Marking parties for marking written assessments

Medical Teacher

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/imte20

Marking parties for marking written assessments: A spontaneous community of practice

Emma Vaccari, Joyce Moonen-van Loon, Cees Van der Vleuten, Paula Hunt & Bruce McManus

To cite this article: Emma Vaccari, Joyce Moonen-van Loon, Cees Van der Vleuten, Paula Hunt & Bruce McManus (02 Oct 2023): Marking parties for marking written assessments: A spontaneous community of practice, Medical Teacher, DOI: 10.1080/0142159X.2023.2262102

To link to this article: https://doi.org/10.1080/0142159X.2023.2262102

Published online: 02 Oct 2023.

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2023.2262102

Marking parties for marking written assessments: A spontaneous community of practice

Emma Vaccari[a], Joyce Moonen-van Loon[b], Cees Van der Vleuten[b], Paula Hunt[a] and Bruce McManus[a]

[a]Faculty of Medicine, University of Southampton, Southampton, UK; [b]School of Health Professions Education, Maastricht University, Maastricht, The Netherlands

ABSTRACT

In programmes of assessment with both high and low-stakes assessments, the inclusion of open-ended long answer questions in the high-stakes examination can contribute to driving deeper learning among students. However, in larger institutions, this would generate a seemingly insurmountable marking workload. In this study, we use a focused ethnographic approach to explore how such a marking endeavour can be tackled efficiently and pragmatically. In marking parties, examiners come together to individually mark student papers. This study focuses on marking parties for two separate tasks assessing written clinical communication in medical school finals at Southampton, UK. Data collected included field notes from 21.3 h of marking parties, details of demographics and clinical and educational experience of examiners, examiners' written answers to an open-ended post-marking party questionnaire, an in-depth interview and details of the actual marks assigned during the marking parties. In a landscape of examiners who are busy clinicians and rarely interact with each other educationally, marking parties represent a spontaneous and sustainable community of practice, with functions extending beyond the mere marking of exams. These include benchmarking, learning, managing biases and exam development. Despite the intensity of the work, marking parties built camaraderie and were considered fun and motivating.

KEYWORDS

Assessment; written assessment; undergraduate phase of education; staff development

Introduction

Faculty who construct assessment systems within competency-based medical education are faced with conflicting demands they need to balance. Among these is the tension between needing to drive learning behaviour and needing to make progression decisions (Tavakol and Dennick 2017). Many programmes of assessment in medical schools use a combination of high and low-stakes assessments. In some cases, progression decisions are based on a small number of high stakes assessments, although some are moving away from this model, such as those using programmatic assessment, which require these decisions to be made in competence committees, taking into account multiple data points (Heeneman et al. 2021). Low-stakes assessments often have a greater formative component, aiming to give useful feedback and direction. However, students tend to focus on passing the high-stakes examinations and adapt their behaviour accordingly (Cilliers et al. 2010). As it is possible to produce Multiple Choice Questions that assess higher-order cognitive functions (Hift 2014), their comparative ease of marking may sway larger organisations to choose this type of assessment over other types of written assessment. However, their perception as a simple form of assessment tends to drive students to more superficial learning than longer open-ended type questions (Cilliers et al. 2010).

Assessment can help direct student learning towards important clinical tasks, for example written clinical communication. Clear, concise and accurate written communication is a skill that medical students may struggle with (Rawson et al. 2005), but one which is very important in clinical practice (Michell et al. 2012; ACSQHC 2017). The 2018 Ottawa consensus statement on good assessment suggested that 'difficult to measure' competencies, like record keeping, should be included in systems of assessment (Norcini et al. 2018). Assessments of clinical documentation, for example completing a patient note following a standardised patient encounter, can be used to assess clinical reasoning (Yudkowsky et al. 2015). Indeed, increasing the authenticity in assessment may contribute to narrowing the uncomfortable gap between what students feel they should learn to pass the exam and what they feel they should learn to become better clinicians (Cilliers et al. 2010). However, including authentic-feeling assessments of written clinical communication skills brings the problem of how to mark them. The experience with the Patient Note exercise in the USA has shown that there is little agreement between

Practice points

- Marking parties can bring fun into the potentially tedious task of marking long answer questions.
- Marking parties play a role in the development of shared standards, examiner training and benchmarking, and managing bias.
- Marking parties can be seen as spontaneous communities of practice.

CONTACT Emma Vaccari vaccari.emma@gmail.com Faculty of Medicine, University of Southampton, Academic Centre, College Keep 4-12 Terminus Terrace, Southampton SO14 3DT, UK.

individual clinician-raters when scoring patient notes (Boulet 2004). This is to be expected given the complexity of the skill being assessed and reflects the tension between authenticity and standardisation (Govaerts et al. 2019). To mitigate for this, traditionally each answer to an open-ended question would be marked by a single examiner (Downing 2010). While this would increase reliability in a single setting, the examiner variability would then play out in a loss of equivalency across different cycles of testing. To address the challenges presented by the need to balance validity, reproducibility and equivalence (Norcini et al. 2011, 2018), some institutions (Clauser et al. 2008; Wilcox et al. 2020) choose to assess written clinical communication skills over a number of different 'stations', in a similar way to an Objective Structured Clinical Examination (OSCE) (Khan et al. 2013).

Given the impact of high-stakes assessments on learning behaviour, it seems important to include assessments that are as authentic as possible and related to learning in the workplace. So why are high-stakes assessments of written clinical communication not more mainstream? It is possible that this has to do with a perceived lack of acceptability, particularly in terms of workload. Marking hundreds or even thousands of instances of clinical documentation may seem like a daunting or even impossible task, especially for medical schools, where examiners are often clinicians, already very busy with their clinical tasks. In the field of Business Studies, where discursive assessments are more commonly used, Price (2005) found that 'marking bees' were considered to be time-saving and effective by module leads who had tight deadlines to work to. In such meetings, marking and moderation are conflated: examiners individually mark papers, but as they do it in the same room, they have the option to engage in discussion with each other. It is possible that this kind of meeting also contributes to the development of a community of practice within medical education, with a shared knowledge base, beliefs, values and experiences. Faculty development communities of practice have been suggested as powerful tools to raise the profile of education in academic institutions. In recognition of this, communities of practice are being intentionally set up, modelled on what were originally spontaneous networks of people (de Carvalho-Filho et al. 2020). Cruess et al. (2018) describe how communities of practice in medicine, as in other areas, have three main characteristics: domain, community and practice. In this paper, we examine the marking process in a medical school that uses 'marking bees' or 'marking parties', as they are known locally, to mark a high-stakes assessment of clinical written communication skills. By examining the acceptability of a local solution to the problem of marking high numbers of written assessments, we aim to share our learning with and inspire other institutions grappling with this problem. As well as looking at the explicit function of marking parties, we are interested in investigating a potential unintended but advantageous aspect of marking parties, that is, whether they have the characteristics of a community of practice.

Methods

Setting

The Clinical Summary Exam (CSE) is part of the programme of assessment in the Final Year at Southampton medical school, UK. The CSE assesses students' ability to synthesize clinical information, presented in either written or audio-visual format, and recommunicate it in a variety of different written formats, such as medical note entries, discharge summaries or referral letters (Wilcox et al. 2020). The CSE is made up 6 written tasks. Each task has a marking rubric and model answer. Examiners give numerical marks to each domain heading of the marking rubric. The overall score for that task is determined by a weighted sum of these marks. Examiners also give a 'global impression' score on an A–F scale. This 'global impression' score is used to set the overall numerical pass mark for that task, using the borderline regression method (Schoonheim-Klein et al. 2009). As over 200 students sit the exam every year, several 'marking parties' are organised to mark the output. This study focuses on the marking of two tasks from the CSE.

Study design

An ethnographic approach was used to explore the nature and function of marking parties. EV took field notes during the marking parties for the two tasks. The field notes focused on the interactions between examiners and verbal exchanges were recorded as accurately as possible. EV acted as a 'complete participant' (Pope 2005), as this was her natural position, having previously been involved in marking parties and task design for the CSE. Alongside the primary observations, during and after the marking parties, EV noted reflections and early interpretations (Reeves et al. 2013). The field notes were triangulated with data from a demographics questionnaire completed before the marking parties, an open-ended questionnaire that was sent out to all participating examiners after the marking party and a single in-depth interview. The participant for the in-depth interview was selected though purposive intensity sampling (Palinkas et al. 2015), as someone who could provide rich examples of major emerging themes from the initial analysis. As a complete participant, EV was able to provide a sort of strict reciprocity (Johnson and Rowlands 2012) which is helpful in in-depth interviews. Due to the timing of the COVID-19 pandemic and the fact all participants were clinicians, it was only possible to arrange one such interview. All examiners consented to participate in the observational aspect of the study and completed the demographics questionnaire. A proportion of examiners also consented to completing the open ended questionnaire and an in depth interview. Throughout the study, a research diary was kept to record an audit trail of decisions made and researcher reflections.

In addition the qualitative data, we collected the marks given during the marking parties which were studied. Qualitative analysis was iterative in nature. Field notes, answers to the open ended questionnaire, reflections, and the verbatim transcribed interview were thematically analysed using NVIVO software (QSR International Pty Ltd. 2020). Analysis continued until inductive thematic saturation was achieved (Saunders et al. 2018).

The study received ethical approval from the University of Southampton (ERGO ID 52834).

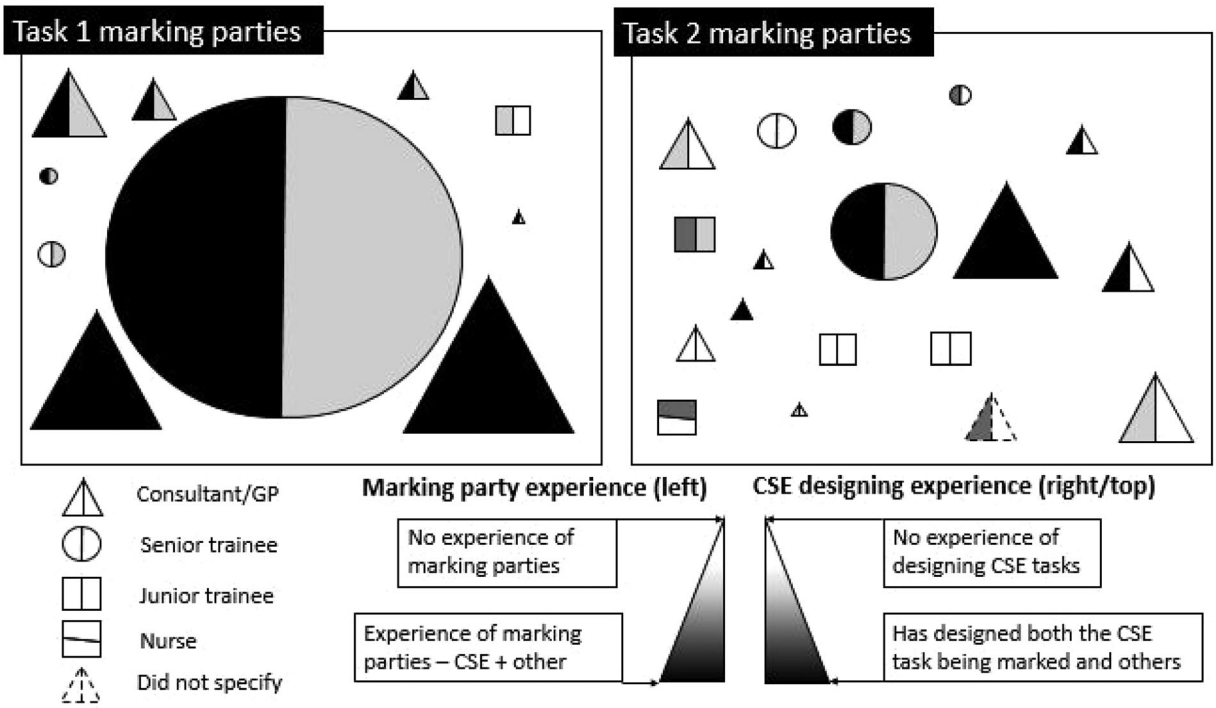

Figure 1. Diagram showing the participants in the marking parties for the two tasks. Each icon represents a participant. The shape shows their clinical seniority. The size is proportional to the amount of time spent in the marking party. The colour on the left of the icon is related to the marking party experience (black: experience of marking parties both in CSE and other exams; dark grey: experience of CSE marking parties; light grey: experience of other marking parties; white: no experience of marking parties). the colour on the right/top of the icon relates to experience designing the CSE (black: has designed the task being marked and others; dark grey: has designed the task being marked; light grey: has designed other CSE tasks; white: no CSE design experience).

Table 1. Distribution of global impression scores by task.

	Task	
Overall	Task 2	Task 1
A (excellent)	8	19
B (good)	48	53
C (clear pass)	82	91
D (borderline pass)	63	32
E (borderline fail)	11	9
F (clear fail)	2	0

Results

Description of marking parties

EV participated in and took detailed field notes over 4 d of marking parties, two for each task. In addition to the field notes, the final qualitative data set was formed of 11 returned email questionnaires, one in depth interview, reflective notes following each marking party and an iterative research diary. The participant was chosen for the in-depth interview due to their ability to provide a different perspective to that of EV as they held a significantly more senior clinical position, but had relatively less involvement with the Clinical Summary Exam. In total, the two Task 1 marking parties lasted 13.3 h and the two Task 2 marking parties lasted 8 h. 10 examiners attended the Task 1 marking parties, each staying between 25 min and 11.3 h (mean 4.9 h). 18 examiners attended the Task 2 marking parties, each staying between 1 and 7.2 h (mean 2.9 h). Figure 1 shows the composition of the marking parties in terms of time, level of clinical experience, and previous experience of marking parties. In both marking parties, some senior examiners dedicated some of the time to exam-related activities other than marking, for example replying to emails or moderating papers marked as fails.

204 Task 1 and 214 Task 2 papers were marked in marking parties, taking an average of 14.5 and 14.8 min per paper, respectively. The overall grades awarded for Task 1 were generally higher than those awarded for Task 2, χ^2 (5) = 17.3, p = .004, as shown in Table 1.

The two marking parties ran differently. For Task 2, the majority of examiners arrived together, viewed the video which students had based their answer on, and were briefed together. The Task 1 marking parties were more fluid, with examiners coming and going at different times, and each examiner being given a brief, tailored to their needs, by a senior faculty member upon arrival. There was no other training outside the marking parties. On both days, the changing mix of examiners and perceived time pressures affected how much discussion occurred. Most of the interactions in the marking parties was related to the marking process, with the sharing of questions, reflections and extracts from student papers, announcing marks, and generally discussing standards. However, the discussion was not limited exclusively to the marking.

Practice: functions of marking parties

Shared standard development

> K: 'I think some examiners marking alone or with less confidence may have followed the [...] model answer too closely, whereas there was agreement in the room that [the model answer] was perhaps too detailed'. [questionnaire. Task 1. Day 2]

An important function of the marking parties was the development of shared standards. This is a complex process that warrants a detailed description beyond the scope of this overview. Broadly, interaction between examiners facilitated the interpretation of written standards in the context of students' actual answers, in conjunction with examiners' personal standards. The seniority and expertise

of examiners determined how much weight their personal standards had in this process. However, examiners who were present for the majority of marking for a particular question provided continuity to the process by recalling previous discussions.

Benchmarking

There was a tension between moderating the dove-hawk effect in the marking party and acknowledging that this process would be done formally during 'moderation', through psychometric analysis and double marking. In two instances, during the task 2 marking parties two examiners suggested they wanted to give a specific mark (a 'B' and an 'F' respectively) and a senior member of faculty suggested that they give the global impression score they felt was appropriate as this would then be moderated. On other occasions, examiners had an opportunity to review the mark they would have given as a result of interaction.

> E: 'I'd like to have a good one … '
>
> G: 'look at my first one'
>
> After looking over it, E comments: 'it's all right'.
>
> G: 'what would you have given it?'.
>
> E: 'probably a B. It's not perfect, is it?'.
>
> B: 'perfect is not in the description for an excellent'. [Field notes. Task 1. Day 2]

In many other instances, marks were announced and did not lead to an indication of whether the mark was appropriate or not, but did give an indication of the spread of marks being given.

> X: 'B!'
>
> D: 'B is good! Which is better than C, a clear pass, which is better than D, a borderline pass etc'
>
> X: 'So most people should get Cs or Ds?'
>
> G: 'Don't worry, mine will make up for it. I've failed a few!' [Field notes. Task 2. Day 1]

Learning

In many interactions, as shown above, discussion allowed personal marking standards to be updated, influencing the marking process. In others, however, there were instances of more overt teaching and learning, as in this exchange:

> AA, reading from a student paper: 'Furrowed brow?'
>
> E and G furrow their brows.
>
> D explains how counting the furrows used to be a clinical sign. [Field notes. Task 2. Day 1]

In other cases, the student papers triggered learning for the whole group:

> O, reading from a student paper: 'walking corpse syndrome?'
>
> Blank expressions around the room.
>
> E googles out loud: 'walking corpse syndrome … It's Cotard's!' [Field notes. Task 2. Day 1]
>
> D: 'and we've all learned something new!'

In addition to clinical learning, there was a degree of ad hoc examiner training. This was often directly related to a student paper (e.g. how to apply written standards, what to do when faced with a paper labelled as written by a student with an additional educational requirement). Other times, training was more general and brought forward experience from previous marking parties, as this example from the introduction to the first Task 2 marking party.

> '[The leader] suggested people start reading the answers from the end, at the impression/diagnosis section as she has found this helpful when marking [this] task in the past.' [Reflections. Task 2. Day 1]

Managing bias

There were several instances where discussion helped manage personal bias. Papers of students with dyslexia or other learning difficulties were marked with a sticker. Examiners picking up a paper of a student with dyslexia, marked with a sticker, would often discuss with the group how this would affect marking. However, interaction allowed more hidden sources of bias to be revealed. In one case, the use of a particular terminology was reattributed to the student belonging to the international cohort of students, rather than it being considered a mistake.

> A: 'am I being thick here? Morbus parkinsons??'
>
> E: 'I think it's old fashioned'
>
> K: 'I bet they're German. What do their [number] ones look like?'
>
> E, after googling: 'yes, it is German'. [Field notes. Task 1. Day 2]

During the task 2 marking parties, the group leader made a point of highlighting potential areas of bias in marking:

> G and T share some comments on how handwriting affects their initial impressions.
>
> D: 'Please be aware of our conscious and unconscious biases' [Field notes. Task 2. Day 1]

Exam development

A small, but qualitatively significant number of interactions related to developing the exam question that was being marked. For example:

> E: 'I'm looking at the model answer and I think I would have marked it down on several things from the criteria I have formed in my head'
>
> K: 'agreed'
>
> A: 'you can rewrite it for us if you want … ' [field notes. Task 1. Day 2]

Community: a friendly atmosphere

The atmosphere in all marking parties was friendly. In their post-marking party questionnaires, examiners identified that marking parties were more 'fun' [E], 'motivating' [S] and 'sociable' [C] than individual marking. In the interview, F suggested that the frequent discussions could be 'irritating', and further elaborated:

> 'I tend to be able to tend to work better if I'm in a quiet room, focused. So any sort of an interruption definitely slowed me down, probably more … perhaps disproportionately, 'cause I felt like I had to listen to what was being said before being able to carry on.' [Interview 1]

However, F also commented that 'humour' and 'camaraderie' contributed to the feeling of wanting 'to do the job and to do it well'. This seemed to trump any frustration experienced as a result of the interruptions to the marking task:

> 'It was actually a really enjoyable experience for me. In terms of activities that you can do to tick your [Continuing Professional Development] box, I think this should be high up there, as a way of meeting peers and learning something and contributing to education. […] I think there's a lot of other benefits to the marking party in terms of […] peer support, and that feeling of camaraderie and that sense of belonging that] …] are positive. And these are things that tend to be lacking in the NHS at the moment and in training in general … ' [Interview 1]

Humour was used frequently throughout the marking parties and contributed to the positive and enjoyable atmosphere. Humour was often related directly to unusual or amusing turns of phrase or approaches to answering the question by students:

> '"Blood work": this one has watched too much ER!' [field notes. Task 1. Day 2].

In a couple of instances, humour was not linked to the marking process at all, but in most others it was and it included mock shock reactions 'good' papers or examiner standards, sharing disappointment at 'poor' performance and jokes about the exam question, student answers and even making friendly humorous remarks about examiners' personal circumstances.

There were many instances where emotions were shared with the group, for example relief and even joy when marking a good paper or disappointment when confronted with poorer examples. In some instances, discussion allayed junior examiners' fear of the consequences of failing a particular paper:

> E: '[my first] fail!'
>
> M: 'what happens if someone fails?'
>
> A: 'it's only 10% of the mark, don't worry. You're not ending someone's career'
>
> M: 'it's a lot of pressure. I want them to get it right!' [field notes. Task 1. Day 1]

Discussion

Marking parties can be considered examples of spontaneous communities of practice, as described by de Carvalho-Filho et al. (2020) and demonstrated by the 'feeling of belonging' that results from marking party membership. Although spontaneous communities of practice require less effort to sustain, viewing marking parties in this light opens avenues for optimisation. Figure 1 demonstrates the involvement of both experts and novice examiners in the CSE marking parties. This ensures the community of practice is rich in expertise (memory) and innovation (new ideas) (de Carvalho-Filho et al. 2020). The friendly and welcoming atmosphere of the marking parties create a fertile ground for the community to develop and, in the word of one of the participants, make members 'want to do the work and do it well'. The work of the marking party, as shown above, extends beyond pure marking and contributes to the development of individuals as examiners and more generally as educators. There are even some instances of clinical learning, which nods to the fact clinicians are part of multiple communities of practice including education communities, clinical specialty communities and others (Cruess et al. 2018). Moving from the periphery to the centre of a community of practice requires the development of a shared identity, knowledge and skills. Much of this is tacit knowledge (Cruess et al. 2018). Due to the primary task of the marking party, much of the knowledge and skills developed are related to examining. Examiners in our study seemed to develop a shared and sometimes unspoken understanding of what quality looks like, as well as a shared ideal of what a 'good' range of marks would be. On some occasions, the shared knowledge in the room allowed individuals' biases to be identified and moderated. If the more explicit aspects of examiner training were viewed through the formal lens of instructional design (Van Merriënboer and Kirschner 2017), it could be said that it allows the provision of supportive information at the exact point the novice examiner requires it, as happens in the case of marking papers of students with learning difficulties. However, the practice of developing new examiners must be balanced with the need for the exam to be fair and reliable. How new and established members of the community of practice influence the development of shared standards for medical students is something that is worth exploring further.

As communities of practice, marking parties would appear to be an acceptable approach to marking hundreds of exam papers. Where complex skills are assessed, such as the CSE, examiner variability is to be expected (Singh 2021). There are parallels to be drawn between the CSE and OSCEs. In large medical schools, a number of different examiners will mark the same station/task for different students (Yeates et al. 2021). At a speculative level, it is possible that the interactions between examiners that occur in the CSE marking party contribute to mitigating the effect of individual-examiner and examiner-cohort variability. Both OSCEs and CSE assess the application of skills and knowledge. The fact that some of our participants experienced anxiety around failing students is at least in part a testament to the complexity of the tasks being required of students, the quality of which could not be reduced to a simple set of explicit written criteria. Group interactions moderated these anxieties. The 'failure to fail' phenomenon has been widely documented in the context of workplace based assessments (Yepes-Rios et al. 2016) and has been discussed in the context of OSCEs (Shulruf et al. 2018). However, it is not typically something that is discussed regarding written assessments. The marking party provides a potential avenue to manage this phenomenon in the context of written assessment. These considerations may have some applicability in the marking of large national assessments of written clinical skills, such as within the United States Medical Licencing Examination (Clauser et al. 2008). Finally, examiner uncertainty can be particularly prevalent and consequential for borderline performances (Shulruf et al. 2018). It would be interesting to investigate whether individual examiners preferentially bring borderline cases to the attention of the group for discussion.

As with all research, there are some limitations to our study. In terms of qualitative methodology, EV's role as a

complete participant required a balance between observing, documenting field notes and actually participating in the marking effort. As described by Andreassen et al. (2020), compared to traditional ethnography, focussed ethnography requires a high intensity of work during the observation phase. The trade-off between the intensity of observations and the acceptability of EV's presence as a researcher was compensated for by triangulating the field notes with other methods of data collection. Of course, a greater number of interviews would have been preferable, to be sure sampling was sufficient for theoretical saturation (Saunders et al. 2018), but the COVID pandemic redirected priorities elsewhere. Despite these limitations, our study gives an example of a practical novel approach to marking, which takes the form of a spontaneous community of practice. Educational communities of practice are particularly important in medical education, as clinical practice may often take priority over education. In addition to this, our paper widens the debate around how to address the 'failure to fail' phenomenon and how to moderate examiner biases in high-stakes assessments.

Disclosure statement

Emma Vaccari has attended a course on ADHD, with expenses paid for by Takeda.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Notes on contributors

Emma Vaccari is training in General Psychiatry.

Joyce Moonen-van Loon and ***Cees Van der Vleuten*** work in Educational Development and Research at Maastricht, where Cees is a Professor of Education.

Paula Hunt and ***Bruce McManus*** lead the Southampton Faculty of Medicine Assessment Team. Paula is a Senior Clinical Lecturer and Bruce is a Professor Fellow of Medical Education.

References

Andreassen P, Christensen MK, Møller JE. 2020. Focused ethnography as an approach in medical education research. Med Educ. 54(4): 296–302. doi: 10.1111/medu.14045. 31850537.

Australian Commission on Safety and Quality in Health Care (ACSQHC). 2017. Improving documentation at transitions of care for complex patients. Sydney, NSW: ACSQHC; p. 3–26.

Boulet JR, Rebbecchi TA, Denton EC, McKinley DW, Whelan GP. 2004. Assessing the written communication skills of medical school graduates. Adv Health Sci Educ Theory Pract. 9(1):47–60. doi: 10.1023/B:AHSE.0000012216.39378.15.

Cilliers FJ, Schuwirth LW, Adendorff HJ, Herman N, Van der Vleuten CP. 2010. The mechanism of impact of summative assessment on medical students' learning. Adv Health Sci Educ Theory Pract. 15(5):695–715. doi: 10.1007/s10459-010-9232-9.

Clauser BE, Harik P, Margolis MJ, Mee J, Swygert K, Rebbecchi T. 2008. The generalizability of documentation scores from the USMLE step 2 clinical skills examination. Acad Med. 83(10 Suppl):S41–S44. doi: 10.1097/ACM.0b013e318183cd1d.

Cruess RL, Cruess SR, Steinert Y. 2018. Medicine as a community of practice: implications for medical education. Acad Med. 93(2): 185–191. doi: 10.1097/ACM.0000000000001826.

de Carvalho-Filho MA, Tio RA, Steinert Y. 2020. Twelve tips for implementing a community of practice for faculty development. Medical Teacher. 42(2):143–149. doi: 10.1080/0142159X.2018.1552782.

Downing S. 2010. Assessment in health professions education. London: Routledge Ltd. doi: 10.1080/0142159X.2018.1552782.

Govaerts MJ, van der Vleuten CP, Holmboe ES. 2019. Managing tensions in assessment: moving beyond either–or thinking. Med Educ. 53(1):64–75. doi: 10.1111/medu.13656.

Heeneman S, de Jong LH, Dawson LJ, Wilkinson TJ, Ryan A, Tait GR, Rice N, Torre D, Freeman A, van der Vleuten CPM. 2021. Ottawa 2020 consensus statement for programmatic assessment–1. Agreement on the principles. Med Teach. 43(10):1139–1148. doi: 10.1080/0142159X.2021.1957088.

Hift RJ. 2014. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ. 14(1):249. doi: 10.1186/s12909-014-0249-2.

Johnson JM, Rowlands T. 2012. The interpersonal dynamics of in-depth interviewing. In: Gubrium, JF, Holstein JA, Marvasti AB, McKinney KD, editors. The SAGE handbook of interview research: the complexity of the craft. Thousand Oaks: Sage; p. 99–113.

Khan KZ, Gaunt K, Ramachandran S, Pushkar P. 2013. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part II: organisation & administration. Med Teach. 35(9):e1447–e1463. doi: 10.3109/0142159X.2013.818635.

Michell V, Tehrani J, Liu K. 2012. Are clinical documents optimised for patient safety? A critical analysis of patient safety outcomes using the EDA error model. Health Policy Technol. 1(4):214–227. doi: 10.1016/j.hlpt.2012.10.003.

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, et al. 2011. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 33(3):206–214. doi: 10.3109/0142159X.2011.551559.

Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D. 2018. 2018 Consensus framework for good assessment. Med Teach. 40(11): 1102–1109. doi: 10.1080/0142159X.2018.1500016.

Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health. 42(5):533–544. doi: 10.1007/s10488-013-0528-y.

Pope C. 2005. Conducting ethnography in medical settings. Med Educ. 39(12):1180–1187. doi: 10.1111/j.1365-2929.2005.02330.x.

Price M. 2005. Assessment standards: the role of communities of practice and the scholarship of assessment. Assess Eval High Educ. 30(3):215–230.

QSR International Pty Ltd. 2020. NVivo (released in March 2020). https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.

Rawson RE, Quinlan KM, Cooper BJ, Fewtrell C, Matlow JR. 2005. Writing-skills development in the health professions. Teach Learn Med. 17(3):233–238. doi: 10.1207/s15328015tlm1703_6.

Reeves S, Peller J, Goldman J, Kitto S. 2013. Ethnography in qualitative educational research: AMEE Guide No. 80. Med Teach. 35(8):e1365–e1379. doi: 10.3109/0142159X.2013.804977. 23808715.

Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, Burroughs H, Jinks C. 2018. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant. 52(4):1893–1907. doi: 10.1007/s11135-017-0574-8.

Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, Van Der Vleuten C, Van der Velden U. 2009. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. Eur J Dent Educ. 13(3):162–171. doi: 10.1111/j.1600-0579.2008.00568.x.

Shulruf B, Adelstein B-A, Damodaran A, Harris P, Kennedy S, O'Sullivan A, Taylor S. 2018. Borderline grades in high stakes clinical examinations: resolving examiner uncertainty. BMC Med Educ. 18(1):272. doi: 10.1186/s12909-018-1382-0.

Singh T. 2021. Principles of assessment in medical education. New Delhi: Jaypee Brothers Medical Publishers.

Tavakol M, Dennick R. 2017. The foundations of measurement and assessment in medical education. Med Teach. 39(10):1010–1015. doi: 10.1080/0142159X.2017.1359521.

Van Merriënboer JJ, Kirschner PA. 2017. Ten steps to complex learning: A systematic approach to four-component instructional design: New York: Routledge.

Wilcox CR, Vaccari E, Whitehurst L, Davey M, McManus B, Hunt P. 2020. The clinical summary exam: a novel approach to assessing information processing skills and improving preparedness for starting work among final-year medical students. Med Teach. 42(5):591–592. doi: 10.1080/0142159X.2019.1646416.

Yeates P, Moult A, Cope N, McCray G, Xilas E, Lovelock T, Vaughan N, Daw D, Fuller R, McKinley RKB. 2021. Measuring the effect of examiner variability in a multiple-circuit objective structured clinical examination (OSCE). Acad Med. 96(8):1189–1196. doi: 10.1097/ACM.0000000000004028.

Yepes-Rios M, Dudek N, Duboyce R, Curtis J, Allard RJ, Varpio L. 2016. The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. Med Teach. 38(11):1092–1099. doi: 10.1080/0142159X.2016.1215414.

Yudkowsky R, Park YS, Hyderi A, Bordage G. 2015. Characteristics and implications of diagnostic justification scores based on the new patient note format of the USMLE Step 2 CS exam. Acad Med. 90(11 Suppl):S56–S62. doi: 10.1097/ACM.0000000000000900.