Exploring differences in individual and group judgements in standard setting

Peter Yeates, Natalie Cope, Eva Luksaite, Andrew Hassell and Lisa Dikomitis

How absolutes vary

Exploring differences in individual and group judgements in standard setting

Authors:

Peter Yeates, Lecturer in Medical Education Research, Keele University School of Medicine / Consultant in Acute and Respiratory medicine, Pennine Acute Hospitals NHS Trust.

Natalie Cope, Lecturer in Clinical Education (Psychometrics), Keele University School of Medicine.

Eva Luksaite, Teaching Fellow in Sociology of Health, Keele University School of Medicine.

Andrew Hassell, Professor of Medical Education / Head of School, Keele University School of Medicine / Consultant Rheumatologist, the Haywood Hospital, Midlands Partnership Foundation NHS Trust.

Lisa Dikomitis, Senior Lecturer in Sociology of Health, Keele University School of Medicine / Research Institute for Primary Care and Health Sciences.

Communicating Author:

Peter Yeates

School of Medicine, David Weatherall Building, Keele University, Staffordshire
ST5 5BG.

Email: p.yeates@keele.ac.uk

Tel: +44 1782733930

Keywords

Evaluation / assessment of clinical performance

Testing / Assessment

Qualities/Skills/Values/Attitudes

Qualitative Research Methods

Abstract

Background:

Standard setting is critically important to assessment decisions in medical education. Recent research has demonstrated variations between medical schools in the standards set for shared items. Despite the centrality of judgement to criterion referenced standard setting methods, little is known about the individual or group processes which underpin them. This study aimed to explore the operation and interaction of these processes in order to illuminate potential sources of variability.

Methods: Using qualitative research, we purposively sampled across UK medical schools who set either a low, medium, or high standard on nationally-shared items, collecting data by observation of graduation-level standard setting meetings and semi-structured interviews with standard setting judges. Data were analysed using thematic analysis based on the principles of grounded theory.

Results: Standard setting occurred through the complex interaction of: institutional context; judges' individual perspectives; and group interactions. Schools' procedures, panel members and atmosphere produced unique contexts. Individual judges formed varied understandings of the clinical and technical features of each question, relating these to their differing (sometimes contradictory) conceptions of minimally-competent students, by balancing information and making suppositions. Conceptions of minimal competence variously comprised: limited attendance, limited knowledge, poor knowledge application, emotional responses to questions, "test-savviness", or a strategic focus on safety. Judges experienced tensions trying to situate these abstract conceptions in reality, revealing uncertainty.

Groups constructively revised scores through debate, sharing information, often constructing detailed clinical representations of cases. Groups frequently displayed conformity, illustrating a belief that outlying judges were likely to be incorrect. Less frequently judges resisted change, using emphatic language, bargaining or rarely "polarization" to influence colleagues.

Conclusions: Despite careful conduct through well-established procedures, standard setting is judgementally complex and involves uncertainty. Understanding whether or how these varied *processes* produce the previously observed variations in *outcomes* may offer routes to enhance equivalence of criterion-referenced standards.

299 words (max 300)

Background:

Standard setting procedures are critical to the validity of assessments(1,2) as they determine which candidates will pass or fail(3). Consequently, they have the potential to impact both educational fairness and patient safety(4) making it vital that their selection can be justified. Within medical education, criterion-based standards derived from group judgements on the difficulty of questions (for example Angoff or Ebel processes(5)) are recommended(6) or even required(7) for high-stakes assessments to avoid the potential for standards to be compromised by variations in the abilities of student cohorts(8). Recent research, however, has shown substantial variations in the standards set for graduation-level knowledge testing items shared between different UK medical schools(9). These variations were sufficient to potentially result in important differences in the level of knowledge required to graduate at different medical schools, and as such they challenge the validity of standard setting judgements. Despite the centrality of both individual and group decisions within criterion-based standard setting processes, little is known about how such judgements are made.

An extensive scholarship on standard setting is available. Several studies have considered whether the choice of standard setting methods (i.e., Angoff, Ebel, Hofstee, Cohen) influences the resulting cut-score(10). Whilst differences have occurred between standard setting methods(11,12), the findings vary across studies suggesting that the context rather than the method *per se* may determine these effects.

Procedural modifications have shown more consistent influences. Meta-analysis has shown that group discussion was associated with increased cut scores whereas provision of “reality check” data (i.e. data from prior performance of items) was associated with decreased cut scores(13). Clauser and colleagues(14) found that judges’ opinions became more consistent but no more predictive of students’ actual performances when they discussed questions, but that reality check data increased the predictive validity of their judgements. Consistently, Fowell et al(15) found that 10 judges were required prior to discussion to achieve a set threshold of reliability, falling to just 6 judges after discussion.

Very little research has focused on the behaviour or judgement/decision processes of standard setting panel members. Whilst variability in assessors’ judgement processes are well described in other contexts (16–19), we did not find any studies focused on individual standard setting judgements. Boursicot and Roberts(20) found that whilst judges could easily articulate good or very poor performance, they described uncertainty around borderline or minimally competent performance and sometimes disagreed strongly over whether students should pass or fail if they didn’t know applied knowledge relating to a particular item. Focusing on group function, Duenk et al(21) investigated “polarization” when an individual who privately disagrees with an evolving consensus moves their judgement away from the group average in order to pull the resulting average towards their favoured position. They described this phenomenon on 17% of observed occasions of standard setting(21).

A diverse literature has considered how groups function in occupational settings(22). Time pressure, group size, perceptions of individuals’ status and members’ degree of engagement all have a bearing on the processes and outcomes of group functioning(23), whilst failures to identify and use relevant information; or a tendency to behave irrationally due to social pressures add further complexity(24). Groups can resolve differences through conformity (where individuals change their position due to discomfort with being unusual); persuasion (where the presented information changes individuals thinking); or compromise (reaching an intermediate decision)(24). Group members who are unfamiliar expend effort on developing social roles and norms which reduces their productivity(25). Two recent reviews in medical education have considered these theories in relation to the function of clinical competence committees(26,27). Amidst a variety of potential implications, they both caution against "group-think" in which important information is ignored by members of a group in an attempt to maintain harmony.

In summary, whilst standard setting processes are vital to assessment decisions, recent work has highlighted important variation between medical schools in the standards set for graduation level items. These differences potentially threaten the validity of standard setting and warrant explanation. A diverse range of individual, interpersonal and institutional processes could influence the group-decisions on which criterion-based standard setting rely. Despite this, only very limited prior work has considered these issues. Understanding how both individual and group level processes operate within standard setting may illuminate the potential origins of variability in standard setting, and could in due course lead to an understanding which informs targeted interventions to enhance equivalence of the standards which result. This study therefore sought to explore the following questions:

1. What processes (thoughts, deliberations, interpersonal interactions) do individual judges and standard setting committees use whilst standard setting, and how do they interact?
2. How might differences in these processes give insight into the potential origins of variability in standard setting

Methods

Study Design

We used qualitative research using the principles of social constructivist grounded theory(28,29), collecting data through a combination of observation of standard setting meetings and semi-structured interviews with individual standard setting judges.

Population and Sampling

We used purposive maximum divergence sampling of UK medical schools, including two medical schools from those who set "low", "medium", and "high" cut scores in the 2016 common content exercise(9). These categories were determined by a third party statistician based on difference from the median. The common content exercise is a collaborative endeavour to share applied knowledge testing items between the 33 medical schools who comprise the UK's Medical Schools Council Assessment Alliance (MSC-AA). Within included schools, we sampled by observing a single graduation-level standard setting meeting and interviewing a convenience sample of individual standard setting judges, sampling until theoretical saturation occurred(30).

Recruitment procedures, consent and ethics

As common content data are not in the public domain, schools in each sampling category (low, medium, high) were approached and invited on our behalf by the MSC-AA. Details of schools who agreed to participate were passed to researchers. To assist neutrality, researchers were blind to whether schools had set a "high", "medium" or "low" standard in the common content exercise(9).

To avoid potential coercion, researchers asked all judges who were present at observed meetings to complete a form either providing written consent or declining to participate, enabling redaction of contributions by non-participating judges.

All standard setting judges were invited to participate in semi-structured interviews Participants signed a consent form before the interviews and again at the end to indicate ongoing consent to use their data. Ethical approval for the study was granted by Keele University Research Ethics Committee (Ref ERP2301).

Data collection

Two researchers from different disciplinary backgrounds (variously PY, educationalist, clinician; NC, educationalist, psychometrician; EL or EC, both anthropologists) unobtrusively observed, and audio recorded, each standard setting meeting. Researchers sat separately to each and either behind or to the side of the meeting, noting features which they considered pertinent to the research question and making field notes of their observations(29). During breaks, researchers asked the chair person or other judges about features of the meeting which weren't immediately clear from their vantage points – for example to understand what written information was provided to judges, or what information was being displayed on screens. In order to produce some commonality of focus, the researchers (of whom 3 are experienced ethnographers) jointly developed their observation approach through pilot observation and team discussions. Observation focused only on parts of meetings where single-best answer (SBA) questions were considered.

We chose individual semi-structured interviews(31)in order to enable individual judges to express their views freely without group dynamics. Researchers developed an interview topic guide from a sensitising literature review (exemplified within the background) about 1/.standard setting, 2/.individual and 3/.group judgement. Interviews initially focused on the judge's conceptions of "just-safe" or "borderline" students, exploring its specificity, origins and clarity. In order to enable situated demonstration of their reasoning, researchers then asked participants to standard set the same six single-best-answer (SBA) multiple choice questions, whilst describing their thoughts. To ensure a naturalistic approach, judges used their chosen standard setting method (Ebel or Angoff), which could resultantly vary between settings. The latter part of the interview focused on issues of group function and interaction. Researchers used the topic guide to create a conversation, exploring issues which arose and working with the participant to construct an understanding of the phenomenon. All interviews were audio recorded.

Audio data was redacted for sensitive content (i.e. substantive portions of multiple choice question items which might enable reconstruction of questions) before transcription by an external agency. Data were loaded into QSR NVivo 11(32) to aid analysis.

Data Analysis

All analysts began by reading and re-reading an initial data sample (one observation and two interviews), inductively labelling their insights and searching for concepts suggested by our sensitising literature review. Team members met to discuss coding and an initial coding framework was developed.

Three members of the team (PY, NC, EL) worked independently to open code the data corpus (observation and interview transcripts). PY performed secondary reading and analysis on a subset of data to enhance consistency. Analysis was iterative, alternating between coding and further data collection(30). Researchers used micro-analysis(34) to interpret meaning from the data and memo writing to capture emergent theoretical ideas(35). Further codes (with definitions) were developed as relevant processes were perceived, and shared amongst the researchers.

The research team met regularly to discuss progress, plan further data collection and discuss emerging theoretical concepts. Theoretical ideas were tested in existing and new data and refined. As analysis progressed, axial codes were developed to organise open codes and label emergent theoretical ideas(35). Findings were triangulated between interview and observation data and researchers' field notes. Through this process, constant comparison was used to test theoretical ideas, looking for consistency and divergence across participants and settings(34). Some theoretical ideas from early iterations were refuted whilst new theoretical ideas were interpreted. Comparative matrices were used to explore differences between schools and settings, guiding a search for similarities and differences. Data saturation was judged to have been achieved when the coding frame was sufficient to label all findings in the final school(30).

Results

The data corpus comprised over 265,000 words (approximately 40 hours of audio recording), including 6 observations of standard setting meetings and 18 individual interviews. All participants in observed meetings consented to inclusion of their data. As well as fulfilling our purposive sampling criterion, our sample also contained a diversity of: new and established medical schools; size of student cohorts; and geographical locations.

The standard setting process was complex, involving the interplay of three distinct themes: the standard setting context in the school (the atmosphere, procedures used and the panel composition); the variety of processes involved in forming individual judgements; the interactions which arose between judges within the standard setting groups.

Context: atmosphere, procedures and panel

Each school performed standard setting within the comparatively unique context which arose from the combination of the procedures they employed; the knowledge and beliefs of the panel members who were present; and the atmosphere within the meeting and school itself. Schools were evenly split between the Angoff and Ebel methods (three each), with no obvious relationship to the score which they gave. Despite using ostensibly the same standard setting processes (e.g. the Angoff method), schools varied substantially in the practices which they employed.

Meeting length differed substantially between schools: some were under two hours, whilst others took several days. In some schools judges individually rated all questions before the meeting, whilst in others they read them for the first time during the meeting. Some schools made judges' initial scores on questions visible to the other panel members, whilst others kept these private, referring instead to average scores or the spread of opinion. Schools differed in the timing of the standard setting meeting in relation to exams. Four schools standard-set before the exam, one during, and one after the exam. In the latter case, judges were able to view item-level data on students' performance from the current exam whilst judging questions. In other schools no such data were available, although some judges described the past performance of some repeated items. Schools selected questions for discussion differently; some discussed all questions, whilst others selected questions based on judges responses (usually exceeding a defined threshold of agreement), and used varied procedures to select judges to open discussions on each question. Judges in one school commented sequentially on whether they would like to alter their (visible) initial scores; others schools invited discussion from the most discrepant judges; whilst in one school judges could privately alter their judgements on laptop computers without comment.

Schools differed in the make-up of the panels: differences in group size (the smallest was 4 judges, the largest was 11 judges); the clinical speciality mix; the inclusion (and seniority) of trainees; and the presence of non-clinical academics. The gender balance of all panels was roughly equal. Most schools acknowledged difficulties in keeping consistency of group membership between successive meetings, with procedural specialities (e.g. cardiology and surgery) often less represented.

The atmosphere varied between meetings. Some were highly collegiate; others were more discursive and opinionated. Some groups were relatively light-hearted and conducive to discussion whilst others were more formal with limited interaction.

The combined influence of these different procedures and rituals produced a comparatively unique standard-setting culture within each school, which appeared, through format and expectations, to influence both the individual judgement process and the interactions of the panel during the meeting.

Forming Individual Judgements

Judges appeared to form individual or private initial judgements by relating their understanding of individual questions to their conceptions of minimally-competent students.

<u>Understanding the question:</u>

Using the approaches with which they were familiar, judges deliberated on specific clinical features of the items. Questions featuring common conditions; with important consequences; or featuring typical presentations tended to be perceived as easier (*the reason I thought it was quite easy was that the words that are there are textbook descriptions of the condition* School 2, Interview 2). Comparatively esoteric topics (e.g., rheumatology antibodies) were typically viewed as more difficult than foundational topics (e.g., basic ECG interpretation, immediate treatment of breathlessness, consent).

Judges' variably considered technical features of questions: information volume; the number of required inferential "steps"; the plausibility of the distractor options (i.e., the 4 incorrect answers which were available); and likelihood of construct-irrelevant heuristics.

Judges' own clinical knowledge appeared to importantly influence their perceptions of question difficulty:

> *so I think this is quite tricky, so I'd probably put it as moderate, but that probably reflects my lack of specialist knowledge in ophthalmology.*
>
> School 2, Interview 2.

Judges drew from their current practice and recollections of their knowledge at different stages of training. In schools which set a higher standard, the judges more frequently appeared to personally find the questions easy. Through appraisal of the questions' clinical content and technical features in the light of their own current and previous knowledge, judges developed somewhat different understandings of the questions.

Conceptions of minimally-competent students

Judges related their understanding of the question to their conceptions of minimally competent students. Judges used various terminologies to describe minimal-competence, including: "just safe", "just adequate", "borderline", "struggling", "not very good" or even simply "undergraduates". The language used to conceptualise minimal-competence varied between and within schools.

Judges articulated a variety of different conceptions of the essence of minimal-competence. Some described students who had had poor patterns of attendance:

> *You don't really see them and, you know, the reason they're struggling is because they don't attend.*
>
> School 2, Interview 2

Judges with this conception tended to view experience-based (rather than conceptual) topics as more difficult for minimally-competent students. Other judges articulated conceptions of students who possessed rote-learnt foundational knowledge without depth. Judges variously articulated that such students would rely on guess work for less core topics, or struggle to apply knowledge to practice. This might have manifested as difficulty recognising or interpreting cues in questions, making these questions more difficult.

Some judges' articulations of minimal competence described students with a strategic focus on safety: minimally competent students would know safety-related knowledge, but might lack more esoteric or theoretical knowledge:

> *I think where there are crucial patient safety issues, in relation to, say, maybe a management of emergencies, or a management of, you know, crucial drug issues, … recognition of red flag symptoms in histories, …That's a minimum level.*
>
> School 6, Interview 1.

Some judges conceived minimally competent students as impulsive or emotional (*I just thought they were going to absolutely freak and it would just be a, just best move on…* School 1, observation), tending to make questions requiring detailed processing or larger volumes of information seem more difficult. In other conceptions, judges' articulated notions of students who had poor exam technique, did not read questions thoroughly, or who missed key information in the questions.

Judges' articulations of minimally-competent students sometimes contradicted each other. For example: students who would struggle to reason through multiple inferential steps; versus savvy students with common sense (if not detailed knowledge). Judges often described uncertainty in their conceptions of minimally-competent students, and variously tried to situate these abstract notions in more concrete experience. This could draw from student teaching; performance data; or clinical exams. Judges varied in their access to and use of information which helped to situate their judgements in concrete experience. Several commented that they personally had not been minimally competent, making their own experiences of assessment difficult to interpret. Some judges described recalling minimally-competent peers at medical school. A judge from one school which set a high standard commented that they rarely personally encountered minimally-competent students, potentially suggesting that judges' perceptions of minimal competence could be influenced by typical performance within their student cohort. Collectively, these various articulations suggested that judges conceptualised minimal-competence in a variety of different ways.

Relating the question to minimally-competent students

Judges appeared to relate their individual understandings of the question to their individual conceptions of minimal-competence in order to produce judgements on particular questions. Doing this involved two additional processes: *balancing information* and *supposition*.

Judges balanced information from different sources: some which influenced their understanding of the question; and some which situated their conceptions of minimally-competent students. These could have contradictory influences on their evolving judgements:

> *I'm thinking this is easy because that's what I see every day. So I kind of have to separate myself from that and think about then what's a student who isn't doing that every day going to think or going to... so I know that they find death certificates difficult, but it's really important and it's something that, at [name of university], we emphasise a lot and they get taught a lot on it. They've also not yet though done their assistantship…*
>
> School 6, Interview 2.

Judges frequently made suppositions whilst trying to determine how minimally-competent students would think whilst answering a particular question.

> *So I thought the borderline candidate would look at the word "curtain" and decide TIA, in fact I thought it was going to end up being quite hard for them.*
>
> School 4, observation.

> *I think that poorer students would word match and go for the otosclerosis and assume that because he's got recurrent left sided otorea that that's related, even though it's not..*
>
> School 5, Interview 3.

Suppositions were sometimes highly specific, and included emotional responses students might have to questions or pitfalls they might encounter. Relating the question to minimally-competent students was complicated by a tension which many judges acknowledged: a difference between what they desired that a minimally competent student *should* know and what (in reality) they actually *would* know. Many judges knew that the Angoff procedure asks judges to focus on what students "would" know, but nonetheless slipped into focusing on "should" thereby tending to see questions as easier and set harder standards as a result. They also struggled at times to reconcile a tension between the clinical importance or urgency of a particular treatment in a specific scenario with their perception that new doctors would be very unlikely to manage such scenarios without supervision. Similarly, tension arose from an awareness of differences between guideline based care and experiential learning.

By relating their understandings of the question to their conceptions of minimally competent students, judges produced individual judgements on the items. Sometimes these judgements varied substantially, for example different judges gave the same question individual Angoff estimates of 25% to 80%, whereas Ebel processes showed individual judges offering judgements of question difficulty and relevance in maximally different categories.

Group Processes

Standard setting meetings occurred within the context created by the atmosphere within the meeting, the standard setting procedures which were employed, and the composition of panel members. Chairpersons influenced group functioning in different ways: sometimes drawing the group's attention to situating information; arbitrating discussions, encouraging or legitimising minority perspectives; timekeeping and sometimes using humour to set the tone. As a result they importantly contributed to the atmosphere of the group, and each had their own style.

Changing perspective through discussion and debate

Individual judges shared their varied perspectives of individual questions. Some groups extensively debated the implications of clinical features of questions, in order to construct a shared understanding of the clinical material. Sometimes this occurred with only limited reference to minimally-competent students. The extent of this type of discussion varied between schools.

As well as revealing judges' different understandings of the questions, group discussions revealed judges' different (sometimes contradictory) conceptions of minimal-competence. Similar to individual judgements, judges within the groups often made suppositions as they illustrated their perspectives, and collectively balanced information from several perceptions:

> *Judge 1: I think the things that are included in the stem, push you quite firmly towards a vascular cause and there is only one...*
>
> *Judge 2: because both those conditions are quite common … I just think that the distractors are quite good for this potentially...*
>
> *Judge 3: I think the borderline candidate would see severe pain after eating, and could well think ulcer...*
>
> *Judge 1: Yes, that was kind of where I was going…*
>
> *Judge 4: I am not sure the distracters are great, you know … senior would be coming, and this is something that they are taught in Year 4 … So I don't think it's an easy question for a just passing undergrad, but nor do I think it is a hard question either.*
>
> *Judge 5: I have been the most hawkish. I think this is just the sort of thing an F1 is going to get called to, and depending on the level of support, may have to make a certain decision.*
>
> School 4, observation, successive comments in relation to the same question

Many episodes of discussion were constructive, and caused judges to reflect and revise their judgements. Judges described being reassured that they had reached a better judgement, particularly in areas where they felt relative uncertainty.

Conforming with colleagues' perspectives

In contrast to this constructive reflective process, groups also showed a moderate tendency to conform. Groups usually appeared to assume that outlying judges were less likely to be correct than those near to the group average. Outliers often seemed uncomfortable in their position and moved their judgements towards the centre of the group:

> *you sort of want to fit in with the trend … unless you're really quite sure that you're, confident that you're right, I think you'd probably end up nudging towards the mean*
>
> School 1, Interview 1.

> *if I've got a strong view I'll say it I suppose my mind set is if I am an outlier the onus is on me to say why I'm not going to shift more to the mean*
>
> School 3, Interview 2.

Conforming often appeared to occur in response to uncertainty. Occasionally, judges altered their scores without significant discussion, apparently simply because they were outliers. Whilst normative tendencies were fairly ubiquitous, their extent varied between schools. One school's procedures required the group to reach a predefined level of agreement on each question, and correspondingly showed the strongest tendency to conformity.

<u>Strategies of influence:</u>

Less frequently, instances arose where judges disagreed more strongly despite discussion, and (in contrast to conforming) seemed reluctant to settle for a compromise

> *there may be… times where you think you want to try and convince someone about why you feel the way you do about it,*
>
> School 1, Interview 1.

Judges sometimes appeared to feel determined to either convince their colleagues of their position or to minimise any compromise to their stance. In some instances judges may have defended their choice to avoid an inferred sense of personal criticism. Overt displays of hostility, anger or authority were absent. One junior doctor described a sense that it could be difficult to assert their opinion with more senior colleagues (*you do, as a junior, you do feel a bit… hmm, inferior's the wrong word. Um.. a bit less bold.* School 1, Interview 1). Instead, judges used a variety of strategies to influence their colleagues. Sometimes emphatic language was used to emphasise their point:

> *Judge 1: You've got to really think about it and picture it in your mind to get there. And it's anatomy. It's been four years… no-one does this on clinical placement*
>
> *Judge 2: If they say anything other than [mumbles], then they need shooting.*
>
> School 1, observation.

Rarely a judge who was at odds with the group (who might therefore be expected to conform) appeared to bargain, offering a partial re-alignment of their score (*I would be happy to move to a 6, but definitely not 5.* School 4, observation). We observed a single instance of "polarization" (see background) in which a judge adjusted their score *away* from the group mean (rather than towards it) in response to a disagreement. This tended to move the final group average towards their position. Whilst these strategies were only used in a minority of instances, they appeared to influence the group's final judgement on these questions.

<u>Final judgements</u>

Through the combination of discussion and debate leading to changes in perspectives, frequent conformity, and the less frequent use of strategies of influence, groups produced final judgements for each question, and a final standard for the exam.

Discussion:

Summary of results:

Standard setting occurred in the comparatively unique contexts which arose from each school's choice of procedures, selection of panel members and the atmosphere within the meeting. Judges formed individual judgements by developing varied understandings of the clinical and technical features of each question, and relating these (using the processes of 'balancing information' and 'supposition') to their varied conceptions of minimally-competent students. Differing (and sometimes contradictory) conceptions of minimal competence variously described limited attendance, limited knowledge, difficulty with application, emotional or irrational responses to questions, "test-savviness" or a focus on practical safety without detailed theoretical knowledge. Tensions arose for judges when trying to situate their conceptions of minimal competence which revealed their uncertainty in these concepts.

Standard setting groups debated questions, shared information, and often constructed detailed representations of clinical material. Groups also used supposition and balancing information and constructively revised scores through discussion. Groups often displayed a tendency to conformity, illustrating a belief that outlying judges were likely to be incorrect. Outliers frequently changed their scores in response to this normative influence, sometimes with very limited discussion. By contrast, judges less frequently resisted change, using emphatic language, bargaining or very rarely using polarization to influence colleagues. Whilst exhibiting combinations of these varied processes, groups arrived at a final judgement for each question and, as a result, a standard for the exam

Theoretical consideration of findings

Our study has provided the first in-depth description of the ways that institutional context, individual judgements and group interactions operate within standard setting for knowledge testing in medical education. Whilst these variations in *process* illuminate many potential avenues for further exploration, it remains to be determined whether they are responsible for the observed variations in the *outcomes* of standard setting (i.e. variable standards(9)), or whether other unknown factors either partly or wholly determine this variability. Equally, it is important to acknowledge that since no gold standard method exists for standard setting(3,5), and the optimal approach remains a matter of debate(36) different medical schools have very legitimately interpreted standard setting methods in their own ways to arrive at the varied procedures described. For an individual school, a priority in standard setting would be to arrive at a process which is familiar to participants, produces consistent standards year on year, and whose standards assure faculty of competent graduates.

These caveats notwithstanding, it may be informative to speculate how the variations in context, judgement and interaction we have described could potentially operate to cause variations in the standard which a school sets. Individual judges varied substantially in their judgements for individual questions, and so the final score for each item might potentially have been influenced by the mix of judges who were present. This influence, if confirmed, could suggest that differences between schools in the standards which are set may not be fixed, but might arise from within-school variability due to panel composition.

Schools had well-developed, but different, procedures and rituals within their meetings, which can be considered to a kind of organisational memory(37,38), and may in turn arise from the wider culture within the school. These different contexts are likely to represent fairly fixed differences between schools, which could potentially influence standard setting. The influence of some procedural differences (for example the presence or absence of reality-check data) can be theorised from prior empirical research(13) whilst others (e.g. the way judges are selected to speak) will require further investigation to understand whether they have a significant bearing on outcomes or not. Nonetheless the combination of different procedures which each school employs appears to contribute to a relatively static context which could potentially influence standards.

A judge at one of the high-standard schools described their cohort of students and junior doctors as more capable than is typical in other schools. By inference, they suggested that this might situate their conceptions of minimal competence around a comparatively high ability, thereby influencing the standard they set. Whilst resonating with other domains of assessment(39), it is beyond the limitations of our method to determine whether this suggestion is true; further research is needed to investigate this possibility.

Judges varied conceptions of minimal-competence may be important, principally because they are likely to emanate from differing experiences of *different* students who are minimally-competent for varied reasons. The complexity which arises from such legitimately-different perspectives may be best managed using “distributed cognition”(40,41) in which different members of a team process separate aspects of a complex decision and then share information to reach a decision. This might involve deliberately acknowledging and legitimising different judges’ varied conceptions, encouraging consideration of their implications for specific items, and then weighing these different positions to reach a collective decision.

Judges used supposition whilst considering students’ thinking. . Despite this, judges commented that they lacked experience of directly observing minimally-competent students whilst they answer knowledge testing questions, making these suppositions potentially difficult to justify. Perhaps surprisingly, there is an almost complete lack of empirical study of the decision processes of minimally competent students whilst answering multiple-choice question.

Collectively our findings illustrate that there is significant judgemental complexity in the procedures used to establish criterion-referenced standards: judges are sometimes uncertain; their conceptions of minimal competence vary, sometimes contradict and may reflect a plurality to the phenomenon of minimal competence; they make suppositions; and there is suggestion that their judgements may be influenced by normative comparisons.

Suggestions for practice:

Whilst further empirical study is required before making recommendations, our findings suggest some avenues which could be explored as potential means to enhance the equivalence of standard setting between schools. Exploration could include harmonisation of panel size and composition (i.e. the mix of specialities) between schools; and sharing and use of national performance data on common content items in an attempt to situate judgements in a common frame of reference. Chairpersons might be encouraged to further notice and legitimise different conceptions of just-safe students, and encourage the group to explore how these might explain differences on some questions. Whether this might reduce the normative inclination which groups sometimes exhibited could be explored.

No objective means currently exist to determine criterion-referenced standards. Whilst it remains conceptually plausible that such standards may exist, our findings illustrate the judgemental complexity of the procedures used to determine them. As a result it may be reasonable for academic debate to consider pragmatic means of supporting or scaffolding the judgements of panels. Such debate might consider greater use of performance data, statistical equating procedures, or even potentially combining criterion and norm referenced methods in some form of hybrid procedure.

Strengths and Limitations:

The study was broadly sampled and triangulated across both observations of standard setting meetings and participants interviews with standard setting judges. The analysis has been conducted rigorously and in depth. Despite these strengths, the study has a number of limitations. The sampling approach observed a single standard setting meeting in each school and so it is not possible for us to comment on how stable our observations may have been over time. Participants were aware that they were being observed, so despite our assurances of confidentiality and non-judgemental exploration, some Hawthorne effect (or the tendency for performance to be altered through observation(42)) could have occurred. We recruited a convenience sample of judges in each school. Whilst this doesn't influence the credibility of the range of perceptions they described, we cannot exclude the potential that other judges held different perceptions. Our sampling reached theoretical saturation(30) with respect to the processes in our model, but, despite this further procedural variations may occur outside of the schools which we sampled.

Future research:

Future empirical study should seek to describe the variety of approaches which minimally-competent students employ when answering multiple-choice questions, and how these differ from the strategies of more capable students. Experimental study could compare the influence of some of the procedural variations we have described on group functioning, or seek to determine their effect on standard setting outcomes. Further research might explore the merits of our suggestions for practically enhancing standard setting.

Conclusions:

Standard setting procedures produce fixed criteria through a complex interplay of the institutional context, judges' varied individual perceptions and group interactions. Whilst variations between schools in standard setting for shared items could potentially arise from a variety of these, or other, processes, further work is needed to establish any such influence on standard setting outcomes, or to understand their implications for enhancing equivalence.

Word count: 5704

Acknowledgements:

We would like to thank Efun Coker for assisting with initial data collection and analysis; the Medical Schools council Assessment Alliance for commissioning the study and assisting with sampling and recruitment; each of the six participating medical schools for allowing us to observe their standard setting; and to all of the participating standard setting judges.

Funding:

This work was commissioned and funded by a grant from the Medical Schools Council Assessment Alliance (MSC-AA). The MSC-AA board reviewed and commented on the manuscript prior to submission.

Peter Yeates is an NIHR Clinician Scientist Award holder and receives funding for his ongoing research. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflicts of Interest:

None declared

Author contributions:

PY led and substantially contributed to the design of the study, planning, development, data collection, analysis and manuscript drafting.

NC substantially contributed to planning, development, data collection, analysis and contributed to manuscript drafting.

EL substantially contributed to data collection, analysis, and contributed to manuscript drafting.

AH substantially contributed to the design of the study, and contributed to the analysis and manuscript drafting.

LD substantially contributed to planning, analysis and manuscript drafting.

References:

1. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7.

2. Kane MT. Validating the Interpretations and Uses of Test Scores. J Educ Meas. 2013;50(1):1–73.

3. Cizek GJ, editor. Setting Performance Standards Foundations, Methods and Innovations. 2nd ed. New York and London: Routledge; 2012.

4. Epstein RM, Hundert EM. Defining and Assessing Professional Competence. JAMA. 2002;287(2):226–35.

5. Downing SM, Tekian A, Yudkowsky R. Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education Procedures for Establishing Defensible Absolute Passing Scores on Performan. Teach Learn Med. 2006;18(1):50–7.

6. Wass V, Vleuten C Van Der, Shatzer J, Jones R. Medical education quartet Assessment of clinical competence. Lancet. 2001;357:945–9.

7. General Medical Council. Assessment in undergraduate medical education [Internet]. 2009. Available from: http://www.gmc-uk.org/Assessment_in_undergraduate_medical_education___guidance_under_review_0815.pdf_56439668.pdf

8. Fowell SL, Bligh JG. Recent developments in assessing medical students. Postgrad Med J. 1998;74(867):18–24.

9. Taylor CA, Gurnell M, Melville CR, Kluth DC, Johnson N, Wass V. Variation in passing standards for graduation-level knowledge items at UK medical schools. Med Educ. 2017;51(6):612–20.

10. Cusimano MD. Standard setting in medical education. Acad Med. 1996;71(10):S112-20.

11. Kaufman DM, Mann K V, Muijtjens AMM, der Vleuten CPM. A comparison of standard setting procedures for an OSCE in undergraduate medical education. Acad Med. 2000;75:267–71.

12. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, Van Der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Med Educ. 2003;37(2):132–9.

13. Hurtz GM, Auerbach M a. A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus. Educ Psychol Meas. 2003;63(4):584–601.

14. Clauser BE, Mee J, Baldwin SG, Margolis MJ, Dillon GF. Judges’ Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. J Educ Meas. 2009;46(4):390–407.

15. Fowell SL, Fewtrell R, McLaughlin PJ. Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? Adv Heal Sci Educ. 2008;13(1):11–24.

16. Gingerich A, Ramlo SE, van der Vleuten CPM, Eva KW, Regehr G. Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. Adv Heal Sci Educ. 2017; 22(4):819–38.

17. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. Med Educ. 2016;50(5):511–22.

18. Lee V, Brain K, Martin J. Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications. Acad Med. 2017;92(6):880–7.

19. Eva KW. Cognitive Influences on Complex Performance Assessment: Lessons from the Interplay between Medicine and Psychology. J Appl Res Mem Cogn. Society for Applied Research in Memory and Cognition; 2018;7(2):177–88.

20. Boursicot K, Roberts T. Setting Standards in a Professional Higher Education Course : Defining the Concept of the Minimally Competent Student in Performance- Based Assessment at the. High Educ Q. 2006;60:74–90.

21. Deunk MI, van Kuijk MF, Bosker RJ. The Effect of Small Group Discussion on Cutoff Scores During Standard Setting. Appl Meas Educ. 2014;27(2):77–97.

22. Ilgen DR, Hollenbeck JR, Johnson M, Jundt D. Teams in Organizations: From Input-Process-Output Models to IMOI Models. Annu Rev Psychol. 2005;56(1):517–43.

23. Levine JM, Moreland R l. Progress in small group research. Annu Rev Psychol. 1990;41(1):585.

24. Kerr NL, Tindale RSS. Group performance and decision making. Annu Rev Psychol. 2004;55:623–55.

25. Guzzo RA, Dickson MW. Teams in organizations: Recent research on performance and effectiveness. Annu Rev Psychol. 1996;47:307–38.

26. Hauer K, ten Cate O, Boscardin C, Iobst W, Holmboe E, Chesluk B, et al. Ensuring Resident Competence: A Narrative Review of the Literature on Group Decision Making to Inform the Work of Clinical Competency Committees. J Gr Med Educ. 2016;8(2):156–64.

27. Chahine S, Cristancho S, Padgett J, Lingard L. How do small groups make decisions? Perspect Med Educ. 2017 Jun 22;6(3):192–8.

28. Watling CJ, Lingard L. Grounded theory in medical education research: AMEE Guide No. 70. Med Teach. 2012;34(10):850–61.

29. Kennedy TJT, Lingard LA. Making sense of grounded theory in medical education. Med Educ. 2006;40(2):101–8.

30. 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.

31. Kennedy S, Wilkerson L. Topics for discussion reported by students using clinical cases in a problem-based tutorials during a basic science course. Acad Med. 1993;68(10):S31–3.

32. QSR. NVivo qualitative data analysis Software. 2014.

33. Miles MB, Huberman AM, Saldaña J. Fundamental fo Qualitative Data Analysis. 3rd ed. Qualitative Data Analysis: A Methods Sourcebook. Sage Publications Ltd; 2014.

34. Strauss A, Corbin J. Basics of Qualitative Research Techniques and Procedures for Developing Grounded Theory. 2nd ed. Thousand Oaks, California: Sage; 1998.

35. Charmaz K. Constructing grounded theory: a practical guide through qualitative analysis. Book. 2006. p208.

36. Homer M, Darling JC. Setting standards in knowledge assessments: Comparing Ebel and Cohen via Rasch. Med Teach. 2016;38(12):1267–77.

37. Edmondson AC, Bohmer RM, Pisano GP. Disrupted Routines: Team Learning and New Technology Implementation in Hospitals. Adm Sci Q. 2001;46(4):685.

38. Argote L, Guo JM. Routines and transactive memory systems: Creating, coordinating, retaining, and transferring knowledge in organizations. Res Organ Behav. 2016;36:65–84.

39. Yeates P, Cardell J, Byrne G, Eva KW. Relatively speaking: contrast effects influence assessors' scores and narrative feedback. Med Educ. 2015;49:909–19.

40. Mathieu JE, Heffner TS, Goodwin GF, Salas E, Cannon-Bowers JA. The influence of shared mental models on team process and performance. JApplPsychol. 2000;85(2):273–83.

41. Durning SJ, Artino AR. Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide no. 52. Med Teach. 2011;33(3):188–99.

42. Adair JG. The Hawthorne effect: A reconsideration of the methodological artifact. J Appl Psychol. 1984;69(2):334–45.