and tutorials, keeps up to date with work, notifies staff if expecting to be absent from duty.						
Initiative and Enthusiasm – self-motivated, able to identify the needs of the job, follows up without being prompted, thinks and plans ahead, asks questions of supervisors.						
Aware of limitations of expertise.						
Orders case specific investigations, uses healthcare resources appropriately.						
Health Advocacy	1	2	3	4	5	n/a
Encourages patients and communities to improve and maintain health, and prevent disease.						
Teaching and Learning	1	2	3	4	5	n/a
Takes responsibility for own learning – evidence of reading up on cases, attends hospital clinical meetings and tutorials, keen to discover new knowledge and integrate it into practice.						
Delivers meaningful education to patients, the wider community and colleagues.						
Meets teaching responsibility.						
Accepts feedback and guidance and shows commitment to improvement.						
Ethics	1	2	3	4	5	n/a
Shows understanding and judgment of ethical issues.						
Complies with hospital policies and legal legislation.						

Summary of the main strengths of this trainee's performance (based on the above 4 ratings)
Summary of the trainee's areas most in need of further development (based on the above ratings 1 – 3)

Trainee portfolio reviewed	☐	☐	☐	Log of procedures verified	☐	☐	☐
RLP reviewed and discussed	☐	☐	☐	Next RLP completed at this meeting	☐	☐	☐

Total number of CbDs satisfactorily completed for this year to date:	
Total number of ProDAs satisfactorily completed for this year to date:	
Total number of Derm-CEXs satisfactorily completed for this year to date:	

The overall performance of the trainee in this rotation has been assessed as:	
SATISFACTORY	DEVELOPMENT REQUIRED → PIF Required – DoT must be notified

A trainee is deemed unsatisfactory if they:

- Receive '2' or below from two or more CSs on two or more criteria
- Have not completed the required number of Derm-CEXs or ProDAs commensurate with their year of training.

If a trainee is deemed unsatisfactory, a Performance Improvement Plan is required.

Trainee's Signature	
Supervisor's Signature	

Both signatures confirm that a face-to-face meeting took place to discuss this report.

Please return this form to the SoT by uploading via the e-Portfolio.

Dermatology clinical evaluation exercise (Derm-CEX)



THE AUSTRALASIAN COLLEGE
OF DERMATOLOGISTS

DERMATOLOGY CLINICAL EVALUATION EXERCISE (Derm-CEX)

Date:

Trainee Name:			
Training Year:		Rotation:	
Training Hospital:			
Supervisor of Training:			

Trainees and Supervisors should refer to the Curriculum, specifically the Fundamentals of Clinical Practice in Dermatology learning outcomes and the relevant Specialised Content Topic Areas.

Specialised Content Topic Area:			
Full Skin Examination:	☐ Yes	☐ No	Note: Trainees in their first year must include an appropriate, satisfactory full skin examination in at least one Derm-CEX.

CASE INFORMATION

Setting:	☐ Inpatient	☐ Outpatient	☐ Emergency	☐ Other:

ASSESSMENT GUIDE

Please rate the trainee's performance for each area. The trainee's performance should be assessed against the standard expected at the current stage of training.

Red zone	Amber zone	Amber+ zone	Green zone				
1	2	3	4				
Very limited in clinical knowledge and skills	Limited in clinical knowledge and skills	Moderate in clinical knowledge and skills	Competent in clinical knowledge and skills				
			1	2	3	4	N/A
Establishes a relationship with patient and/or carers (initial rapport, eye contact, acknowledges issues raised in referral letter, actively listens)							
Obtains and records a thorough and accurate history							
Optimises examination environment to ensure patient comfort and privacy, safe work practice and to minimize error.							
Prepares the patient for the examination by explaining the reason for, nature and extent of the examination and obtains informed consent							
Examines the skin of the patient in a respectful and sensitive manner with consideration of the individual patient							
Examines the problem presented by the patient and other relevant areas of the skin, mucosae, hair and nails							
Identifies incidental findings during the examination that are unrelated to the patient's complaint							

Conducts systemic examination, where relevant					
Performs specialised examination, where relevant, e.g. dermatoscope, Wood's light.					
Records examination findings that are clinically significant and uses diagrams, photos and other visual documentation as appropriate					
Outlines the immediate management plan with the patient and arranges follow-up, as required.					
Completes medical notes accurately, legibly and with follow up instructions if applicable					
Strengths (based on the above 4 ratings)					
Suggestions for Development (based on the above ratings 1 – 3)					
Trainee's reflective comments on their performance:					
Overall Clinical Competence					
☐ SATISFACTORY			☐ DEVELOPMENT REQUIRED		
Trainee's Signature					
Supervisor's Signature:					

Please submit this form by uploading it into your e-portfolio.

Procedural Dermatology Assessment (ProDA) example



THE AUSTRALASIAN COLLEGE
OF DERMATOLOGISTS

PROCEDURAL DERMATOLOGY ASSESSMENT (PRODA) - BIOPSY

Date:

Trainee Name:			
Training Year:		Rotation:	
Training Hospital:			
Supervisor of Training:			

Please rate the trainee's performance for each area. The trainee's performance should be assessed against the standard expected at the current stage of training

Red zone	Amber zone	Amber+ zone	Green zone					
1	2	3	4					
Trainee required detailed guidance and instruction from	Trainee required guidance and instruction from	Trainee was able to perform	Trainee was able to perform the step competently.					
Procedure Type:	☐ Eyelid	☐ Lip/Mucosa	☐ Nail	☐ Punch	☐ Scalp	☐ Shave		
					1	2	3	4
Selects the appropriate method of biopsy (punch, shave, saucerisation, excisional, incisional)								
For large lesions, performs multiple biopsies in order to avoid sampling error								
Identifies any contraindications to anaesthetic agents								
Identifies any contraindications to specific procedure planned								
Documents condition, communicates with patient								
Obtains informed consent								
Prepares patient and procedural environment								
Selects the appropriate site for biopsy (anatomical and within lesion)								
Accurately documents biopsy site, wherever possible with photography								
Marks out area to be biopsied, including appropriate margin if excisional biopsy								
Applies appropriate aseptic technique								
Administers appropriate anaesthesia								
Performs the selected biopsy method: - To the most appropriate size and depth for tissue diagnosis - Considers underlying and adjacent important anatomical structures - Handles tissue with care to avoid crushing artefact - Places tissue in appropriate transport medium								
Ensures that specimen is observed to be within correctly labelled specimen jar. Ideally 2 observers confirm this.								

Achieves haemostasis by electrosurgical, chemical, suture, haemostatic dressings or wound closure				
Ensures appropriate wound closure, if necessary, for punch, incision or excision biopsy				
Applies wound dressing				
Provides patient with verbal and/or written advice with regard to appropriate after care				
Ensures that there is a record in a biopsy log book and that ALL pathology reports are seen				
Organises follow up visit/phone call to communicate pathology, to remove sutures if applicable and to arrange any further treatment as indicated				
Demonstrates professionalism				
Strengths (based on the above 4 ratings)				
Suggestions for Development (based on the above ratings 1 – 3)				
Trainee's reflective comments on their performance:				
Overall Clinical Competence				
☐ SATISFACTORY		☐ DEVELOPMENT REQUIRED		
Trainee's Signature				
Supervisor's Signature				

Please submit this form by uploading it into your e-portfolio.

Case based discussion (CbD)



THE AUSTRALASIAN COLLEGE
OF DERMATOLOGISTS

CASE-BASED DISCUSSION (CbD)

Date:

Trainee Name:			
Training Year:		Rotation:	
Training Hospital:			
Name of Supervisor:			
CASE INFORMATION			
Setting:	☐ Inpatient	☐ Outpatient	☐ Emergency Other:
Specialised Content Topic Area/Long Case:			
Patient Age:		Patient Gender: ☐ Male	☐ Female
Case Complexity:	☐ Low	☐ Medium	☐ High

Trainees and Supervisors should refer to the curriculum, specifically the Fundamentals of Clinical Practice in Dermatology learning outcomes and the relevant specialised content topic areas.

Please rate the trainee's performance for each area. The trainee's performance should be assessed against the standard expected at the current stage of training	Red zone	Amber zone	Amber+ zone	Green zone
	1 Very limited in scope and detail	2 Limited in scope and detail	3 Moderate in scope and detail	4 Competent in scope and detail

To be satisfactory the trainee must achieve a 4 in all areas

	1	2	3	4
Record keeping	☐	☐	☐	☐
History taking	☐	☐	☐	☐
Development of differential diagnosis	☐	☐	☐	☐
Development and implementation of investigation plan	☐	☐	☐	☐
Interpretation and application of clinical findings	☐	☐	☐	☐
Establishment of working diagnosis	☐	☐	☐	☐
Development and implementation of treatment management	☐	☐	☐	☐

Strengths (based on the above rating 4)	
Suggestions for Development (based on the above ratings 1 – 3)	
Trainee's Reflective Comments	
Overall clinical competence	
☐ DEVELOPMENT REQUIRED	☐ SATISFACTORY (must be all 4's)
Trainee's Signature	
Supervisor's Signature:	

Please submit this form by uploading it into your e-portfolio.

Appendix F Interview schedule

Assessor grades and feedback: private thoughts and public judgements

How and why does reluctance to deliver negative assessment messages manifest in dermatology training in Australia?

Date:

Phone

In person

Plain language statement given

☐

Consent form completed

☐

Recording check

☐

Code name for participant:

.....

Demographics

- Number of years in practice
- Number of years assessing
- CS or SOT?
- State
- Gender

Background – In this interview, I am particularly interested in the clinical performance assessments - the SITA and the WBA (Derm-CEX, Pro-DA, CbD), so please answer the questions with these in mind.

- Can you tell me what it is like being an assessor?
- What are the positive things about being an assessor?
- Have you found any challenges with the role?
- Which formats do you personally conduct? How many per year (estimate)?

Assessor views on the performance assessment process

- What do you think of the clinical performance assessment formats?
- What are the aims of the WBA? What are they trying to achieve?
- What are the aims of the SITA? What are they trying to achieve?
- How well do you think they achieve these aims?

The process of delivering decisions

There has been a lot of research and study on various aspects of these assessments.

I want to talk in particular about the part of the assessment process when you **deliver** the judgment that you have formed on a trainee's level of performance.

- In general, how do you go about doing this for a WBA?
- When do you communicate it, where, how...
- Is there any difference depending on how well or poorly you judge that the trainee has done?
- In general, how do you go about doing this for a SITA?
- When do you communicate it, where, how...
 - If SOT - do you prepare for this in advance? If so, how?
 - Is this different depending on how well or poorly you judge that the trainee has done?
- Does delivering assessment judgements have any impact on you?
- Do you notice any physical or cognitive/emotional responses?
 - If so, how long do these responses last for?
- Do you think this process has any parallels to delivering clinical results to patients?

Communicating negative assessment news

I'd like to focus on the situation where you have formed an opinion that a trainee has performed poorly – either on a WBA or SITA – somewhere where you felt that their performance would reflect a low rating or some negative comments or feedback.

- Can you think of such a situation where a trainee performed poorly in a clinical performance assessment and you had formed a negative opinion on their performance?
- What was the situation? What happened – in what way did they perform poorly?

We know, both from the literature and from my previous research that sometimes in such situations, assessors do not end up delivering their thoughts – these particular negative ratings or feedback – on the trainee's level of performance.

- Have you ever been in a situation where you **didn't** deliver your thoughts – the feedback or ratings that you felt reflected the trainee's level of performance?
- Why did you not deliver the ratings/feedback?
- What did you do instead?

Scenarios

I'd like you to look at the scenarios below. They are purely fictitious – I made them up for these interviews. Could you please read them and then we'll discuss them?

(go through the two scenarios and questions/prompts at the bottom of each)

1.

Tara is half-way through her second year of training and asks her consultant, Ramesh to complete a WBA with her in clinic. When he agrees, she comments she's very grateful to him as she finds these very difficult to arrange and hasn't completed any so far this year.

Tara performs the WBA and Ramesh has concerns about her level of competence in several areas, thinking that she's really at the 'moderate' or '3' level for these.

In the end, he gives her all 4's so that she is marked as satisfactory, but documents all his concerns in the "suggestions for development" box.

Discussion points:

Have you ever found yourself in a similar situation? What did you do?

Do you think this sort of situation is common/occurs? If this is not a common or realistic situation, can you suggest a more realistic one?

Why might Ramesh mark her competent on the assessment when he didn't really think that she was?

Do you understand Ramesh's scoring?

Are there other things that he could have done?

Are some areas of WBAs more important than others? I.e. should only some be able to trigger a 'development required' outcome?

2.

Ben is a third-year trainee and Mio is one of the clinical supervisors completing a SITA form for him.

Mio thinks that Ben is a fairly good trainee – really nice, mostly reliable but sometimes a bit lacking in terms of clinical knowledge.

Mio does not think that Ben is below standard in any area on the SITA, though she does think he is borderline in several. In the end, she makes most of her ratings 'meets expectation' and gives a couple of '5's' also.

Discussion points:

Have you ever found yourself in a similar situation? What did you do?

Do you think some (other) assessors mark in this way?

Why would Mio do this?

Does it really matter if Mio gives Ben 4's and 5's instead of 3's?

Are there other things she could have done?

What if Mio did think that Ben was in the red zone in a few areas?

- What should she do?

- Should she contact the other CS to ask their opinion and how they are going to mark?

Does this happen? What does happen?

Finishing

- Are there other situations that you have experienced or heard about that are more common or likely to occur than the ones above?

- Do you think that assessors need help in delivering their messages to trainees?

- What could help assessors to deliver their messages?

- What has been tried so far and found not to work?

- What do you believe should be tried?

- What advice would you give to a clinician who is thinking about becoming an assessor?

- Is there anything else you would like to add/mention/discuss?

Thank you very much for your participation in the interview.

Once transcribed, I will forward it to you for checking.

Appendix G Summary of the SITA process



Summative-in-Training Assessment (SITA)

What is it?	The SITA process is designed to assess whether a Trainee is achieving satisfactory performance levels.
Areas of Assessment	Assess whether a Trainee's performance in the previous six months has met required standards relative to their stage of training. Determine whether a Trainee requires more supervision. Inform Trainees who are not meeting the required standards that their performance needs to improve. Provide a formal mechanism for documenting unsatisfactory performance, if applicable.
The Process	- In June and December each year, the Trainee provides a SITA Assessment form to each of the Supervising Assessors nominated on the RLP at the beginning of the rotation, and asks that the assessment be completed. These CS must have direct supervision of the Trainee. Usually the SoT, a CS and a Surgical Supervisor should complete the forms. - The SoT collates the information on the three SITA Assessment forms - The SoT determines whether the assessment is 'Satisfactory' or 'Unsatisfactory'. A Trainee's performance will be deemed unsatisfactory if the SITA Summary form details ratings of "Below expected standard" or "unsatisfactory" from two or more supervisors on two or more criteria. - Both the SoT and the Trainee sign the completed summary form - It is the Trainee's responsibility to ensure that all the forms are complete, that they have the required meeting with their SoT to discuss the SITA Summary form, and that the form is sent to the College office, and the DoT by the due date. For details regarding an Unsatisfactory SITA refer to the TPH, Chapter 4

Summative In-Training (SITA) Process

Supervisor of Training (SoT):

1. distributes SITA forms to Clinical Supervisors of their choosing
2. collates information from all three SITA forms onto the SITA Summary form
3. meets with Trainee to review Rotation Learning Plan and Trainee Portfolio, complete and discuss SITA Summary
4. sends copies of the three SITA forms to College **ONLY** if a PIF is invoked

SITA forms completed and returned to SoT

Trainee keeps SITA Summary copy in Trainee Portfolio, sends copies to DoT, SoT and original to College

UNSATISFACTORY SITA SUMMARY

Director of Training (DoT):

1. organises a meeting with the Trainee and SoT within two weeks to complete a Performance Improvement Plan (PIF). NTC is notified and the Trainee is on probation.
2. sends original PIF to College and copies to Trainee and SoT
3. organises follow-up meeting SoT and Trainee three months later to review PIF using the PIF Follow-up form.
4. sends original PIF Follow-up to College and copies to Trainee and SoT
5. if necessary repeats the PIF Follow-up process in another three months.

SATISFACTORY result from PIF/PIF Follow-up, the Trainee is off probation.

UNSATISFACTORY result from PIF/PIF Follow-up:

1. Referred to the National Education Committee
2. Training may not be accredited
3. Conditions must be met for future training time to be accredited.

Appendix H Supplementary data – questionnaire – assessors

Table H.1 Assessor self-report of number of assessments performed each year

Number of assessments completed per year	Derm-CEX n (%)	Pro-DA n (%)	CbD n (%)	SITA n (%)
1–2	18 32.7%	18 32.7%	12 21.8%	24 43.6%
3–4	13 23.6%	9 16.4%	8 14.6%	16 29.1%
5–6	4 7.3%	4 7.3%	7 12.7%	7 12.7%
7–10	1 1.8%	3 5.5%	2 3.6%	3 5.5%
11–15	0	0	0	2 3.6%
16–20	2 3.6%	0	1 1.8%	0
>20	2 3.6%	3 5.5%	4 7.3%	0
None	15 27.3%	18 32.7%	21 38.2%	3 5.5%

Table H.2: Mean PADS scores by demographics across all statements

Mean PADS		Difference	95 % CI	P-value
Junior dermatologist	Senior dermatologist			
2.65 (n = 7)	2.42 (n = 42)	.232	–.207, .622	.277
Junior assessor	Senior assessor			
2.51 (n = 15)	2.42 (n = 34)	.090	–.256, .461	.642
Female	Male			
2.41 (n = 20)	2.48 (n = 29)	–.073	–.474, .296	.742
SoT	Not an SoT			
2.29 (n = 14)	2.52 (n = 35)	–.231	–.646, .149	.266
Reported sharing bad assessment news within last 12 months	Not shared bad news			
2.28 (n = 23)	2.60 (n = 26)	–.320	–.691, .044	.095

Table H.3: Comparison of mean discomfort levels by assessor gender in relation to telling male and female trainees their performance needs to improve

	Mean PADS		Difference	95% CI	P-value
	Male assessor	Female assessor			
Telling a male trainee that his performance needs to improve	2.467 (n = 30)	2.565 (n = 23)	.099	-.524, .721	.752
Telling a female trainee that her performance needs to improve	2.828 (n = 29)	2.350 (n = 20)	-.478	-1.141, .186	.154

Table H.4: PADS scores by demographics in response to statement – “Warning a poorly performing trainee that unless performance improves, you will support a recommendation for his or her removal from the training program”.

PADS for statement		Difference	95 %CI	P-value
Junior dermatologist	Senior dermatologist			
4.14 (n = 7)	3.64 (n = 42)	.5	-.420, 1.433	.29
Junior assessor	Senior assessor	.412	-.169, 1.036	.176
4 (n = 15)	3.59 (n = 34)			
Female	Male	-.11	-.704, .547	.74
3.65 (n = 20)	3.76 (n = 29)			
SoT	Not an SoT	-.8	-1.476, -.135	.027
3.14 (n = 14)	3.94 (n = 35)			

Table H.5: PADS scores by demographics in response to statement – “Supporting a recommendation that a trainee be removed from the training program”

PADS for statement		Difference	95% CI	P-value
Junior dermatologist	Senior dermatologist			
4.00 (n = 4)	3.47 (n = 43)	.535	-.216, 1.293	.182
Junior assessor	Senior assessor	.4	-.234, 1.062	.247
3.81 (n = 16)	3.41 (n = 34)			
Female	Male	.054	-.748, .816	.893
3.57 (n = 21)	3.52 (n = 29)			
SoT	Not an SoT	-.155	-1.144, .834	.756
3.43 (n = 14)	3.58 (n = 36)			

Table H.6: PADS scores by demographics in response to statement – “A trainee’s accusing you of playing favourites in the rating of trainees”

PADS for statement		Difference	95 %CI	P-value
Junior dermatologist	Senior dermatologist			
3.143 (n = 7)	3.405 (n = 42)	-.262	-1.350, .656	.623
Junior assessor	Senior assessor			
3.270 (n = 15)	3.412 (n = 34)	-.145	-.915, .729	.723
Female	Male			
3.500 (n = 20)	3.276 (n = 29)	.224	-.572, 1.018	.589
SoT	Not an SoT			
3.286 (n = 14)	3.400 (n = 35)	-.114	-.9733, .732	.801

Table H.7: Comparison of mean number of MUM items agreed with in subscale by respondent seniority as dermatologist

MUM	Mean number of items agreed with in subscale		Mean difference (Junior dermatologist – Senior dermatologist)		
	Junior dermatologist	Senior dermatologist	Estimate	95% CI	P-value
MUM avoid	1.5 (n = 2)	.619 (n = 21)	.881	-.905, 2.682	.662
MUM distort	3.5 (n = 2)	1.35 (n = 20)	2.15	1.250, 3.042	.073
MUM other	0 (n = 2)	.263 (n = 19)	-.263	-.556, -.053	.056

Table H.8: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by respondent seniority as dermatologist

Percentage agreeing with having performed at least one behaviour			Difference in percentages (Junior – Senior)		
	Junior dermatologist	Senior dermatologist	Difference	CI	P-value
MUM avoid	50% (n = 2)	38.1% (n = 21)	11.9%	-.338, .560	.640
MUM distort	100% (n = 2)	80% (n = 20)	20%	-.468, .416	.662
MUM other	0% (n = 2)	21.1% (n = 19)	-21.1%	-.433, .459	.648

Table H.9: Comparison of mean number of MUM items agreed with in subscale by respondent seniority as assessor

MUM	Mean number of items agreed with in subscale		Mean difference (Junior assessor – Senior assessor)		
	Junior assessor	Senior assessor	Estimate	95% CI	P-value
MUM avoid	.667 (n = 6)	.706 (n = 17)	–.039	–.938, 1.200	.946
MUM distort	1.667 (n = 6)	1.5 (n = 16)	.167	–1.165, 1.646	.822
MUM other	0 (n = 6)	.333 (n = 15)	–.333	–.692, –.071	.055

Table H.10: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by respondent seniority as assessor

	Percentage agreeing with having performed at least one behaviour		Difference in percentages (Junior - Senior assessor)		
	Junior assessor	Senior assessor	Difference	95 % CI	P-value
MUM avoid	33.3% (n = 6)	41.2% (n = 17)	–7.8%	–.407, .337	.565
MUM distort	83.3% (n = 6)	81.3% (n = 16)	2.1%	–.394, .299	.708
MUM other	0% (n = 6)	26.7% (n = 15)	–26.7%	–.520, .154	.228

Table H.11: Comparison of mean number of MUM items agreed with in subscale by assessor gender

MUM	Mean number of items agreed with in subscale		Mean difference (Female - Male)		
	Female	Male	Estimate	95% CI	P-value
MUM avoid	.308 (n = 13)	1.2 (n = 10)	–.892	–1.930, .146	.110
MUM distort	1.000 (n = 13)	2.333 (n = 9)	–1.333	–2.699, 1.182	.066
MUM other	.154 (n = 13)	.375 (n = 8)	–.221	–.838, .235	.454

Table H.12: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by assessor gender

MUM	Percentage agreeing with having performed at least one behaviour		Difference in percentages (Female – Male)		
	Female	Male	Difference	CI	P-value
MUM avoid	23.1% (n = 13)	60% (n = 10)	–36.9%	–.645, .026	.086
MUM distort	69.2% (n = 13)	100% (n = 9)	–30.8%	–.576, .042	.098
MUM other	15.4% (n = 13)	25% (n = 8)	–9.6%	–.454, .226	.498

Table H.13: Comparison of mean number of MUM items agreed with in subscale by respondent SoT status

MUM	Mean number of items agreed with in subscale		Mean difference (SoT – not SoT)		
	SoT	Not an SoT	Estimate	95% CI	P-value
MUM avoid	.250 (n = 8)	.933 (n = 15)	–.683	–1.412, –.025	.079
MUM distort	.750 (n = 8)	2.000 (n = 14)	–1.25	–2.214, –.324	.021
MUM other	.125 (n = 8)	.308 (n = 13)	–.183	–.571, .250	.406

Table H.14: Comparison of percentage of respondents agreeing with at least one MUM behaviour in subscale by respondent SoT status

	Percentage agreeing with having performed at least one behaviour		Difference in percentages (SoT – not SoT)		
	SoT	Not an SoT	Difference	CI	P-value
MUM avoid	25% (n = 8)	46.7% (n = 15)	–.217	–.510, .188	.290
MUM distort	62.5% (n = 8)	92.9% (n = 14)	–30.4%	–.628, .037	.117
MUM other	12.5% (n = 8)	23.1% (n = 13)	–10.6%	–.396, .271	.502

Table H.15: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by seniority as a dermatologist

Percentage agreeing			Difference in percentages (Junior – Senior)		
	Junior dermatologist	Senior dermatologist	Estimate	CI	<i>P</i> -value
F2F	14.3% (n = 7)	12.5% (n = 40)	1.8%	–.162, .395	.643
GI	16.7% (n = 6)	17.9% (n = 39)	–1.3%	–.214, .394	.714

Table H.16: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by seniority as an assessor

Percentage agreeing			Difference in percentages (Junior – Senior)		
	Junior assessor	Senior assessor	Estimate	CI	<i>P</i> -value
F2F	6.3% (n = 16)	16.1% (n = 31)	–9.9%	–.272, .140	.322
GI	6.7% (n = 15)	23.3% (n = 30)	–16.7%	–.351, .092	.169

Table H.17: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by assessor gender

Percentage agreeing			Difference in percentages (Female – Male)		
	Female	Male	Estimate	CI	<i>P</i> -value
F2F	4.8% (n = 21)	19.2% (n = 26)	–14.5%	–.335, .064	.15
GI	15% (n = 20)	20% (n = 25)	–5%	–.265, .188	.487

Table H.18: Percentage self-report of failure to fail (F2F) or grade inflation (GI) of a trainee by SoT status

Percentage agreeing			Difference in percentages (SoT – not SoT)		
	SoT	Not an SoT	Estimate	CI	<i>P</i> -value
F2F	0% (n = 14)	18.2% (n = 33)	–18.2%	–.344, .054	.103
GI	21.4% (n = 14)	16.1% (n = 31)	5.3%	–.163, .330	.482

Table H.19: Comparison of mean PADS score with self-report of approach in relation to a clinical performance assessment or recently having to share negative assessment information with a trainee

Mean PADS score		Mean difference (First column – second column)		
Direct request made	No direct request made	Estimate	95% CI	P-value
2.630 (n = 5)	2.439 (n = 42)	–.191	–.738, .308	.496
Had to share negative information	Not had to share negative information			
2.282 (n = 23)	2.602 (n = 26)	–.320	–.691, .044	.106

Table H.20: Comparison of number of MUM behaviours self-reported between those assessors who reported by directly approached by a trainee in relation to an assessment outcome with those who did not

	Number of MUM behaviours self-reported		Mean difference (First column – second column)		
	Direct request made	No direct request made	Estimate	95% CI	P-value
MUM avoid subscale	2.667 (n = 3)	.389 (n = 18)	–2.278	–3.700, –.579	.118
MUM SC subscale	2.667 (n = 3)	1.389 (n = 18)	–1.278	–2.778, .550	.281
MUM other subscale	.333 (n = 3)	.222 (n = 18)	–.111	–.882, .421	.779

Table H.21: Comparison of percent of assessor self-reported engagement in at least one MUM behaviour between those assessors who reported by directly approached by a trainee in relation to an assessment outcome with those who did not

	Self-reported at least one MUM behaviours		Mean difference (First column – second column)		
	Direct request made	No direct request made	Estimate	95% CI	P-value
MUM avoid subscale	100%	27.8%	.722	.115, .875	.042
MUM SC subscale	100%	77.8%	.222	–.355, .452	.511
MUM other subscale	33.3%	16.7%	.167	–.187, .638	.489

Appendix I Supplementary data – questionnaire – trainees

Table I.1: Trainee responses to MUM avoidance subscale items

Do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:	Agree n (%)	Disagree/neutral n (%)	Respondent number n
Tell trainees only part of the problem?	15 (48.4)	16 (51.6)	31
Don't go into much detail about the negatives?	17 (54.8)	14 (45.2)	31
Leave out some of the negative information?	17 (54.8)	14 (45.2)	31
Don't point out all the weaknesses?	18 (58.1)	13 (41.9)	31
Avoid the trainee at work so they would not have to talk to him/her?	4 (12.9)	27 (87.1)	31
Ask others to give any negative information to the trainee?	7 (22.6)	24 (77.4)	31
Avoid it altogether?	5 (16.1)	26 (83.9)	31
Put it at the bottom of their "to do" list?	8 (25.8)	23 (74.2)	31
Steer clear of circumstances in which they'll have to give a trainee negative information?	6 (20.0)	24 (80.0)	30
Find ways to get out of sharing negative information with a trainee?	5 (16.1)	26 (83.9)	31

Table I.2: Trainee responses to MUM sugarcoating subscale items

Do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:	Agree n (%)	Disagree/neutral n (%)	Respondent number n
Put the negative news between two positives?	13 (41.9)	18 (58.1)	31
Change the message they have to give a trainee to make it more positive?	13 (41.9)	18 (58.1)	31
Try to make it sounds nicer than it is?	15 (48.4)	16 (51.6)	31
Change the message they have to give a trainee so it is not so negative?	13 (41.9)	18 (58.1)	31
Try to make information more positive than it really is?	12 (38.7)	19 (61.3)	31
Try to make the trainee feel better by emphasising positive things?	16 (51.6)	15 (48.4)	31

Table I.3: Trainee responses to MUM other subscale items

Do you think that when assessors give negative feedback to trainees about their performance in a WBA or on a SITA they:	Agree n (%)	Disagree/neutral n (%)	Respondent number n
Change the feedback are giving during feedback session?	3 (10.0)	27 (90.0)	30
“Unload” the feedback quickly before the trainee has a chance to say anything?	4 (13.3)	26 (86.7)	30
Adopt a very formal manner that doesn’t encourage dialogue?	8 (26.7)	22 (73.3)	30

Table I.4: Mean number of items respondents agreed to in MUM subscales by trainee seniority

MUM subscale (n of items in subscale)	Mean number of items in subscale agreed with		Difference (Junior – Senior)		
	Junior (n = 20/19)	Senior (n = 11)	Estimate	95% CI	P-value
Avoid (10)	2.65 (n = 20)	4.45	–1.80	–4.05, .44	.110
Distort (6)	2.45 (n = 20)	3.00	–.55	–2.46, 1.36	.557
Other (3)	0.42 (n = 19)	0.64	–.36	–.98, .55	.562

Table I.5: Percentage of respondents agreeing to at least one item in MUM subscales by trainee seniority

MUM subscale (n of items in subscale)	Percentage agreeing with at least one item in subscale		Difference in proportions (percentages) (Junior – Senior)		
	Junior (n = 20/19)	Senior (n = 11)	Estimate	95% CI	P-value
Avoid (10)	60% (n = 20)	90.9%	–30.9%	–.535, .030	.106
Distort (6)	60% (n = 20)	81.8%	–21.8%	–.468, .128	.262
Other (3)	26.3% (n = 19)	36.4%	–10.0%	–.418, .208	.687

Table I.6: Mean number of items respondents agreed to in MUM subscales by trainee gender

MUM subscale (n items in subscale)	Mean score		Mean difference (Female – Male)		
	Female (n = 19)	Male (n =)	Estimate	95% CI	P-value
Avoid (10)	3.58	2.80 (n = 10)	.779	–1.739, 3.297	.526
Distort (6)	2.95	1.90 (n = 10)	1.047	–.805, 2.900	.255
Other (3)	.37	.89 (n = 9)	–.520	–1.385, .344	.216

Table I.7: Percentage of respondents agreeing to at least one item in MUM subscales by trainee gender

MUM subscale (n items in subscale)	Percentage agreeing with at least one item in subscale		% difference (Female – Male)		
	Female (n = 19)	Male (n = 10/9)	Estimate	95% CI	P-value
Avoid (10)	73.7%	60.0% (n = 10)	13.7%	–.186, .459	.479
Distort (6)	68.4%	60.0% (n = 10)	8.4%	–.238, .414	.540
Other (3)	21.1%	55.6% (n = 9)	–34.5%	–.630, .020	.112

Table I.8: Trainee responses to statements in Usefulness of feedback subscale

Statements – Usefulness subscale	Agree n (%)	Disagree/neutral n (%)	Response count n
I think feedback from my assessors is vitally important in improving my performance	25 (78.1)	7 (21.9)	32
I will usually reflect on assessor's feedback	28 (87.5)	4 (12.5)	32
I listen carefully when an assessor provides feedback	31 (96.9)	1 (3.1)	32
I'm extremely encouraged by positive feedback from assessors	28 (87.5)	4 (12.5)	32
I think that feedback provides clear direction on how to improve my performance	22 (68.8)	10 (31.2)	32
Feedback from my assessors can be a valuable form of praise	31 (96.9)	1 (3.1)	32
I pay careful attention to instructional feedback	29 (93.5)	2 (6.5)	31
Feedback from my assessors motivates me to improve my performance	29 (93.5)	2 (6.5)	31
Feedback from assessors is a waste of time	1 (3.2)	30 (96.8)	31
I feel relieved when I receive positive feedback	27 (87.1)	4 (12.9)	31

Table I.9: Trainee responses to statements in Response/Sensitivity subscale

Statements – Response/Sensitivity subscale	Agree n (%)	Disagree/neutral n (%)	Response count n
Negative feedback hurts my feelings	14 (45.2)	17 (54.8)	31
I feel threatened by negative feedback	4 (12.9)	27 (87.1)	31
Negative feedback is intimidating	9 (29.0)	22 (71.0)	31
My feelings are not easily hurt by negative feedback from an assessor	7 (22.6)	24 (77.4)	31
It is difficult to “get over” negative feedback	9 (29.0)	22 (71.0)	31
Negative feedback is embarrassing	9 (29.0)	22 (71.0)	31
I tend to dwell on the negative feelings that result from negative feedback	9 (29.0)	22 (71.0)	31
Negative feedback from an assessor increases the stress I feel about future performance	17	14	31

Table I.10: Trainee responses to statements in Confidentiality subscale

Statements – Confidentiality subscale	Agree n (%)	Disagree/neutral n (%)	Response count n
I don't like to receive negative feedback in front of other people	25 (80.6)	6 (19.4)	31
I don't like others to hear the feedback that I am receiving	23 (74.2)	8 (25.8)	31
I don't mind being singled out by feedback from an assessor	14 (45.2)	17 (54.8)	31
I prefer to receive feedback from an assessor in private	29 (93.5)	2 (6.5)	31
I like others to hear the feedback I am receiving from my assessor	0 (0)	31 (100)	31

Table I.11: Trainee responses to statements in Retention of feedback subscale

Statements – Retention subscale	Agree n (%)	Disagree/neutral n (%)	Response count n
I can't remember what assessors want me to do when they provide feedback	5 (16.1)	26 (83.9)	31
I tend to miss out on the details of what assessors want when they provide me with feedback	3 (9.7)	28 (90.3)	31
I typically do not make note of the assessor's corrective comments	3 (9.7)	28 (90.3)	31

Table 1.12: Comparison of mean number of items in subscale agreed with by level of training

Mean number of items agreed with in subscale			Mean difference (Junior – Senior)		
IFOS subscale (n items in subscale)	Junior (n = 20)	Senior (n = 11)	Estimate	95% CI	P-value
Usefulness (10)	9.05	7.818	1.23	–.50, 2.96	.154
Response/Sensitivity (8)	2.70	2.64	.064	–1.542, 1.669	.936
Confidentiality (5)	3.5	4.0	–.50	–1.490, .490	.309
Retention (3)	2.45	1.818	.632	–.170, 1.434	.115

Table 1.13: Comparison of percentage of respondents agreeing with at least one item in subscale by level of training

Percentage agreeing with at least one item in subscale			Difference in proportions (percentages) (Junior – Senior)		
IFOS subscale	Junior (n = 20)	Senior (n = 11)	Estimate	95% CI	P-value
Usefulness (10)	95%	100%	–5%	–.236, .212	1.000
Response/Sensitivity (8)	85%	81.8%	3.2%	–.216, .343	1.000
Confidentiality (5)	90%	100%	–10%	–.301, .169	.527
Retention (3)	95%	90.9%	4.1%	–.160, .330	1.000

Table I.14: Comparison of mean number of items in subscale agreed with by trainee gender

IFOS subscale (n items in subscale)	Mean number of items agreed with in subscale		Mean difference (Female – Male)		
	Female (n = 19)	Male (n = 10)	Estimate	95% CI	P-value
Usefulness (10)	8.63	8.70	-.068	-1.793, 1.656	.935
Response/Sensitivity (8)	2.05	3.4	-1.347	-3.09, .396	.122
Confidentiality (5)	3.58	3.80	-.22	-1.366, .923	.692
Retention (3)	2.16	2.20	-.042	-.78, .700	.907

Table I.15: Comparison of percentage of respondents agreeing with at least one item in subscale by trainee gender

IFOS subscale	Percentage agreeing with at least one item in subscale		Difference in proportions (percentages) (Female – Male)		
	Female (n = 19)	Male (n = 10)	Estimate	95% CI	P-value
Usefulness (10)	94.7%	100%	-5.3%	-.246, .210	1.000
Response/Sensitivity (8)	78.9%	90%	-11.1%	-.348, .218	.743
Confidentiality (5)	89.5%	100%	-10.5%	-.314, .182	.591
Retention (3)	89.5%	100%	-10.5%	-.314, .182	.591

Table I.16: Percentage of respondents agreeing to statements about WBA by level of training

	Percentage agreeing		Difference in percentages (Junior – Senior)		
	Junior (n = 19)	Senior (n = 11)	Estimate	95% CI	P-value
WBA a challenge to arrange	78.9%	63.6%	15.3%	-.154, .462	.417
Feel pressure to be marked competent	89.5%	54.5%	34.9%	.030, .625	.068
Look for assessor more likely to mark competent	42.1%	36.4%	5.7%	-.283, .360	1.00

Table I.17: Percentage of respondents agreeing to statements about WBA by trainee gender

	Percentage agreeing		Difference in percentages (Female – Male)		
	Female (n = 19)	Male (n = 9)	Estimate	95% CI	P-value
WBA a challenge to arrange	73.7%	66.7%	7.0%	-.239, .415	1.00
Feel pressure to be marked competent	78.9%	66.7%	12.3%	-.185, .460	.794
Look for assessor more likely to mark competent	31.6%	44.4%	-12.9%	-.460, .211	.228

Table I.18: Percentage of trainee respondents agreeing to statements by IFOS and MUM responses

	Percentage agreeing		Difference in percentages		
	yes (n =)	no (n =)	Estimate	95% CI	P-value
Agree find it challenging to arrange WBA	IFOS Use agree with at least 1 statement				
	95.5% (n = 22)	100% (n = 8)	-4.5%	-.218, .281	.733
	IFOS Resp agree with at least 1 statement				
	86.4% (n = 22)	75% (n = 8)	11.4%	-.152, .466	.405
	MUM avoid agree with at least 1 statement				
	77.3% (n = 22)	62.5% (n = 8)	14.8%	-.168, .491	.355
	MUM distort agree with at least 1 statement				
	70% (n = 20)	80% (n = 10)	-10%	-.362, .246	.452
	MUM other agree with at least 1 statement				
	77.8% (n = 9)	71.4% (n = 21)	6.3%	-.294, .33	.547
Agree feel pressure to be marked competent	IFOS Use agree with at least 1 statement				
	95.7% (n = 23)	100% (n = 7)	-4.3%	-.21, .313	.767
	IFOS Resp agree with at least 1 statement				
	87% (n = 23)	71.4% (n = 7)	15.5%	-.124, .521	.329
	MUM avoid agree with at least 1 statement				
	72.7% (n = 22)	87.5% (n = 8)	-14.8%	-.380, .226	.377
	MUM distort agree with at least 1 statement				
	75% (n = 20)	80% (n = 10)	-5%	-.311, .289	.571
	MUM other agree with at least 1 statement				
	66.7% (N = 9)	81% (n = 21)	-14.3%	-.475, .156	.343
Agree to look for assessor who will be likely to pass trainee	IFOS Use agree with at least 1 statement				
	91.7% (n = 12)	100 (n = 18)	-8.3%	-.354, .105	.4
	IFOS Resp agree with at least 1 statement				
	91.7% (n = 12)	77.8% (n = 18)	13.9%	-.162, .370	.318
	MUM avoid agree with at least 1 statement				
	36.4% (n = 22)	50% (n = 8)	-13.6%	-.466, .216	.396
	MUM distort agree with at least 1 statement				
	45% (n = 20)	30% (n = 10)	15%	-.209, .433	.35
	MUM other agree with at least 1 statement				
	33.3% (n = 9)	42.9% (n = 21)	-9.5%	-.391, .267	.472