
[Document Abstract]

Context
A recent study has suggested that assessors judge performance comparatively rather than against fixed standards. Ratings assigned to borderline trainees were found to be biased by previously seen candidates’ performances. We extended that programme of investigation by examining these effects across a range of performance levels. Furthermore, we investigated whether confidence in the rating assigned predicts susceptibility to manipulation and whether prompting consideration of typical performance lessens the influence of recent experience.

Methods
Consultant doctors were randomised to groups within an internet experiment. The descending performance group judged videos of Foundation Year 1 (F1; postgraduate Year 1) doctors in descending order of proficiency; the ascending performance group judged the same videos in ascending order. For all videos, participants rated: (i) trainee competence; (ii) rater confidence and (iii) percentage better (the percentage of other F1 doctors who would perform better on the same task).

Results
Overall, the descending performance group assigned lower scores than the ascending performance group (2.97 [95% confidence interval 2.73–3.20] versus 3.50 [95% confidence interval 3.25–3.74]; \(F(1,47) = 9.80, p = 0.003, d = 0.52\)). Pairwise comparisons showed differences were significant for good and borderline performances. The percentage better ratings showed a similar pattern (descending performance mean = 57.4 [95% confidence interval 52.5–62.3], ascending performance mean = 43.4 [95% confidence interval 38.4–48.5]; \(F(1, 46) = 16.0, p < 0.001, d = 0.67\)). Confidence ratings did not vary by level of performance and showed no relationship with the effect of group.

Discussion
Assessors’ judgements showed contrast effects at both good and borderline performance levels. Findings suggest that assessors use normative rather than criterion-referenced decision making while judging, and that the norms referenced are weakly represented in memory and easily influenced. Confidence ratings suggested a lack of insight into this phenomenon. Raters’ judgements could be importantly influenced in ways that are unfair to candidates.