Effect of Exposure to Good vs Poor Medical Trainee Performance on Attending Physician Ratings of Subsequent Performances

Peter Yeates, MBBS, MClinEd
Paul O'Neill, MBChB, MD
Karen Mann, PhD
Kevin W. Eva, PhD

Context Competency-based models of education require assessments to be based on individuals' capacity to perform, yet the nature of human judgment may fundamentally limit the extent to which such assessment is accurately possible.

Objective To determine whether recent observations of the Mini Clinical Evaluation Exercise (Mini-CEX) performance of postgraduate year 1 physicians influence raters' scores of subsequent performances, consistent with either anchoring bias (scores biased similar to previous experience) or contrast bias (scores biased away from previous experience).

Design, Setting, and Participants Internet-based randomized, blinded experiment using videos of Mini-CEX assessments of postgraduate year 1 trainees interviewing new internal medicine patients. Participants were 41 attending physicians from England and Wales experienced with the Mini-CEX, with 20 watching and scoring 3 good trainee performances and 21 watching and scoring 3 poor performances. All then watched and scored the same 3 borderline video performances. The study was completed between July and November 2011.

Main Outcome Measures The primary outcome was scores assigned to the borderline videos, using a 6-point Likert scale (anchors included: 1, well below expectations; 3, borderline; 6, well above expectations). Associations were tested in a multivariable analysis that included participants' sex, years of practice, and the stringency index (within-group *z* score of initial 3 ratings).

Results The mean rating scores assigned by physicians who viewed borderline video performances following exposure to good performances was 2.7 (95% CI, 2.4-3.0) vs 3.4 (95% CI, 3.1-3.7) following exposure to poor performances (difference of 0.67 [95% CI, 0.28-1.07]; $P=.001$). Borderline videos were categorized as consistent with failing scores in 33 of 60 assessments (55%) in those exposed to good performances and in 15 of 63 assessments (24%) in those exposed to poor performances ($P<.001$). They were categorized as consistent with passing scores in 5 of 60 assessments (8.3%) in those exposed to good performances compared with 25 of 63 assessments (39.5%) in those exposed to poor performances ($P<.001$). Sex and years of attending practice were not associated with scores. The priming condition (good vs poor performances) and the stringency index jointly accounted for 45% of the observed variation in raters' scores for the borderline videos ($P<.001$).

Conclusion In an experimental setting, attending physicians exposed to videos of good medical trainee performances rated subsequent borderline performances lower than those who had been exposed to poor performances, consistent with a contrast bias.

JAMA. 2012;308(21):2226-2232 www.jama.com

Author Affiliations: Schools of Translational Medicine (Dr Yeates) and Medicine (Drs O'Neill and Mann), University of Manchester (Drs Yeates, O'Neill, and Mann), University Hospital of South Manchester (Drs Yeates and O'Neill), Manchester, United Kingdom; Division of Medical Education, Dalhousie University, Halifax, Nova Scotia, Canada (Dr Mann); and Department of Medicine, Centre for Health Education Scholarship, University of British Columbia, Vancouver, Canada (Dr Eva).
Corresponding Author: Peter Yeates, MBBS, MClinEd, University Hospital of South Manchester, Southmoor Road, Manchester, M23 9LT, United Kingdom (peter.yeates@manchester.ac.uk).

THE USEFULNESS OF PERFORmance assessments within medical education is limited by high interrater score variability,[1] which neither rater training[2] nor changes in scale format[3] have successfully ameliorated. Several factors may explain raters' score variability,[4] including a tendency of raters to make assessments by comparing against other recently viewed learners, rather than by using an absolute standard of competence.[5] This has the potential to result in biased judgments.

Competency-based models of medical education require that judgments be made against a fixed or absolute level of ability.[6] However, evidence from psychology and behavioral economics suggests that judgments tend to be relational. Two opposite effects are possible. The first is anchoring bias, in which recent experiences remain activated in observers' minds, causing them to pay undue attention to similar features in subsequent experiences.[7] As a consequence, they may offer judgments biased toward recently viewed anchor experiences, so that a recent experience of good performances may tend to increase subsequent scores compared with recently viewed poor performances. The second is contrast (or relativity) bias, in which judgments are in-

fluenced by the perceived relative rank of items in the immediate context[8]; thus, recent experience of good performances may tend to decrease subsequent scores compared with recent experience of poor performances.

Anchoring and contrast effects have been observed in many contexts, including perception of physical objects, interpersonal judgments, and various occupational roles,[9] including clinical medicine.[10] Both contrast and anchoring effects have been observed in performance assessment. Although mediating conditions have been investigated,[11] it is not clear which influence may predominate in medical education. Although studies have demonstrated sequence effects in medical education (ie, score increases over successive performances),[12] these studies cannot disentangle changes in trainee performance from bias in examiner judgments.

We therefore investigated whether raters' recent exposure to good vs poor performances influenced their scores of subsequent borderline performances, as well as the magnitude of any such bias relative to individual raters' tendencies to score stringently or leniently,[13] their sex,[14] and their practice experience.[14]

METHODS

Participants

The study population consisted of consultant physicians from England and Wales (comparable with US specialist attending physicians) working in specialties associated with internal medicine. We included emergency medicine physicians who trained via the examinations system of the Royal College of Physicians because they are frequently involved in the supervision of trainees providing care for patients with emergency presentations of internal medicine problems; this group is highly comparable in their assessments with internal medicine physicians.[15]

Additionally, participants must have worked as consultants in the United Kingdom for at least 2 years, estimated that they assess trainees using the Mini Clinical Evaluation Exercise (Mini-CEX) at least 5 times per year, and indicated that they feel comfortable assessing trainees on internal medicine case material. Mini-CEX assessments are direct observation assessments of trainees' clinical skills, in which a clinical performance is assessed and the trainee is given scores and feedback. Physicians were excluded if they had participated in previous studies by our group.

Recruitment

Recruitment was aimed at all UK consultant physicians. A standard e-mail invitation was sent by the national UK Foundation Programme to regional foundation directors, and then forwarded to foundation tutors in individual hospitals, and subsequently to individual consultants. In anticipation that e-mails may not have reached all areas, and to ensure geographical representativeness, follow-up direct e-mailing was used to increase recruitment in areas from which few participants had volunteered.

Interested and eligible individuals were invited to e-mail the research team, and were subsequently provided with further information, the study's web address, and a password. It is not possible to determine the proportion of UK consultant physicians who were eligible or who received e-mails. Researchers e-mailed 662 physicians directly, although the number of e-mails that were not received or went unread is unknown. Characteristics of the participants appear in TABLE 1.

Ethical approval was obtained from the Yorkshire and the Humber regional ethics committees. Consent was obtained online before beginning the study. Participants were not paid for their involvement in the study.

Design

The study used an Internet-administered experimental design, and randomized participants to 1 of 2 experimental groups. In the intervention phase, one group was primed by viewing 3 videos of good performances (ie, those that were competent for the stage of training, with some features of excellence) by postgraduate year 1 (PGY1) medical trainees. The other group was primed by viewing 3 videos of poor performances (ie, those that were below the standard of competence in multiple ways) by PGY1 medical trainees at the same level. In the comparison phase, both groups viewed 3 identical videos of borderline performances (ie, those that were marginally below the expected standard) in the same order.

To avoid confounding due to assessor × case variations, the same 3 cases were used for both groups, repeating between the intervention and comparison phases, but showing different performance levels. Within groups, all performances were viewed in the same order, and both groups saw the borderline performances in the same order. Participants were instructed to imagine that the trainees in the videos had requested a Mini-CEX assessment and that they were being asked to judge and score the performances accordingly.

Mini-CEX assessments are required for all PGY1 trainees nationally, so all supervising consultants should be familiar with them. Consultants are locally trained in conduct of the Mini-CEX. To preserve the ecological validity of our findings, participants did not undergo additional training. Participants were blinded to the study's premise; they were informed the study would investigate "an aspect of the way assessors make decisions." Participants scored performances consecutively using the UK Foundation Programme standard Mini-CEX scoring format (described below). Demographic data were collected after all cases had been viewed and rated. The study was completed between July and November 2011.

Validation of Videos

The study videos featured scripted performances of PGY1 trainees interviewing simulated patients; the performances did not represent the actual skills of the featured trainees. Videos were based on 3 clinical cases: pleuritic chest pain in a 54-year-old woman (case 1); unexplained loss of consciousness in a 34-year-old man (case 2); sus-

pected upper gastrointestinal bleeding in a 44-year-old man (case 3).

Three different PGY1 trainees (A, B, and C) were featured in the videos, with each working up each case only once. Owing to video content, the good performance–primed group saw all 3 trainees (A, B, and C) in the priming videos; whereas the poor performance–primed group saw trainee C twice and trainee A once during the priming phase. Borderline videos (seen by both groups) featured trainee B for case 1 (pleuritic chest pain) and case 3 (upper gastrointestinal bleeding) and trainee A for case 2 (transient loss of consciousness). Consequently, the PGY1 trainee featured in the borderline video for case 2 was known to both priming groups, whereas the PGY1 trainee in the borderline video for cases 1 and 3 was known only to the good performance–primed group (TABLE 2). To enable participants to view multiple videos without placing undue strain on concentration or memory, previously used videos were shortened to approximately 4 minutes each. Each focused on the history of the presenting complaint and sections that involved explaining likely investigations or provisional diagnoses.

Scripts and videos were developed and validated for a previous study[5] and demonstrated distinct levels of performance. For this study, we assessed whether they still ranked appropriately following editing by having a 6-member expert panel rank order the videotaped performances within each case. Cases 1 and 2 showed complete agreement with the intended order; for case 3, 5 of 6 experts ordered videos as intended, and 1 expert reversed the order of the borderline and poor performances.

Table 1. Participant Demographics[a]

	Overall (N = 41)[b]	Performance-Primed Group		*P* Value
		Good (n = 20)	Poor (n = 22)	
Sex				
Male	28 (68)	13 (65)	16 (71)	.66
Female	13 (32)	7 (35)	6 (29)	
Duration of consultancy, mean (SD), y	10 (7.4)	13 (7.5)	8 (6.5)	.03
Expert raters (>7 y of rating experience)	26 (63)	15 (75)	11 (52)	.13
Specialty				
Acute medicine	4 (9.7)	2 (10.0)	2 (9.5)	.96
Cardiology	4 (9.7)	4 (20.0)	0	.03
Clinical pharmacology	1 (2.4)	0	1 (4.8)	.32
Dermatology	1 (2.4)	1 (5.0)	0	.30
Endocrinology (diabetes mellitus)	3 (7.3)	0	3 (14.3)	.08
Emergency medicine	3 (7.3)	2 (10.0)	1 (4.8)	.52
Internal medicine (general)	2 (4.9)	2 (10.0)	0	.14
Geriatric medicine	9 (22.0)	3 (15.0)	6 (28.6)	.29
Gastroenterology	5 (12.2)	2 (10.0)	3 (14.3)	.68
Palliative medicine	1 (2.4)	1 (5.0)	0	.30
Rehabilitation medicine	1 (2.4)	0	1 (4.8)	.32
Respiratory medicine	5 (12.2)	3 (15.0)	2 (9.5)	.59
Rheumatology	1 (2.4)	0	1 (4.8)	.32
Stroke medicine	1 (2.4)	0	1 (4.8)	.32

[a] Data are presented as number (percentage) unless otherwise indicated.
[b] One participant in the poor performance–primed group was excluded from the primary analysis due to a website problem (did not receive intervention).

Outcome Assessment

Scores were collected using the standard format of the UK Foundation Programme.[16] This required assigning scores to 7 different domains: history

Table 2. Performance Scores by Priming Group

	PGY1 Trainee in Video	Performance-Primed Group, Mean (95% CI)		Between-Group Difference (95% CI)	*t* Statistic (Degrees of Freedom)	*P* Value	Corrected *P* Value[a]
		Good	Poor				
Good performance							
Pleuritic chest pain	A	4.6 (4.3 to 5.0)					
Transient loss of consciousness	B	4.5 (4.2 to 4.7)					
Upper gastrointestinal bleeding	C	4.1 (3.8 to 4.4)					
Poor performance							
Pleuritic chest pain	C		2.1 (1.7 to 2.4)				
Transient loss of consciousness	C		1.8 (1.5 to 2.2)				
Upper gastrointestinal bleeding	A		2.3 (1.8 to 2.7)				
Borderline performance							
Pleuritic chest pain	B	2.9 (2.5 to 3.3)	3.2 (2.9 to 3.6)	0.30 (−0.22 to 0.84)	1.18 (39)	.25	.74
Transient loss of consciousness	A	3.1 (2.8 to 3.6)	4.1 (3.8 to 4.4)	1.00 (0.59 to 1.42)	4.90 (39)	<.001	<.001
Upper gastrointestinal bleeding	B	2.2 (1.8 to 2.5)	2.9 (2.6 to 3.2)	0.70 (0.24 to 1.17)	3.09 (39)	.004	.01

Abbreviation: PGY1, postgraduate year 1.
[a] Bonferroni correction applied for these 3 comparisons.

taking, physical examination, communication skills, critical judgment, professionalism, organization/efficiency, and overall clinical care. Scores are given on a 6-point Likert scale with the options of 1, well below expectations for foundation PGY1 trainee completion; 2, below expectations; 3, borderline; 4, meets expectations for PGY1 trainee completion; 5, above expectations; and 6, well above. There was also an option for unable to comment.

Statistical Analysis

Based on the distribution of trainees' scores in prior research,[17] we considered a difference in scores between groups of 0.5 on the assessments' 6-point scale to be the minimum meaningful difference. A power calculation based on pilot data indicated that a difference in scores of 0.5 between groups could be detected at 80% power with approximately 30 participants. On this basis, we set an a priori recruitment target of 40 participants.

Participants' scores were averaged across the 7 domains to give a single score for each video. Missing data (including unable to comment responses) were excluded from the denominator in the average, so that resulting scores were the average of the available scores. The primary question was addressed through a mixed-design analysis of variance with the dependent variable being the scores participants assigned to each of the 3 borderline performances (case being the repeated measure) and a between-participant factor of the experimental group.

Although Likert items produce ordinal data, the combination of multiple items produces interval data[18] that can be validly analyzed by parametric means.[19,20] Effect sizes were calculated using the Cohen d statistic (large $\geq$0.8, medium $\geq$0.5-$<$0.8, small $\geq$0.1-$<$0.5). The frequencies with which participants categorized performances as consistent with failure (defined as score $<$3) or passing (defined as score $\geq$4) were compared between the 2 priming groups using χ^2 tests.

The clinical case $\times$ group interaction (groups primed with good vs poor performance) was examined to determine if group differences were uniform across cases. Pairwise comparisons were made using *t* tests with Bonferroni correction applied for 3 comparisons; this approach adjusted the *P* values so that the criterion for significance remained at a 2-sided *P* value of .05.[21] Analysis of covariance was used to repeat the analysis with duration of consultancy as a covariate to check for any confounding due to this variable. The frequencies of expert assessors (defined as $>$7 years experience[22]) was determined in each group.

We determined how consistently raters scored stringently (ie, scores lower than participants' mean) or leniently (ie, scores higher than participants' mean) by calculating an intraclass correlation (class 2, accuracy) within each group. This analysis used participants' scores across all 6 observed videos, and treated raters as the facet of differentiation (ie, the numerator in the equation). The resulting coefficient describes the consistency with which raters could be differentiated based on the scores they assigned. We calculated a within-group *z* score (stringency index) for each participant based on their mean ratings assigned to the first 3 videos during the intervention phase. This gave a measure of how comparatively high or low each participant scored the intervention phase performances relative to their group.

Multiple linear regression was used to assess potential moderating factors. The dependent variable was the mean of the scores assigned to the 3 borderline videos by each participant. Independent variables were duration of consultancy (continuous), sex (categorical), the stringency index (continuous), and priming condition (good vs poor; categorical). All 4 independent variables underwent forced entry into the regression.

Statistically significant variables were subsequently entered hierarchically to determine if they offered incremental explanatory power. Because the concept of stringency has been previously described, it was entered first and then the novel variable of priming condition was entered last. Interactions between significant variables were then examined using univariate analysis of variance. All statistical analyses were conducted using SPSS software version 15 (SPSS Inc).

RESULTS

Participants

There are approximately 8500 physicians in the United Kingdom. Eighty individuals agreed to take part by the study's close; 45 were randomized by the time recruitment exceeded the a priori target of 40 participants, and 41 completed the study (FIGURE). Of study participants, 32% were women compared with 31% of secondary care consultants in the United Kingdom (Table 1).[23]

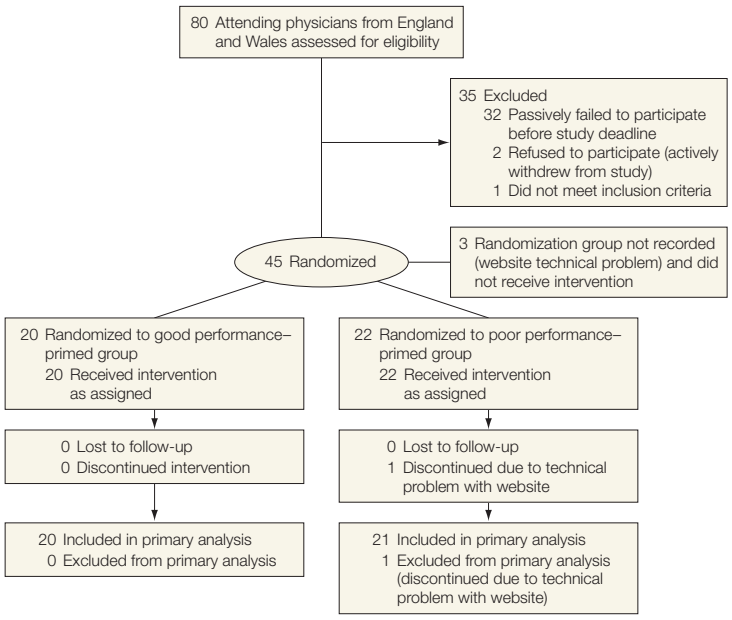
Participants were drawn from 12 of the 15 English and Welsh postgraduate training deaneries (geographical training regions),[24] and included 14 of 18 specialties. Consequently, the sample was broadly representative of the target population. Missing scores accounted for 11% of possible responses in the good performance–primed group and 8% in the poor performance–primed group. Participants in the good performance–primed group had a higher mean duration of consultancy than those in the poor performance–primed group (13 years [95% CI, 10-16 years] vs 8 years [95% CI, 5-11 years]; *P*=.03).

The good performance–primed group contained more cardiologists than the poor performance–primed group (4 vs 0; *P*=.03). The main analysis was rerun with cardiologists excluded, and there was no alteration in the findings. Other differences between groups were not significant.

Comparison of Borderline Videos by Intervention Group

The mean scores assigned to each of the 3 borderline videos were higher for participants who were primed with poor performances compared with those who were primed with good performances

Figure. Randomization Process of UK Attending Physicians in Rating Study



(3.4 [95% CI, 3.1 to 3.7] vs 2.7 [95% CI, 2.4 to 3.0], respectively; difference of 0.67 [95% CI, 0.28 to 1.07]; $P=.001$; Table 2). The Cohen d statistic of 0.63 indicated a moderate effect size due to experimental manipulation. The case × group interaction also was statistically significant ($P=.01$), indicating that the magnitude of the between-group differences varied across cases (case 1, 0.3 [95% CI, −0.22 to 0.84]; case 2, 1.0 [95% CI, 0.59 to 1.42]; case 3, 0.7 [95% CI, 0.24 to 1.17]; Table 1).

Borderline performances were categorized as consistent with failing scores in 33 of 60 assessments (55%) in those exposed to good performances and in 15 of 63 assessments (24%) in those exposed to poor performances ($P<.001$). They were categorized as consistent with passing scores in 5 of 60 assessments (8.3%) in those exposed to good performances compared with 25 of 63 assessments (39.5%) in those exposed to poor performances ($P<.001$).

Post hoc pairwise comparisons showed that group differences were significant for cases 2 and 3, but not for case 1 (TABLE 3). The nonsignificant effect seen in case 1 was in the same direction as the other 2 cases. In the analyses of covariance, there was no significant interaction with duration of consultancy ($P=.38$), and adjusting for this variable altered the main finding only minimally. There was no significant interaction between experimental condition and rater expertise as defined by 7 years or more of consultant experience ($P=.65$).

Participants showed a moderately consistent tendency to be either stringent or lenient compared with the other raters in their group. The intraclass correlation was 0.51 for participants primed with good performances and 0.58 for participants primed with poor performances.

Multivariable Analysis

Neither participants' sex nor their duration of consultancy showed a statistically significant relationship with the mean scores they gave to the borderline performances. The priming condition (good vs poor performances) and the stringency index showed significant relationships, jointly accounting for 45% of the observed variation in raters' scores for the borderline performances ($P<.001$) (Table 3). In the hierarchical regression model, raters' stringency index explained 18% of the observed variance, whereas adding experimental group explained a further 24% of the observed score variance. There was no significant interaction between stringency index and study group by univariate analysis of variance ($P=.46$). Consequently, the influence of recent experience was not different for stringent or lenient raters.

COMMENT

In this study, recently viewed medical trainee performance videos influenced raters' scoring of subsequent performances. The findings were consistent with a contrast effect in that viewing good performances resulted in lower borderline performance scores relative to viewing poor performances. These findings support the notion that recent experience biases raters' performance assessments, and suggest that such biases are not due to anchoring.

The size of this effect in clinical terms is important. In a study by Mitchell et al,[17] the mean Mini-CEX score for UK PGY1 and 2 trainees was 3.91 (SD, 0.38). Had the mean effect we observed (0.67 scale points) occurred in that group of trainees, it would have accounted for a change of 1.76 SDs. Case 2 would have been ranked near the bottom of the cohort on the basis of scores given by participants primed with good performances, but above the middle of the cohort on the basis of scores given by participants primed with poor performances.

Participants in this study were primed with performances that were either consistently good or poor. It is unclear whether an effect would occur with less consistent performance across candidates (ie, mixed performances). Perhaps a single good or poor performance could bias scoring on the subsequent candidate, or conversely, it may be that raters assimilate all recent performances and compare against their average level.

In sequential examination formats, such as objective structured clinical ex-

Table 3. Regression Model

	Forced-Entry Multivariate Analysis			Hierarchical Multivariate Analysis		
	β Coefficient (95% CI)	*P* Value	r^2	β Coefficient (95% CI)	*P* Value	Change in r^2
Sex	0.06 (−0.32 to 0.43)	.77				
Duration of consultancy	0.02 (−0.01 to 0.05)	.14				
Stringency index	0.38 (0.17 to 0.60)	.001		0.36 (0.14 to 0.57)[a]	.002	0.18
Good or poor performance priming	0.78 (0.40 to 1.15)	<.001		0.67 (0.32 to 1.02)[b]	<.001	0.24
Overall		<.001	0.45			

[a] First step in hierarchical regression (stringency index).
[b] Second step in hierarchical regression (good or poor performance priming added to stringency index).

aminations, candidates may follow good or poor colleagues sequentially through a series of clinical stations.[25] If a contrast effect were to occur after a single extreme candidate (good or poor), then the following candidate might receive biased scores that indicate as much about the ability of the preceding candidate as the current examinee.

Furthermore, because trainees often work in consistent pairings or small groups, consultants may consistently compare a given trainee against the same individuals when conducting workplace-based assessments. They may consistently contrast a trainee against either a very good or very poor peer, potentially biasing the judgments of a trainee and consequently the educational feedback provided, thereby having important implications for individual trainees' development.

Raters who have more than 7 years of preceptor experience have more complex assessment-related knowledge structures than more inexperienced raters.[22] Thus it is surprising, given that participants had on average been consultants for approximately 10 years, that viewing just 3 performances might be enough to induce the observed effect and that the effect was unrelated to duration of consultancy. This suggests that despite considerable experience, raters may still not possess well-developed fixed criteria against which to judge observed performance.[26] In other contexts such as clinical reasoning, experts have been shown to be just as susceptible to heuristics as novices,[27] which further supports this finding.

The study findings need to be considered in the context of its limitations. The videos depicted patient interviews within internal medicine and we studied only the judgments of consultant physicians from England and Wales. Consequently, findings may not generalize beyond this setting, although we have no reason to believe populations would differ with respect to the outcomes we measured. Despite demonstrating comparability with the general population, we cannot exclude some respondent bias in that participants who chose to take part in our study may have been more interested in education or assessment than their nonparticipating colleagues. However, because participants were randomized to intervention study groups, this would not have influenced the study's internal validity. It is possible that individuals who are less enthusiastic about education may possess a less developed understanding of the assessment process and thus might have shown an even greater effect.

The study demonstrated the influence of recent experience on borderline performances; whether this influence occurs when evaluating better or worse levels of performance (ie, beyond the borderline cases studied herein) will require further investigation. The observed contrast effect was restricted to cases 2 and 3; no significant difference was seen for case 1, although the direction of the observed effect was consistent. The reason for this is unclear, although context specificity is a well-established phenomenon in medical education.[28]

The observed effect occurred whether or not groups of participants had previous exposure to the trainees in the borderline videos, which supports the robustness of the effect and offers an area for further research aimed at exploring any mediating influence of the individual being observed.

The apparent contrast bias accounted for 24% of the observed score variance in addition to raters' tendency to be consistently stringent or lenient. Raters were only moderately consistent in their stringency or leniency despite common intuition that some examiners are harder graders than others. Neither duration of consultancy nor evaluator sex had any relationship to score variation. Thus, further study into the sources of raters' variability is required.

We recommend that our study be repeated in other contexts, particularly with reference to whether this effect occurs across the spectrum of performance quality, for other examination formats, and for other groups of learners and raters. The mediating role of case specificity and mixed performance (recent performances of differing quality) also should be investigated.

CONCLUSIONS

With the movement toward competency-based models of education, assessment has largely shifted to a system that relies on judgments of performance compared with a fixed standard at which competence is achieved (criterion referencing).[6] Although this makes conceptual sense (with its in-

herent ability to reassure both the profession and the public that an acceptable standard has been reached), the findings in this study, which are consistent with contrast bias, suggest that raters may not be capable of reliably judging in this way.

Author Contributions: Dr Yeates had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Yeates, O'Neill, Mann, Eva.
Acquisition of data: Yeates.
Analysis and interpretation of data: Yeates, O'Neill, Mann, Eva.
Drafting of the manuscript: Yeates.
Critical revision of the manuscript for important intellectual content: Yeates, O'Neill, Mann, Eva.
Statistical analysis: Yeates.
Administrative, technical, or material support: Yeates.
Study supervision: O'Neill, Mann, Eva.
Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Yeates reported receiving reimbursement for travel costs from Medtronic Inc to attend a meeting to discuss future collaboration on an unrelated research project. Dr Mann reported receiving honoraria for her role as chair of the objectives committee for the Medical Council of Canada; receiving grant income from the Social Sciences and Research Council of Canada (coinvestigator) for a project on technology in a new curriculum and from the Society for Academic CME for a project on a study of facilitated reflection; and receiving reimbursement for travel costs (without additional honoraria) for invited guest speaker engagements at Durham University, UK, in July 2012, Utrecht Medical Centre in March 2012, and at University College London, UK, in June 2012. No other authors reported conflicts of interest.
Funding/Support: Dr Yeates received a traveling fellowship award from the Association for the Study of Medical Education that was used in support of this study.
Role of the Sponsor: The Association for the Study of Medical Education had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the preparation, review, or approval of the manuscript.
Additional Contributions: We thank Julie Morris, MSc (University of Manchester, University Hospital of South Manchester), for providing statistical advice to the project. Ms Morris did not receive payment for her work. We also thank our expert review panel for their assistance in video validation, the physicians at the UK Foundation Programme, and simulated patients featured within the videos, the foundation and their associated regional schools for their assistance with recruitment, and the participants who took part in the study. Physicians at the foundation received a small gratuity for their time and simulated patients received professional rates of pay.

REFERENCES

1. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Acad Med*. 2010;85(9):1453-1461.
2. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74-79.
3. Donato AA, Pangaro L, Smith C, et al. Evaluation of a novel assessment form for observing medical residents: a randomised, controlled trial. *Med Educ*. 2008; 42(12):1234-1242.
4. 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; 45(10):1048-1060.
5. 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 [published online May 12, 2012]. *Adv Health Sci Educ Theory Pract*. doi:10.1007/s10459-012-9372-1.
6. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001; 357(9260):945-949.
7. Chapman GB, Johnson EJ. Anchoring, activation, and the construction of values. *Organ Behav Hum Decis Process*. 1999;79(2):115-153.
8. Stewart N, Brown GD, Chater N. Absolute identification by relative judgment. *Psychol Rev*. 2005; 112(4):881-911.
9. Mussweiler T. Comparison processes in social judgment: mechanisms and consequences. *Psychol Rev*. 2003;110(3):472-489.
10. Riva P, Rusconi P, Montali L, Cherubini P. The influence of anchoring on pain judgment. *J Pain Symptom Manage*. 2011;42(2):265-277.
11. Greifeneder R, Bless H. The fate of activated information in impression formation: fluency of concept activation moderates the emergence of assimilation versus contrast. *Br J Soc Psychol*. 2010;49 (pt 2):405-414.
12. Ramineni C, Harik P, Margolis MJ, Clauser BE, Swanson DB, Dillon GF. Sequence effects in the United States Medical Licensing Examination (USMLE) step 2 clinical skills (cs) examination. *Acad Med*. 2007; 82(10)(suppl):S101-S104.
13. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ*. 2006;6(6):42.
14. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ*. 2008;42(4):364-373.
15. Ilgen JS, Bowen JL, Yarris LM, Fu R, Lowe RA, Eva K. Adjusting our lens: can developmental differences in diagnostic reasoning be harnessed to improve health professional and trainee assessment? *Acad Emerg Med*. 2011;18(suppl 2):S79-S86.
16. Norcini JJ. The Mini Clinical Evaluation Exercise. *Clin Teach*. 2005;2(1):25-30.
17. Mitchell C, Bhat S, Herbert A, Baker P. Workplace-based assessments of junior doctors: do scores predict training difficulties? *Med Educ*. 2011;45(12): 1190-1198.
18. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103(2684):677-680.
19. Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Rev Educ Res*. 1972;42(3):237-288.
20. Carifio J, Perla R. Resolving the 50-year debate around using and misusing Likert scales. *Med Educ*. 2008;42(12):1150-1152.
21. Wright SP. Adjusted P-values for simultaneous inference. *Biometrics*. 1992;48(4):1005-1013.
22. Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AMM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract*. 2011;16(2):151-165.
23. General Medical Council. List of registered medical practitioners: statistics. http://www.gmc-uk.org/doctors/register/search_stats.asp. Accessibility verified November 5, 2012.
24. National Health Service. Speciality training. https://www.mmc.nhs.uk/colleges_deaneries/deaneries.aspx. Accessibility verified November 5, 2012.
25. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ*. 2004;38(2):199-203.
26. Schwarz N. Self reports: how the questions shape the answers. *Am Psychol*. 1999;54(2):93-105.
27. Crowley RS, Legowski E, Medvedeva O, et al. Automated detection of heuristics and biases among pathologists in a computer-based system [published online May 23, 2012]. *Adv Health Sci Educ Theory Pract*. doi:10.1007/s10459-012-9374-z.
28. Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ*. 2005;39(1):98-106.