
Article:
Crossley, J.G.M., Groves, J., Croke, D. et al. (1 more author) (2019) Examiner training: A study of examiners making sense of norm-referenced feedback. Medical Teacher, 41 (7). pp. 787-794. ISSN 0142-159X

https://doi.org/10.1080/0142159X.2019.1579902

Examiner training: a study of examiners making sense of norm-referenced feedback

Running head: How examiners experience feedback

Authors:

James GM Crossley MA, MRCPCH, MEd, DM

Professor Crossley is Professor of Medical Education at the University of Sheffield and Consultant Paediatrician at Chesterfield Royal Hospital.

Jeremy Groves BSc, FRCA

Dr Groves is a Consultant in Anaesthesia and Intensive Care at Chesterfield Royal Hospital and an Assessor for the Intercollegiate Committee for Basic Surgical Examinations (ICBSE).

David Croke MA, PhD, FRCPath

Professor Croke is Director of Quality Enhancement at the Royal College of Surgeons in Ireland and Chair of the Internal Quality Assurance Committee of the ICBSE in the UK & Ireland.

Peter A Brennan MD, FRCS, FRCSI, FDSRCS

Professor Brennan is Consultant Maxillofacial Surgeon and Honorary Professor of Surgery, Queen Alexandra Hospital, Portsmouth. He is also chairman of the ICBSE in the UK and Ireland.

Corresponding author:

Professor James Crossley, University of Sheffield, The Medical School, Beech Hill Road, Sheffield S10 2RX

Email: j.crossley@sheffield.ac.uk; Tel: 0114 222 5543

Abstract

Purpose

Examiner training has an inconsistent impact on subsequent performance. To understand this variation, we explored how examiners think about changing the way they assess.

Method

We provided comparative data to seventeen experienced examiners about their assessments, captured their sense-making processes using a modified think-aloud protocol, and identified patterns by inductive thematic analysis.

Results

We observed five sense-making processes: (1) *testing personal relevance* (2) *interpretation* (3) *attribution* (4) *considering the need for change*, and (5) *considering the nature of change*.

Three observed meta-themes describe the *manner* of examiners' thinking: *Guarded curiosity* - where examiners expressed curiosity over how their judgements compared with others', but they also expressed guardedness about the relevance of the comparisons. *Dysfunctional assimilation* – where examiners' interpretation and attribution exhibited cognitive anchoring, personalisation, and affective bias. *Moderated conservatism* - where examiners expressed openness to change, but also loyalty to their judgement-framing values and aphorisms.

Conclusions

Our examiners engaged in complex processes as they considered changing their assessments. The 'stabilising' mechanisms some used resembled learners assimilating educational feedback. If these are typical examiner responses, they may well explain the variable impact of examiner training, and have significant implications for the pursuit of meaningful and defensible judgement-based assessment.

Introduction

Assessor differences in judgement-based assessment

The term judgement-based assessment appears increasingly often in the education literature (Griffin and Robertson 2014). This is because of a growing acknowledgement that tacit aspects of learning and performance require assessors to make subjective judgements. In other words, they cannot be 'objectified'.

Evaluations of judgement-based assessment formats show assessor differences. Often these are so substantial that very large samples of assessors are required before their combined judgements produce reliable assessment metrics. This poses a utility problem affecting, for example, assessment in the workplace (WBA).

From a theoretical perspective Gingerich et. al. (2014) describe three distinct ways of viewing assessors' differences: trainable (applying criteria wrongly), fallible (fundamentally poor at judgement), or meaningfully different (legitimate experienced-based experts).

The 'trainable' perspective

Taking the first perspective, many educators call for more assessor training suggesting that it will improve the reliability of judgement-based assessment (Ringstead et al. 2003). However, empirical studies of assessor training demonstrate a minimal or insignificant impact on assessor reliability even when theoretically informed training interventions have been used (Holmboe et al. 2004; Cook et. al. 2008).

This even holds true for feedback-driven training. In advanced secondary education in the UK, only a sub-group of examiners becomes more consistent at scoring following such feedback (Greatorex and Bell 2008). In a spoken English language assessment, examiners valued the feedback but there was no impact on their behaviour (Knoch 2011). In a written English language assessment, individualised feedback reduced inter-rater variation but also

reduced discrimination (spread) making the overall reliability worse (Elder et al. 2005). Recently, in a medical education context, feedback had exactly the opposite effect. It increased the range of ratings that peer assessors gave within PBL groups (spread) but also increased inter-rater variation such that it had minimal overall impact on reliability (Roberts et al. 2017).

The 'meaningfully different' perspective

To better understand the variable impact of assessor training we need to know more about how assessors think about their judgements and about making changes to them. In particular, since a growing literature shows that assessors' internal heuristics are central to their judgements, we need to understand how assessors compare their own heuristics with the alternatives offered by training or feedback.

Kogan et al. (2015) interviewed 45 experienced assessors about their reactions to an assessor training intervention in a US postgraduate medical context. They reported important barriers to change including 'a preference for holistic assessment over frameworks', and 'difficulty in changing one's approach to assessment'.

Our research question

The evidence suggests that the 'trainable' perspective doesn't provide a comprehensive understanding of, or solution to, examiner differences. Taking the 'meaningfully different' perspective we sought to understand how assessors think about their judgements, how they compare their own heuristics or standards with alternatives, and how they think about changing the way they assess. This kind of understanding has the potential advance meaningful and defensible assessment in two main ways. We may learn better ways to address 'trainable' assessor differences, and we may better understand the meaningful

differences between assessors such that we don't treat them as error in our assessment designs.

Method

We chose a phenomenological approach, using a modified think-aloud protocol to observe the sense-making processes of particular experienced examiners as they received comparative feedback about their judgement-based assessments and were invited to consider the implications.

Study context

The study was conducted in the context of the Intercollegiate Membership of the Royal Colleges of Surgeons (MRCS) OSCE. This qualification is a requirement for moving from core to advanced specialist training as a surgeon in the UK and Ireland. A description is available by following the links from the Intercollegiate Committee for Basic Surgical Examinations (ICBSE) website www.intercollegiatemrcsexams.org.uk/new.

In 2014 the ICBSE decided to implement universal feedback for MRCS OSCE examiners following each examination round to provide them with comparative information about their performance as assessors. This created the opportunity for the present study.

Ethical approval

The ICBSE has professional, lay and trainee representation and represents the Surgical Royal Colleges as the institutional review board for innovations and research in the context of the intercollegiate examinations. The committee considered the study to be exempt from the need for ethical approval.

Feedback to examiners

Data were fed back to examiners in written form only. Each individual feedback 'package' comprised the following four elements:

- an introductory paragraph explaining that the purpose of feedback was to promote reflection and thus contribute to formative development
- an external examiner's comments on the index examiner's station (default of 'satisfactory' unless change was required), and generic candidate comments about the examination as a whole
- charts comparing scores given by the recipient examiner with scores given by his or her peers along with explanations (see figures 1-3 and legends)
- open space to consider the meaning and implications of the feedback

As the figures and legends show, the comparative data is challenging to interpret and has inherent limitations.

[figures 1-2 near here]

Sampling and data collection

The first cycle of interviews was conducted the first time examiners ever received feedback in 2014. Respondents were identified and contacted by the ICBSE manager. Since we were interested in individual experiences, the first round of sampling was driven by investigator locality and not by theoretical or representative considerations. Of eight potential examiners in the English Midlands, five accepted the invitation and three were interviewed. Of a potential four examiners at the Irish College, two accepted the invitation and one was interviewed.

We asked the examiner to 'speak your thoughts out loud as you read and consider the feedback'. We didn't interrupt unless the examiner went quiet – in which case we asked them to 'see if you can speak your thoughts out loud'. When the examiner had finished, we asked him or her to explain anything that we hadn't understood, or about which we wished to know more.

Noting that many examiners struggled to comprehend the feedback the first time they saw it, we decided on theoretical grounds to sample a second cycle of examiners familiar with receiving feedback. The second cycle of interviews took place after two years when all participants had received at least two rounds of feedback. A similar locality-based sampling procedure yielded three interviews from four volunteers of a potential eight in the English Midlands, and three interviews from four volunteers of a potential sixteen in Ireland. A further seven examiners were interviewed during rounds of the Examination in Sheffield, Glasgow and Dublin.

The interview procedure was similar on the two occasions but with an additional step during the second cycle exploring recall and impact. The steps are described below:

- introduction: a brief explanation of the purpose of the interview
- questions about recall and impact of past feedback (second cycle only)
- think aloud exercise on receipt of most recent feedback
- follow up questions

The sample size was convenience driven. However, we did not find any new primary themes arising after the first eleven interviews, suggesting that we may have reached saturation.

Analysis

JC transcribed the interviews and broke them into segments dealing with discrete ideas. He then openly coded and named those ideas as primary themes. The primary themes were then clustered into sense-making processes. Finally, the processes were grouped into meta-themes based on our perceptions of the *manner* in which respondents thought about changing the way they assess. Primary and meta-theme identification was informed by notes made during the interviews. JC, JG and DC discussed all three levels of coding until consensus was reached. We did not undertake member checking.

Results

The sampling strategy resulted in seventeen audio-recorded interviews lasting an average of 26 minutes and 23 seconds. Twelve of the interviews were accompanied by contemporaneous hand-written notes from a second interviewer. Where quotes from these interviews are used below, respondents are identified using an 8-digit code denoting: 6-digit date/interview number that day/gender.

Coding identified eighteen primary themes clustering into five sense-making processes. Table 1 summarises the initial themes and processes, and provides brief illustrative segments or observations.

[table 1 near here]

Our primary interest however is the *manner* in which our respondents thought about changing the way they assess. We reached consensus on three meta-themes characterising manner; they were *guarded curiosity*, *dysfunctional assimilation*, and *moderated conservatism*. Each is discussed in more detail below and the themes, processes and meta-themes are set out systematically in figure 4.

[figure 4 near here]

Guarded curiosity

Curiosity describes some examiners' genuine appetite for feedback (theme 1). Several examiners linked this to the themes of peer comparison (theme 6) and personalisation (theme 10):

> *"The league table attraction is actually very, very powerful. You're looking to see 'how am I doing compared to everybody else'."*
>
> *(05/10/16/3/M)*

Guardedness describes the way examiners used the relevance, meaning and attribution stages of the process as a 'threshold test' regulating whether or not they would consider change. Relevance, meaning and attribution co-functioned in a circular manner: if the examiner wasn't convinced about applicability (theme 2) or believability (theme 3) then he distanced himself from the feedback (theme 5) and made little effort to discern meaning or attribution. For example, the examiner below thought he had received the wrong feedback (believability), but rather than following through on raising this concern with the ICBSE he simply distanced himself from his feedback.

"I contacted the college because there are things which are not clear. And I think they did give me some names – but I keep forgetting." (28/09/16/2/M)

Similarly, if the interpretation, significance, or alignment was unclear (themes 7-9) or the attribution was external (theme 11), then the examiner was likely to stall in vagueness (theme 4). For example, the examiner below struggled with interpretation. Rather than investing additional effort, or seeking help, she settled for a vague and provisional interpretation.

"I mean, I kind of took one look and thought 'ooh, this is above my pay station'." (28/09/16/1/F)

Curiosity and guardedness are in direct tension and this was apparent in the interviews:

"It hits a nerve – particularly with surgeons because we're all psychopaths anyway: 'am I an outlier'; 'am I the best', 'am I not the best'; 'if not, why not'." (05/10/16/3/M)

Dysfunctional assimilation

Several examiners found interpretation (theme 7) to be a complex process that involved sifting through comments and graphical data, understanding the composition of data and the variables and axes of charts, and judging the significance of any variance seen (theme 8):

> *"What threw me was trying to decipher whether my interpretation of these two stations* [the charts] *was correct or was wrong." (05/10/16/2/F)*

Most examiners attempted to integrate diverse data into some kind of singularity (theme 9), but many failed to resolve apparent dissonances and chose one piece of data over another. For example, the examiner below could not reconcile his (erroneous) perception that one chart portrayed him as stringent but another chart as lenient. Rather than reconciling the two, he settled for one of the two options.

> *"I'd expect it to be up this end of things … I put more credence and faith in this one to be honest." (05/10/16/3/M)*

Cognitive anchoring (theme 12) was common, with preconceptions, early observations, or more outstanding observations causing respondents to ignore or misinterpret subsequent data. Affective responses seemed central to these interpretive derailments – as the examiner below indicates.

> *"People get very – obviously – defensive and stuff like that, and can get upset." (14/10/16/1/M)*

By far the commonest reason for strong affective reactions was personalisation (theme 10):

> *"The first thing you look at you think 'arhh, look at me – I'm the outlier'. And again, you think 'hang on, two dots out of the three are outside of the boxes, this confirms that I'm an outlier'. And we all want*

to be good, standard examiners. You don't want to be the softy or the hawk." (12/02/14/1/M)

Moderated conservatism

Conservatism describes examiners' personal investment in their existing means of judgement. Many described their own judging patterns using terms of identity (theme 10) such as 'hawk', 'dove', 'outlier', 'extremist' and 'rogue agent'. They embraced a range of values (theme 17) including 'fairness', 'decisiveness', 'consistency', 'patient-centredness', 'equality', and 'being true to one's initial impression'. Conservatism was often expressed as resistance to change (theme 16):

"Changes? I don't think so. I can't remember but I would have marked them on the basis of what I thought was the right mark for that person at that time." (10/02/14/1/M)

Conservatism also meant that change didn't involve replacing existing means of judgement; rather the new information was mediated through or added to existing means of judgement:

"What it now says to me is you've got to earn your pass with me. I haven't become ferocious or nasty but a little more honest as it were about how I interpret or allocate marks." (05/10/16/3/M)

Moderation describes some examiners' deliberate but 'fragile' openness to information that might challenge their existing means of judgement. Whilst willing to respond to credible information (theme 13), they expressed unease about the possibility of change (theme 15):

"If there were, say, areas of concern or areas to be developed it would help me – certainly help me. It should. But it felt worse than my exit exam!" (05/10/16/1/M)

They also expressed relief if change wasn't required (theme 14).

"this made me feel better than previous charts, because it's more like I think I am" (14/10/16/2/M)

Again, conservatism and moderation are in direct tension. Through this tension, examiners modulated their change threshold.

Discussion

Main findings

The examiners we studied assimilated comparative data about their assessments through critical active engagement rather than passive receipt and change. Indeed, in many cases, there was a guarded tone to their assimilation. This is despite the fact that almost all the examiners welcomed the feedback in principle. For some, questions of relevance, interpretation and attribution seemed to function as a 'threshold test'; a certain threshold had to be reached before they were willing to make the more substantial investment of considering changing the way they assess.

Several examiners finding insufficient evidence to prompt change expressed reassurance. Several who didn't draw clear conclusions expressed anxiety. Some examiners used 'defence mechanisms' such as vagueness or distancing rather than 'doing business with' the comparative data. This may have been a way to avoid the stress or effort associated with the need to change.

Many examiners drew wrong or incomplete interpretations. Their interpretation was sometimes derailed by the substantial cognitive tasks. However, more often, cognitive anchoring and affective responses such as personalisation derailed interpretation.

Finally, those examiners who successfully interpreted the comparative data, and reached the threshold for addressing change, did not plan simply to adjust their practice. Instead, they planned to incorporate the new information provided by the feedback into their existing decision-making frameworks. Several examiners expressed firm convictions about the importance of being true to their judgement. They described the values and aphorisms which help to frame their judgements. The need for change was interpreted through, and incorporated into these values and aphorisms.

Strengths and limitations

We have observed these phenomena in examiners receiving comparative data about their judgements in the form of written feedback. Clearly, this is quite a specific way of influencing examiner performance, and different examiner training interventions are likely to elicit different phenomena. But our premise is that, in all training interventions, examiners engage with information that prompts them to consider how they assess and whether they should change. Consequently, the pheonomena we have observed not only illuminate responses to feedback, but are likely also to illuminate examiners' variable responses to other training interventions.

We have located our study in a specific context and selected particular examiners from within that context using a locality-based sampling procedure and a pragmatic sample size. The examination context will not be representative of all assessments; and the sample of examiners is not representative of all examiners. This is the nature of phenomenology. We are interested in how particular individuals engage with the phenomenon of interest in their context. Our aim is to study the phenomenon in some depth rather than to achieve breadth. Nevertheless, the themes that we have observed have generated some useful hypotheses for further investigation.

Finally, it's important to acknowledge that the comparative data are challenging to interpret and have some inherent limitations. In particular, the position of the examiner's scores on all three charts might have arisen for one of several possible reasons (see chart legends for details). It's possible that clearer or more conclusive data, perhaps gathered over several rounds of examinations, might have produced different examiner responses.

Interpretation

The five sense-making processes we have observed are summarised in figure 4: *testing personal relevance, interpretation, attribution, considering the need for change,* and *considering the nature of change*. These show that sense-making took the form of reflective self-questioning. Schon (1983) explains that this 'aroused' state of reflection emerges when relatively unconscious 'knowing-in-action' reaches its limit and more conscious 'reflection-in-action' is invoked. Malthouse et al. (2015) show that arousal may be driven by cognitive demand (a learning phenomenon), or perceived lack of control over events (situated reflective questioning – a social/organisational phenomenon). Sargeant et al. 2009 have shown that receiving feedback is one of the experiences that triggers this kind of reflection and our findings are consistent with theirs.

The three meta-themes we have observed are summarised in figure 4 – *guarded curiosity*, *dysfunctional assimilation*, and *moderated conservatism*.

The *guarded curiosity* that we have observed in examiners resonates with Watling et. al.'s (2012) observations about educational feedback. Our data show some examiners make credibility and relevance judgements about feedback on their performance in just the same way that learners evaluate the credibility of educational feedback. Watling considers that this credibility test helps learners to decide whether feedback justifies the significant investment of responding. We also found examiners using it as a 'threshold test' before seriously

considering change. This raises the possibility that the similar regulatory frameworks may operate in assessors considering change as in learners.

Our observations about cognitive anchoring and affective responses causing *dysfunctional assimilation* resonate with Eva et. al. (2012) and are part of the socio-affective dimension described by Yang and Carless (2013) from across the feedback literature.

Moderated conservatism describes how some of our examiners sought to align their internal cognitive frameworks with the alternatives frameworks provided by the comparative data. Previous work examining the cognitive variation between assessors has identified relatively abstract themes. For example, Kogan et. al. (2011) found assessors varied in their frames of reference, their use of inference, their means of synthesising information, and their attention to external factors. Whilst Yates et. al (2013) found three themes: differential salience (valuing the same things differently), criterion uncertainty (valuing things that are different), and information integration (constructing singularity differently). However, our data catch assessors in the act of weighing up whether to change their means of judgement. Our findings suggest that some have a strong loyalty to existing means of judgement like Kogan's later (2015) study where assessors disliked training that offered frameworks (perhaps preferring to be left to their own devices), and were reluctant to change their existing means of judgement. Furthermore, our themes suggest that some examiners' judgements are mediated through values, aphorisms, and self-awareness. These are themes that have not, to our knowledge, been described in judgement-based assessment before, and may be worthy of further investigation. Interestingly, they seem to mirror the heuristics that clinicians use in making clinical judgements about patients (Harissis 2009).

Recommendations

The phenomena we have observed may be typical of assessors considering how they make judgements and whether they should change, or they may be restricted to feedback-driven interventions. Further work is required to establish their generality.

However, even on the basis of this data, we would suggest that the assessor is best viewed as a complex cognitive and affective entity. Taking the 'coexistent' view that assessors are simultaneously trainable, fallible *and* meaningfully different, any efforts at assessor training almost certainly need to use a face-to-face, two-way format that can incorporate individuality and affective responses. When using feedback, the 'emotions, content, outcome' (ECO) model provides a possible approach. Sargeant et al (2011), used the ECO as a model to guide multisource feedback for GPs and found: "exploring emotions and clarifying content appeared integral to accepting and using the feedback." Trainers should be prepared to help assessors analyse their means of judgement critically rather than simply provide them with alternatives. Assessors may need support with deciding how to incorporate any new information alongside their existing means of making assessment judgements.

Beyond the research question, the role of values, aphorisms, attention and self-knowledge in mediating judgement-based assessment is an intriguing new insight which deserves further exploration.

Practice points

- Assessors make their judgements in meaningfully different ways (already known)
- Assessor training may well not improve assessment metrics (already known)
- Some reasons for the failure of assessor training may include guardedness towards assessment-modifying information, dysfunctional assimilation of such information, and a tendency to conserve existing means of judgement

- We recommend more sophisticated, examiner-centred, face-to-face training interventions and further research

Acknowledgments: The authors wish to thank Lee Smith, ICBSE manager, for co-ordinating examiner recruitment.

Declaration of interests: The ICBSE funded one investigator's travel to Ireland to undertake interviews. The authors do not consider that this raises any conflict of interest and have no other relevant information to report.

References

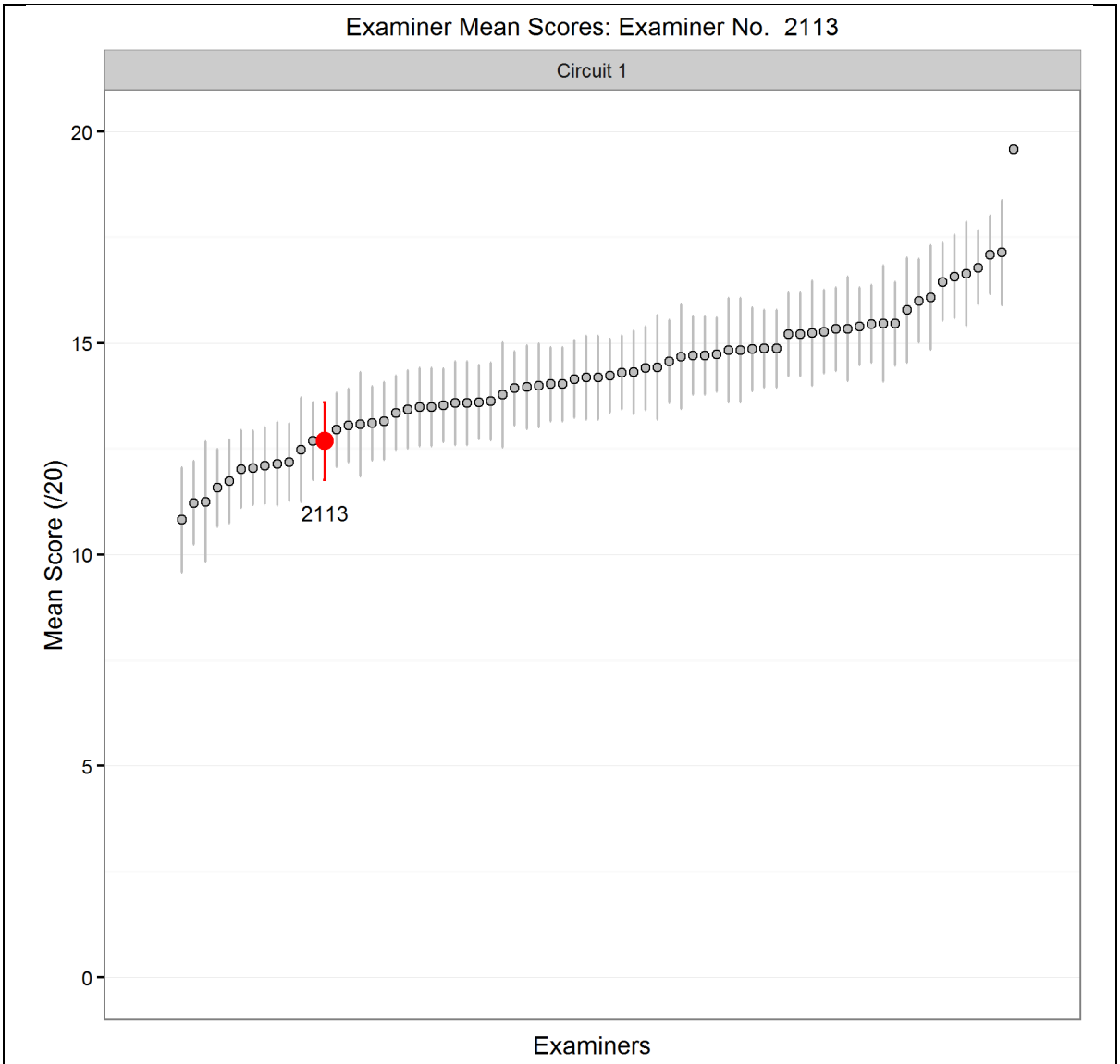
1. Cook D, Dupras D, Beckman T, Thomas K, Pankratz S. Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. J Gen Intern Med. 2008 Dec:24(1):74–79.
2. Elder C, Knoch U, Barkhuizen G, von Randow J. Individual feedback to enhance rater training: Does it work? Language Assessment Quarterly: An International Journal 2005 July;2(3):175-196.
3. Eva K, Armson H, Holmboe E, Lockyer J, Loney E, Mann K, Sargeant J. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. Adv Health Sci Ed. 2012 Mar;17(1):15-26.
4. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. Med Educ. 2014 Nov;48(11):1055–1068.
5. Greatorex J, and Bell JF. What makes AS marking reliable? An experiment with some stages from the standardisation process. Res Papers in Educ. 2008 Sept;23(3):333-355.
6. Griffin P, Robertson P. Judgement-based assessment. In: Griffin P, ed. Assessment for Teaching. Cambridge: Cambridge University Press; 2014:107.
7. Harissis H. Heuristics and evidence: there is room for both. BMJ 2009 June;338:b1860
8. Holmboe E, Hawkins R, Huot S. Effects of training in direct observation of medical residents' clinical competence: a randomised trial. Ann Intern Med. 2004 June;140(11):874–81.
9. Knoch U. Investigating the effectiveness of individualized feedback to rating behaviour – a longitudinal study. Lang test. 2011 Apr;28(2):179-200.

10. 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):1048-1060.
11. Kogan J, Conforti L, 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.
12. Malthouse R, Watts M, Roffey-Barentsen J. Reflective questions, self-questioning and managing professionally situated practice. *Res in Educ*. 2015 Nov;94(1):71-87.
13. Ringsted C, Ostergaard D, Ravn L, Pedersen J, Berlac P, & van der Vleuten C. A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. Med Teach. 2003 Nov;25(6):654–658.
14. Roberts C, Jorm C, Gentilcore S, Crossley J. Peer assessment of professional behaviours in problem-based learning groups. Med Educ. 2017 Apr;51(4):390-400.
15. Sargeant J, Mann K, van der Vleuten C, & Metsemakers J. Reflection: a link between receiving and using assessment feedback. Adv Health Sci Educ. 2009 14(3):399-410.
16. Sargeant J, McNaughton E, Mercer S, Murphy D, Sullivan P, & Bruce D. Providing feedback: exploring a model (emotion, content, outcomes) for facilitating multisource feedback. Med Teach. 2011 33(9):744-749.
17. Schön D. The reflective practitioner: How professionals think in action. Basic Books, New York; 1983.
18. Watling C, Driessen E, van der Vleuten CP, Lingard L. Learning from clinical work: the roles of learning cues and credibility judgements. Med Educ. 2012 Feb;46(2):192–200.
19. Yang M, Carless D. The feedback triangle and the enhancement of dialogic feedback processes. Teach Higher Educ. 2013 18(3):285-297.

20. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly observed performance assessments. Adv Health Sci Educ Theory Pract. 2013 Aug;18(3):325–41.

Figures

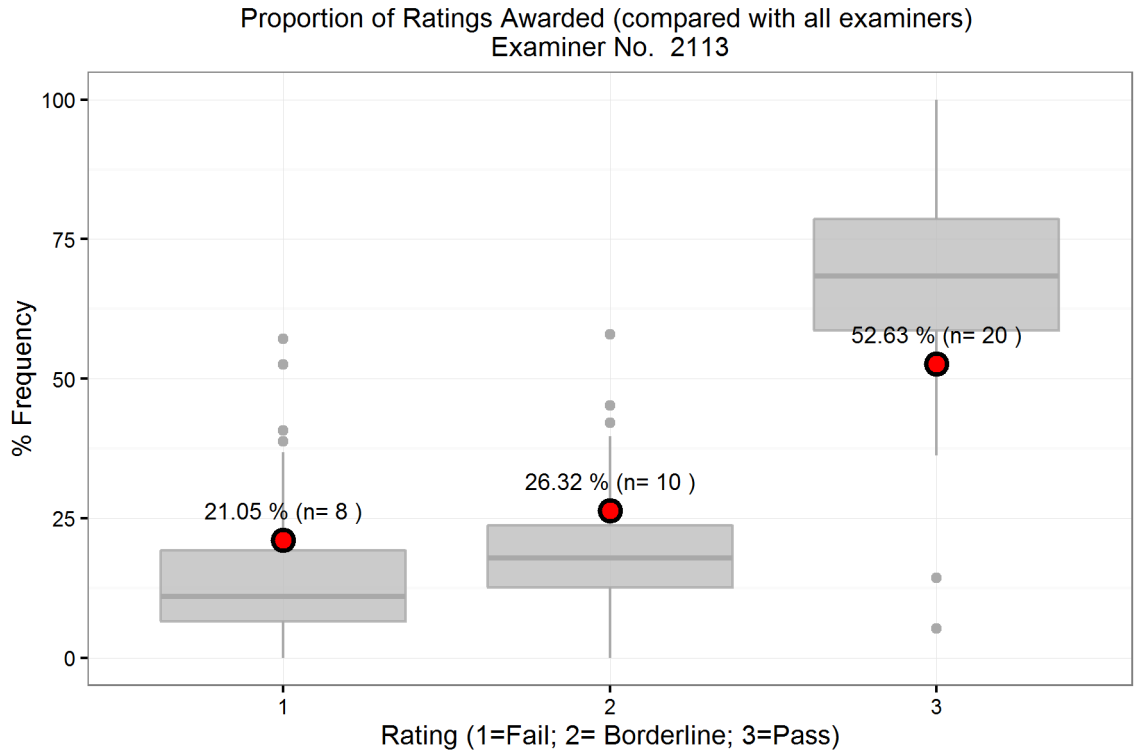
Fig 1: example of error-bar chart and explanation used for feedback



This graph shows the mean score (non-pilot stations only) provided by each examiner across all the candidates they examined (the dots), and the 95% confidence interval either side of that mean score (the whiskers). Your mean score and confidence interval feature in red on the graph. If you find yourself towards either extremes of the scale, you may wish to reflect on your own leniency.

Legend: The error bar chart displays the overall mean and spread of every examiner's scores. Examiners are ranked along the x-axis from the lowest scoring (left-hand end) to the highest scoring (right-hand end). Examiners are compared across stations. Consequently, a relatively low mean score (left of centre) might result from relatively stringent marking or from marking a relatively difficult station, and vice versa.

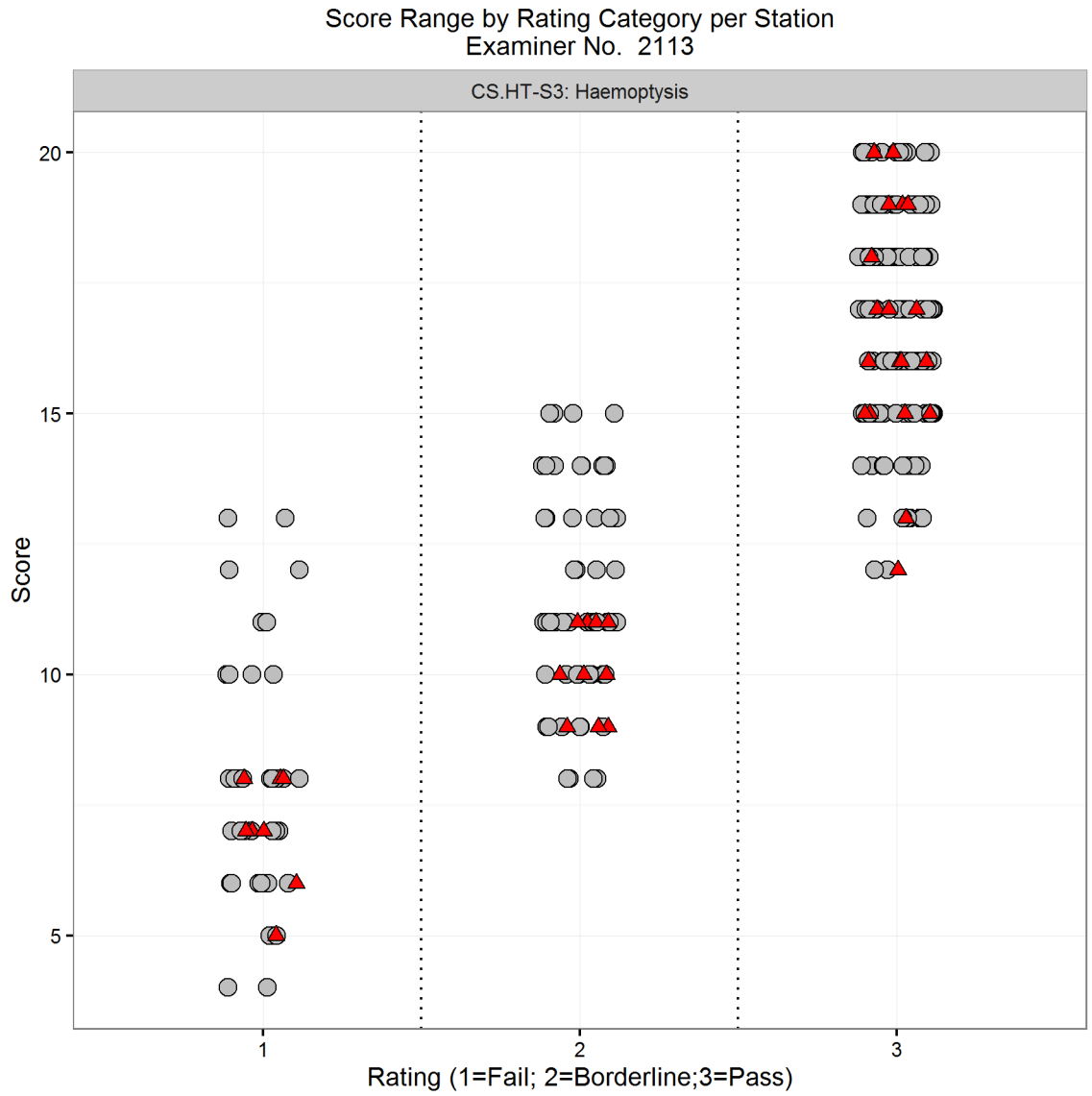
Fig 2: example of box and whisker chart and explanation used for feedback



This graph shows a box plot for each global judgement. The line in the middle of each box represents the median number of global judgements awarded and the top and bottom of the box represent the interquartile range. The whiskers indicate the range of marks excluding outlier examiners. The number of global judgements you awarded are represented by the red dots. If you are outside the box for your global judgements (even more so if you are an outlier) you may wish to reflect on your marking.

Legend: The box and whisker plot displays the proportion of global judgements (red dot) that the examiner allocated to each of: pass, borderline and fail (categories on the x-axis). For comparison, the boxes and whiskers summarise the allocations from all examiners. Again, the plot compares examiners across stations so their scores may reflect either their marking or the difficulty of their station. These global judgements determine the final cut-score but do not differentially affect the outcome for the candidates the examiner marked.

Fig 3: example of dot plot and explanation used for feedback



The red triangles show the scores obtained by candidates who were examined by you, the grey dots those examined by others, on each station. Assuming that the candidates you examined were representative of the candidature as a whole, we would expect the distribution of the marks you awarded to be similar to the marks awarded by other examiners. If this is not the case you may wish to reflect on whether the standard you were applying in the station was similar to other examiners or whether the candidates you examined did not reflect the whole cohort.

Legend: The dot plot combines both scores and global judgements station by station. Examiners could see (with difficulty) the proportion (relative number of triangles) that they allocated to pass, borderline and fail and relate these proportions to other examiners at the same station. They could also relate score (y-axis position) to global category (x-axis position) to check that they allocate higher-scoring candidates to higher categories and to see if their thresholds were similar to other examiners.

Table 1: primary themes clustered in to process with illustrative quotes

Process	Primary theme descriptor	Illustrative example
Testing personal relevance	expresses appetite for feedback	"pleased to get this, and I thought: I'll give this serious reading" (05/10/16/3/M)
	questions applicability of feedback	"are these comments really about me?" (05/10/16/1/M)
	questions believability of feedback	"surprised because I don't remember failing anyone in 2015" (05/10/16/1/M)
	makes conclusions vague	"I'm not a particularly high scorer, but I don't know what to say about that really" (14/10/16/4/M)
	distances self from conclusions	"I took one look and thought, ooh, this is above my pay station" (28/09/16/1/F)
Interpretation	compares self with peers	"trying to work out if I was middling, in between, or totally out of tangent" (05/10/16/2/F)
	tries to interpret chart	"why is the dot up there and not down there in that box" (28/09/16/1/F)
	judges proportion or significance	"bit unsure about this one – I take it we should be in the box" (08/11/16/1/F)
	integrates several pieces of information	"on occasion I'm too hard, on occasion I'm too easy; overall it balances out" (08/11/16/2/M)
Attribution	attributes to self	"I first look at myself" (28/09/16/2/M)
	attributes to something else	"I've been examining for the last 10 years; the COPD was a very easy station" (14/10/16/2/M) "God, these people know their anatomy well" (02/11/16/1/M)
	cognitive anchoring	"you think 'hang on, two dots out of the three are outside of the boxes, this confirms that I'm an outlier'." (12/02/14/1/M)
Considering the need for change	expresses relief	"phew I'm not an outlier; that's reassuring." (28/09/16/3/M)
	expresses unease	"it felt worse than my exit exam!" (05/10/16/1/M)
	learns something new	"the trends are I am a strict examiner compared with what I believed" (28/09/16/2/M)
Considering the nature of change	expresses resistance to change	"I hope it wouldn't change me, I don't want to have to think about it." (28/09/16/1/F)
	change via values and aphorisms	"change my behaviour? Yes, it's reasonable to think: 'you've got to earn your pass'" (05/10/16/3/M)
	change via self-regulation	"I'm less inclined to put somebody on borderline - more definitive judgement" (08/11/16/2/M)

Fig 4: themes, meta-themes and their relationships