# **Title:**

# Exploring the Relationship between Examiners’ Memories for performances, domain separation and score variability.

**Short Title:**

# Remembering to differentiate

**Authors:**

Natasha Cleaton

Peter Yeates

Garth McCrae

Accepted by *Medical Teacher*, 5th April 2018.

DOI: 10.1080/0142159X.2018.1463088, available at: <https://doi.org/10.1080/0142159X.2018.1463088>

**Name of Institution:**

Keele University School of Medicine

**Corresponding details:**

Peter Yeates, School of Medicine, David Weatherall Building, Keele University, University Road, Staffordshire ST4 6QG.

Tel: +441782733930

Fax: n/a number

Email: p.yeates@keele.ac.uk

**Practice points**

* Examiners’ vary considerably in their ability to recollect students’ performances.
* Weak relationships exist between examiners’ recollection accuracy and their ability to accurately score performances or separate different domains of performance.
* Examiners’ episodic memories may be a useful focus for future research aiming to better understand the processing and integration phases of judgement.

**Notes on contributors**

NC is a Year 4 medical student at Keele University School of Medicine. She is interested in assessment of Medical Students competence and practical methods to improve the reliability of this process.

PY is a lecturer in medical education research at Keele University School of Medicine, and a consultant in Acute and Respiratory Medicine at Pennine Acute Hospitals NHS Trust. He is interested in assessor cognition and innovative ways of improving the dependability of OSCEs.

GM is a medical statistician at the Institute for Primary Care and Health Sciences, Keele University, and a fellow of the Royal Statistical Society. His main interests are the measurement of child development, psychometric modelling of judgement data, and creation of short-form PROMS.

**Glossary**

Episodic memory: A component of long-term memory that that enables human beings to remember past experiences. Storing information in episodic memory involves attention, encoding, and processing and then later retrieval to relive the memories. A variety of processes can interfere with these processes, so episodic memory is incompletely accurate (Tulving, 2002).

**Abstract (max 200 words)**

Background: OSCE examiners’ scores are variable and may discriminate domains of performance poorly. Examiners must hold their observations of OSCE performances in “episodic memory” until performances end. We investigated whether examiners vary in their recollection of performances; and whether this relates to their score variability or ability to separate disparate performance domains.

Methods: Secondary analysis was performed on data where examiners had: 1/scored videos of OSCE performances showing disparate student ability in different domains; and 2/performed a measure of recollection for an OSCE performance. We calculated measures of “overall-score variance” (the degree individual examiners’ overall scores varied from the group mean) and “domain separation” (the degree to which examiners separated different performance domains). We related these variables to the measure of examiners’ recollection.

Results: examiners varied considerably in their recollection accuracy (recognition beyond chance -5% to +75% for different examiners). Examiners’ recollection accuracy was weakly inversely related to their overall score accuracy (R=-0.17,p<0.001) and related to their ability to separate domains of performance (R=0.25,p<0.001).

Conclusions: Examiners vary substantially in their memories for students’ performances which may offer a useful point of difference to study processing and integration phases of judgement. Findings could have implication for the utility of feedback.

**Background**

The General Medical council states that medical schools must ensure students are trained to be proficient, ethical and competent in their medical practice (General Medical Council 2017). On a more societal level, this is essential as clinical incompetence may place patients at heightened levels of risk (Cooke et al. 2006). Consequently, assessments to judge whether trainees can apply their knowledge and skills competently and ethically in a practice are critical to medical education.

A range of assessment types are used to assess medical students’ learning. Whilst written assessment formats are useful for assessing applied knowledge, meaningful assessment of practical skills requires observation and judgement of students’ performances. The most prevalent format for these judgements in high stakes assessments are “Objective Structured Clinical Examinations (OSCEs)”, which involve a qualified practitioner making an observational judgement on students’ performances whilst they demonstrate their skills in standardised, simulated clinical tasks such as communication skills, physical examinations, or practical procedures (Newble 2004). Whilst standardised marking schemes help to direct examiners’ focus, OSCEs remains subjective, relying on a judgements by examiners as to whether students are capable of providing safe, effective, clinical care (Gingerich et al. 2014). This subjectivity introduces the potential for substantial inter-examiner score variability, which is problematic for these types of assessment.

Prior research has investigated the magnitude of examiner variability in OSCEs. Brannick et al (2011) systematically reviewed and meta-analysed reports of OSCE reliability. They found that average OSCE reliability was just 0.66 (95% CI0.62-0.70). Within a communication-focused OSCE, Harasym et al (2008) demonstrated that variance due to examiner stringency (0.86) was more than four times greater than variance due to candidate ability (0.20). Even with two examiners per station, examiner variability still has the potential to influence pass or fail decisions for a minority of candidates (McManus et al. 2006). Collectively these studies suggest that examiner variability has a detrimental influence on the utility of OSCEs in high-stakes situations.

A further issue is apparent in assessors’ judgements. Several assessment types, including OSCEs, involve simultaneous assessment of different domains. A number of studies have shown that the scores awarded in one domain influence the scores awarded in another domain (Volkan et al. 2004; McKinley and Boulet 2005; Cook et al. 2010), suggesting that examiners struggle to differentiate between domains of performance. Attempts to address these issues (variability and poor domain discrimination) have generally either used examiner training (Cook et al. 2009; Pell, Homer, and Roberts 2008) or enhanced marking formats (Donato et al. 2008). Unfortunately only modest gains have generally been achieved. In response, a new field called “assessor cognition” has arisen to provide insight into assessors’ judgements processes (Gauthier et al. 2016; Lee et al. 2017).

A recent review of assessor cognition research (Gauthier et al. 2016) has defined the judgment process by three stages: Observation, Processing and Integration. In this model, observation is the most understood while there is much disagreement about the processing and integration stages. The observational stage depends upon ‘attention’; essentially only information that is attended to by the examiner during the observation can go through the processing stage. Processing gives meaning to the observed information which is then translated into a score in the integration phase. Recent research by Tavares’ et al(2014), has described the role of ‘attentional capacity’ within assessment judgements: the mental workload of judging performances is high and examiners experience cognitive overload. As a result, examiners have been shown to use “selective attention” during the observation phase of judgement, in order to manage the mental demands which this task imposes. The use of selective attention appears to limit the information which is available for the processing and integration stages of judgement which may consequently lead to inaccuracies in the judgement of performance.

Short term memory (comprising both sensory memory and working memory) is the only completely accurate part of the human memory system, but it decays after around 30 seconds (Baddeley, Eysenck, and Anderson 2015). As typical OSCE stations last between 5-10 minutes, we may deduce that examiners use an alternative memory system, episodic memory, to store their observations (Dickerson & Eichenbaum 2010). Episodic memory relies on a series of processes: attention, encoding, processing and retrieval (Tulving 2002). These processes are imperfect, meaning that observations may not be accurately recalled a few minutes later. Episodic memory is impacted by several factors: high mental workload can impair encoding, reducing recall (Baddeley 1992). Increasing delay between observation and recall reduces the completeness of recollections (Wang 2014) whilst several more specific processes (blocking, absentmindedness, misattribution, false memory and re-writing) (Schacter 2003) all interfere with the accuracy of recollections. Women have previously been found to perform better at episodic memory tasks then men (Herlitz et al. 1997). Any interruption of examiners’ episodic memories of performances could mean that the processing and integration phases of judgement may be based on incomplete information. As a result, the ability to store observations in episodic memory may be important to the judgement process.

Two separate effects might arise if examiners find it difficult to accurately recall their observations: firstly, examiners’ final judgements of the overall standard of the performance may be based on an incomplete, selected portion of observations. This may influence the scores that are chosen, contributing to score variability, and could go some way to explaining why assessors have been observed to adopt one of a small number of distinctly different perspectives of performances (Gingerich et al. 2017). As a result, examiners who recall performances comparatively poorly may tend to give less accurate (i.e. more variable) scores. Secondly, the ability to distinguish between domains of performance may depend on being able to accurately retain specific observations about performance in each domain. Consequently, examiners who have comparatively poor recall of performances may find it difficult to distinguish domains of performances, therefore showing less domain discrimination.

As a result, understanding the relationship between examiners’ memories for performance and their scoring tendencies may give insight into the little studied aspects of processing and integration of information within assessors’ judgements. In this initial exploratory research, we investigated whether:

1. Examiners vary in their ability to recollect students’ performance and which factors relate to this.
2. Examiners memory relates to overall score variability.
3. Examiners memory relates to their ability to separate domains of students’ performance.

**Methods:**

**Study Design**

This study used secondary analysis of data from a prior study (Yeates et al. 2017), which collected examiners’ scores for three standardised performances as well as a measure of the examiners’ recollections of one of those performances.

**Participants**

Participants in the original study were all: doctors registered to practice in the UK, who had: examined summative OSCE exams in the UK within the last two years; received training as an OSCE examiner; and were comfortable at assessing both knowledge and communication.

**Design of original study**

Participants (n=159) in the original study watched the same 3 sequential videos of OSCE performances. Each performance demonstrated a scripted scenario, in which the same young woman attended a General Practitioner to discuss a new diagnosis of type 1 Diabetes mellitus. Students undertaking the OSCE were required to demonstrate their knowledge of the condition and their empathy and communication skills. Performances were all scripted and validated as described in the original study. Performances varied in the standard of factual knowledge (K) and communication (C) skills which they displayed: one performance showed good knowledge, but poor communication skills (K+/C-); one showed poor knowledge but good communication skills (K-/C+) and one showed a mixture of good and poor instances of knowledge and communication (Mixed). As a result two videos showed divergent performance across different domains (K+/C-, and K-/C+) whereas one showed comparatively homogenous performance across domains (Mixed). Good communication involved features such as displaying empathy, using open questions, summarising and checking understanding; poor communication involved behaviours such as interrupting, being abrupt, lacking empathy, using closed questions. Good factual knowledge involved describing factually accurate information whilst poor factual knowledge involved describing factually inaccurate information. The standards of performance in each subdomain in each videoed performance were validated by a panel of expert educators as part of the original study.

Following each video, participants individually scored the performances and provided written feedback. After watching all three videos, all participants completed a distraction task (a lexical decision task (for an example see Sinclair and Kunda 1999)) lasting a few minutes and then immediately undertook a recollection test for the mixed performance. Following these processes, participants provided demographic details and debriefed.

While all of the participants watched the same three OSCE performances, the original study counter-balanced the presentation order of the videos. Half of participants watched the performances in the order: K+/C-, K-/C+, Mixed, and, as a result, had a short delay between viewing the mixed performance and performing the recollection task (around 5mins); whereas the other half of participants watched performances in the opposite order: Mixed, K-/C+, K+/C-, and consequently had a long delay between viewing the mixed performance and performing the recollection task (around 22 mins).

**OSCE performance scoring**

Performances were scored on a number of domains, with brief lists (5-6 items) of desirable behaviours listed for each domain. Participants scored each domain using end-anchored 7-point Likert scales ranging from “no items done” to “all items done well”. Two domains related to communication skills and two domains related to knowledge. Average scores for communication and knowledge domains were calculated for each participant. Additionally, participants scored the overall quality of the performance on a 7 point scale: fail (points 1 & 2); borderline (3); satisfactory (4); good (5); excellent (6&7). Participants could score the performances at any time whilst observing or immediately after each performance, until they moved to the next performance. Details of the score sheet are provided as an appendix to the original study (Yeates et al. 2017).

**Recollection Testing**

The recollection test probed participants’ memories of the Mixed performance i.e. the performance that displayed a mix of good and poor instances of performance in both the knowledge and communication domains. Participants were asked to read 40 statements which were ostensibly excerpts from the mixed performance. For each statement they were asked to indicate whether it was “True” (it was a direct quote from the mixed performance) or “False” (it was a plausible but invented statement which did not occur in the mixed (or any other) performance). Half of the statements were real (i.e. they should have been marked True) and half were invented (i.e. they should have been marked false) and within both real and invented statements an equal number related to factual knowledge and to communication.

**Calculation of variables**

To address the study research aims, we calculated three new variables from the existing data, prior to using them to test specific hypotheses.

Firstly we calculated a measure of participants’ “overall-score variance” for each performance. This measure expressed the distance (regardless of whether positive or negative) of each participant’s overall score from the mean overall-score for that performance. No objective means exists to determine the “correct” and therefore most accurate score for a performance. Instead the mean of a group of examiners is the best estimate of the notional true score (Streiner and Norman 2008, p167) which exists. Resultantly this measure was centred around the mean of examiners’’ judgements. As such this variable was a measure of the degree to which each participant’s overall score differed from the mean, with higher values indicating greater variation (akin to “inaccuracy”). To overcome problems of positive or negative sign we used a standard variance calculation procedure, namely:

Overall-score variance = √(𝑥𝑖−𝑥̅)2

Where:

𝑥𝑖 is the value of the overall-score for a given performance by a particular participant

𝑥̅ is the mean overall-score given to that performance by all examiners

Next we calculated the degree to which participants separated disparate domains of performance for each of the two divergent performances (K+/C-, and K-/C+). We called this the “domain separation score”. This was calculated by subtracting the communication score from the knowledge score for the K+/C- performance for each participant, and subtracting the knowledge score from the communication score for the K-/C+ performance for each participant. This gave a measure of the degree to which each participant separated disparate domains of performance for both K+/C- and K-/C+, with higher values indicating greater domain separation and lower values indicating less domain separation. We believe that this measure is a novel means of expressing the degree to which disparate domains of performance are separated by assessors.

Finally, we calculated a measure of the accuracy of participants’ recollections of the Mixed performance which we termed “recollection accuracy”. To do this we used individual participants’ scores from the recollection test. We defined recollection accuracy as: The proportion of real statements marked true minus the proportion of invented statements marked true. As such it had the potential to vary from -1.0 to +1.0, with a score of zero indicating recollection equal to chance.

 **Derivation of hypotheses and statistical analysis**

To address our first research aim, we examined the variable “recollection accuracy”. We compared whether recollection accuracy was influenced by the demographic variable of participant sex (male vs female) using univariate ANOVA, or related to their duration in years as an OSCE examiner, using Pearson’s correlation. Additionally, because of the theoretical likelihood that participants’ recollection accuracy may be influenced by a long versus short interval between observing the Mixed performance and completing the recollection test, we compared the influence of long versus short delay on participants’ recollection accuracy using univariate ANOVA.

To address research aim 2, we specified the following hypothesis:

**Hypothesis 1: Examiners’ recollection accuracy (i.e. memory) will show a negative relationship with their overall-score variance (i.e. their degree of inaccuracy / distance from the group median) on all performances.**

We tested this relationship using multiple regression analysis, with participants’ “overall-score variance” as the dependant variable, and the following independent predictor variables: “delay” (2 level factor; whether participants had a long delay or a short delay between observing the mixed performance and completing the recollection task), “performance” (3 level factor; whether the score related to the K+/C- performance, the K-C+ performance, or the mixed performance), “duration” (continuous variable; the number of years participants had examined OSCEs) and “recollection accuracy” (continuous variable, the measure of recollection as described earlier). The model entered all predictors concurrently.

Next, to address research aim 3, we specified the following hypothesis:

**Hypothesis 2: Examiners’ recollection accuracy (i.e. memory) will show a positive relationship with their domain separation score (i.e. the difference between the scores for knowledge and communication domain) on the K+/C- and K-/C+ performances.**

We tested this relationship using multiple regression analysis, with participants’ “domain-separation” score as the dependant variable, and the following independent predictor variables: “delay” (2 level factor; whether participants had a long delay or a short delay between observing the mixed performance and completing the recollection task), “performance” (2 level factor; whether the score related to the K+/C- performance or the K-C+ performance), “duration” (continuous variable; the number of years participants had examined OSCEs) and “recollection accuracy” (continuous variable, the measure of recollection as described earlier). The model entered all predictors concurrently.

Analyses pertaining to RQ1 (univariate ANOVA and Pearson’s correlations) were conducted using IBM SPSS Statistics for Windows, Version 21.0; multiple regression analyses relating to hypothesis 1 & 2 were performed using R “stats” package (Development Core Team, 2017) and variable relative importance statistics were calculated using the R package “relaimpo” (Grömping, 2006). Pearson’s r values were obtained from relative R values by calculating their square roots.

**Ethics**

Ethical approval for this secondary data analysis was granted by the Keele School of Medicine ‘School Student Project Ethics Committee’.

**Results**

**Descriptive data:**

Data from all 159 participants in the original study were included in the analysis. The measure of recollection accuracy was normally distributed, with a mean of 0.40, and a standard deviation of 0.15. This variable had the potential to range from -1.0 to 1.0, with a value of 0.0 being equivalent to random chance. Individual examiners ranged in their recollection accuracy values from -0.05 to 0.75.

The measures of the overall score variance variables for each of the 3 performances all appeared skewed. Their medians and inter-quartile ranges were K+/C- median=0.84, IQR=1.0; K-C+ median=0.65, IQR=1.0; Mixed median=0.81, IQR=0.62.. The measures of domain separation were normally distributed. Their means and standard deviations were: K+/C- mean=3.12, sd=1.46; and K-/C+ mean=2.65, sd=1.39. The order in which examiners viewed the performances influenced their recollection accuracy: short delay 0.45 (95% CIs 0.41 – 0.48); long delay 0.35 (0.32 - 0.38, F(1,157) = 16.3 P<0.001). Examiners recollection accuracy did not differ based on their sex (men= 0.38 (95% CIs 0.34 – 0.42); women 0.41 (0.38-0.44, *F*=1.40, p=0.24). Examiners’ recollection accuracy was negatively correlated with the number of years they had examined OSCEs (Pearson’s r=-0.22, p=0.005).

**Hypothesis 1: Examiners’ recognition accuracy will show a negative relationship with their overall score variance on all performances.**

Regression data for hypothesis 1 is displayed in table 1. When variables are jointly modelled, there was no statistically significant relationship between performance order (long or short delay) and overall score variance. The overall score variance differed for the three different performances (K+/C-β=0.00; K-/C+ β=0.17, p=0.02; Mixed β-0.14, p=0.046). The number of years participants had examined OSCE exams was not related to overall score variance. Consistent with hypothesis 1, examiners’ recollection accuracy showed a negative, statistically highly significant, relationship with overall score accuracy (β=-0.007, R=-0.17, p<0.001). The adjusted R-Squared for this model was 0.06 and relative proportions of the R-Squared can be found in table 1 for each variable. The relative R-Squared explained by recollection accuracy was 0.03.

**Hypothesis 2: Examiners’ recognition accuracy will show a positive relationship with their domain separation score on the K+/C- and K-/C+ performances.**

Regression data for this hypothesis is displayed in table 2. There was no statistically significant relationship between performance order (long or short delay) and domain separation. Domain separation differed for the two different performances on which this variable was calculated (K+/C-β=0.00; K-/C+ β=-0.52, p<0.001). The relationship between domain separation and the number of years participants had examined OSCE exams was close to statistical significance (β=-0.03, R=-0.13, p=0.07). Consistent with hypothesis 2, examiners’ recollection accuracy showed a positive, statistically highly significant, relationship with domains separation (β=0.02, R=0.25, p<0.001). The adjusted R-Squared for this model was 0.11 and relative proportions of the R-Squared can be found in table 1 for each variable. The relative R-Squared explained by recollection accuracy was 0.06.

**Discussion**

**Summary of findings:**

These findings show that experienced OSCE examiners’ varied in their ability to recognise exerts of a recently-viewed OSCE performance. This variation was substantial, ranging from recognition which was marginally worse than chance, to a corrected recognition rate of 75%. Overall, examiners’ memories for performances were imperfect, with an average recognition rate of 40% beyond chance. In addition, examiners’ varied in their overall accuracy (i.e. distance of their score from the group mean) and the degree to which they separated disparate domains of knowledge and communication.

Examiners’ abilities to recollect performances were (as expected) influenced by the duration of delay between the performance and the test of recollection: recollection accuracy was greater after a 5 minute delay than after a 22 minute delay. Examiners’ recollection accuracy reduced as the number of years they had examined OSCE increased.

Consistent with hypothesis 1, examiners’ overall-score variance was negatively related to examiners’ recollection accuracy, whilst, consistent with hypothesis 2, examiners’ domain separation scores were positively related to examiners’ recollection accuracy. In that respect, we found relationships which were consistent with both hypotheses. However, relatively little variance in the dependant variables was explained by recollection accuracy, and as such such its importance as a predictor of examiner behaviour (i.e., accuracy and domains separation) is not strong. That said, it is predictive relative to other factors examined in the two models.

**Theoretical Implications:**

These exploratory findings are based on secondary analysis of existing data, and as such any attempt to derive theory form them must be viewed as highly provisional. Nonetheless, in the event that they can be sufficiently replicated, they may suggest a number of interesting theoretical implications.

Substantial differences occurred between OSCE examiners in the accuracy of their recollections for a performance. This could potentially have arisen either due to differences in the way in which examiners paid attention whilst observing performances or their subsequent processing or integration of information. The potential for differences in attention resonate with prior work (Yeates et al. 2013; Gingerich et al. 2017). Either way, it suggests that differences in assessors’ memory for performances may constitute a point of difference between assessors, which may give a window into attention or processing differences.

Our hypotheses were supported, but only weakly. Had we found stronger relationships we might have been tempted to tentatively conclude that some assessors pay attention more fully, process more deeply and as a result: score performances more accurately; separate competencies more clearly; and recollect more fully. The existence of weak relationships leaves open the potential that this pattern may occur for some examiners, but with relationships of the magnitudes we observed this theory seems insufficient to fully explain our observations. At least some examiners must have scored accurately and separated domains well, but performed poorly on the recollection test. At least two alternative explanations of this observation may be suggested.

Firstly it is conceivable that some examiners were able to judge both the overall ability of performances and appropriately discriminate disparate domains of performances without the detailed processing of information which tends to result in its incorporation into episodic memory (Tulving 2002). This could have occurred if some examiners used a predominantly intuitive (or system 1-based) approach to judgement (Norman 2009). Notably this suggestion is somewhat at odds with the work of Tavares et al (2016) which suggested a more deliberative approach.

Alternatively some examiners could conceivable have processed information in detail, encoded it into episodic memory and used this whilst scoring, but then rapidly forgotten that information as they engaged in subsequent tasks. This could have occured due to retroactive interference, in which new information interferes with previously stored information (Bäuml 1996) and is consistent with doctors known ability to “task shift”( pausing a task to handle an incoming task)(Walter et al. 2014). Equally it might be that each of these processes occurred to a greater or lesser degree for different examiners.

**Practical implications:**

Whilst it is premature to make recommendations for practice until further work has been conducted, a few potential implications may be considered. The observation that assessors memories for performances declines with a fairly small increase in time (17 extra minutes), plus the observation of two further performances, has implications for assessment practice. This queries whether it is reasonable to allow examiners to amend scores or feedback as an OSCE progresses, and underscores the importance in less structured assessments (such as workplace based assessments) of ensuring that feedback is provided shortly after observation (Lefroy et al. 2015).

Feedback given following the assessment is an important part of the student learning process (Harrison et al. 2014). Studies have confirmed that students benefit from specific feedback, particularly looking at communication and professionalism (Junod Perron et al. 2016). Examiners’ episodic memories for performances may impact on the quality of the feedback provided by different examiners; those with less good recollection may potentially struggle to provide appropriately specific feedback. This may lead to unfair variation in feedback quality, and reduced learning.

**Limitations**

This exploratory study has a number of limitations. It used secondary analysis of data not specifically intended for this study. Data from the original study were pooled across groups because findings indicated that the groups were identical in all measures. Nonetheless, replication of these findings in a prospective study exploring these issues is required. The study tested its hypotheses by examining associative relationships. It is important to note that association does not indicate causation, and as such all theorising from these data is tentative.

In addition, participants only scored 3 videos and so we can only judge participants scoring on a limited amount of data. Participants only scored 2 domains of performance (knowledge and communication) whilst in real life scenarios various other domains such as practicable ability must also be assessed. Future research in this area will require a larger study involving more domains of assessment.

Participants’ episodic memories for performances were measured with a recognition-based test. Whilst this was useful in that it provided directly comparable responses, it only tested participants’ memories for the items which were posed; participants may have remembered numerous other features of the performances beyond that which we queried. As a result future research might use an open ended free-recall based method to see whether similar findings occur.

**Future Research**

Further research should seek to replicate these findings prospectively, across different stimuli, different marking sheets and with different groups of participants. In particular, experimental research might seek to manipulate examiners memory (potentially by deliberately incurring a delay between viewing and scoring performances, or by means of cognitive interference) to test whether examiners’ episodic memory is causally related to either score accuracy or domain discrimination.

As part of research to further understand the processing and integration phases of assessors’ judgements, research could examine the inter-relation of other factors known to influence memory such as transience, misattribution (Schacter 2003) or retroactive interference (Bäuml 1996). All of this will develop our theoretical understanding of the processing and integration phases of assessor judgement. Work to understand the implications of examiners’ episodic memory on their ability to recognise (or forget) important focal areas of weak performance and its implications for feedback will help to clarify practical implications.

**Conclusions**

Examiners vary substantially in how well they can recollect students’ performances, with all examiners showing imperfect recollections. Examiners ability to accurately score performances and differentiate disparate domains of students’ performances appears to be weakly related to their ability to recollect performances. These exploratory findings should stimulate the assessor cognition community to consider how differences in examiners’ episodic memories for performances influence the processing and integration phases of judgement.

**Acknowledgments section**

None

**Declarations of interest**

None declared

**References**

Baddeley A. 1992. Working Memory - the Interface between Memory and Cognition. J Cogn Neurosci. 4:281–288.

Baddeley A, Eysenck M, Anderson M. 2015. Memory. 2nd ed. London and New York: Taylor Francis.

Bäuml K. 1996. Revisiting an old issue: Retroactive interference as a function of the degree of original and interpolated learning. Psychon Bull Rev. 3:380–384.

Brannick MT, Erol-Korkmaz HT, Prewett M. 2011. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 45:1181–1189.

Cook D a, Beckman TJ, Mandrekar JN, Pankratz VS. 2010. Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability. Adv Health Sci Educ.; 15:633–45.

Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. 2009. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med. 24:74–9.

Cooke M, Irby DM, Sullivan W, Ludmerer KM. 2006. American medical education 100 years after the Flexner report. N Engl J Med. 355:1339–44.

Development Core Team, R. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Dickerson BC, Eichenbaum H. 2010. The episodic memory system: Neurocircuitry and disorders. Neuropsychopharmacology. 35:86–104.

Donato AA, Pangaro L, Smith C, Rencic J, Diaz Y, Mensinger J, Holmboe E. 2008. Evaluation of a novel assessment form for observing medical residents: a randomised, controlled trial. Med Educ.; 42:1234–42.

Gauthier G, St-Onge C, Tavares W. 2016. Rater cognition: review and integration of research findings. Med Educ. 50:511–522.

General Medical Council. 2017. Outcomes for Graduates [Internet]. Available from: https://www.gmc-uk.org/education/undergraduate/undergrad\_outcomes\_overarching.asp

Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. 2014. Seeing the “black box” differently: assessor cognition from three research perspectives. Med Educ. 48:1055–68.

Gingerich A, Ramlo SE, van der Vleuten CPM, Eva KW, Regehr G. 2017. Inter-rater variability as mutual disagreement: identifying raters’ divergent points of view. Adv Heal Sci Educ.22(4):819-838.

Grömping, U. (2006). Relative Importance for Linear Regression in R : The Package relaimpo. Journal of Statistical Software, 17(1).

Harasym PH, Woloschuk W, Cunning L. 2008. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ. 13 (5):617–32.

Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten C. 2014. Barriers to the uptake and use of feedback in the context of summative assessment. Adv Heal Sci Educ.:229–245.

Herlitz A, Nilsson LG, Bäckman L. 1997. Gender differences in episodic memory. Mem Cognit. 25:801–811.

Junod Perron N, Louis-Simonet M, Cerutti B, Pfarrwaller E, Sommer J, Nendaz M. 2016. The quality of feedback during formative OSCEs depends on the tutors’ profile. BMC Med Educ. 16:293. Available from: http://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-016-0815-x

Lee V, Brain K, Martin J. 2017. Factors Influencing Mini-CEX Rater Judgments and Their Practical Implications. Acad Med. 92:880–887.

Lefroy J, Watling C, Teunissen PW, Brand P. 2015. Guidelines: the do’s, don’ts and don’t knows of feedback for clinical education. Perspect Med Educ. 4:284–299.

McKinley DW, Boulet JR. 2005. Using factor analysis to evaluate checklist items. Acad Med. 80:S102-5.

McManus IC, Thompson M, Mollon J. 2006. Assessment of examiner leniency and stringency (’hawk-dove effect’) in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 6 (42). Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1569374&tool=pmcentrez&rendertype=abstract

Newble D. 2004. Techniques for measuring clinical competence : objective structured clinical examinations. Med Educ. 38:199–203.

Norman G. 2009. Dual processing and diagnostic errors. Adv Health Sci Educ Theory Pract. 14 Suppl 1:37–49.

Pell G, Homer MS, Roberts TE. 2008. Assessor training: its effects on criterion‐based assessment in a medical context. Int J Res Method Educ. Available from: http://www.tandfonline.com/doi/abs/10.1080/17437270802124525

Schacter DL, Chiao JY, Mitchell JP. 2003. The Seven Sins of Memory Implications for Self. Ann N Y Acad Sci. 7:226–239.

Sinclair L, Kunda Z. 1999. Reactions to a black professional: motivated inhibition and activation of conflicting stereotypes. J Pers Soc Psychol. 77:885–904.

Streiner D, Norman G. 2008. Health Measurement Scales. 4th ed. Oxford: Oxford University Press.

Tavares W, Eva KW. 2014. Impact of rating demands on rater-based assessments of clinical competence. Educ Prim care. 25:308–18.

Tavares W, Ginsburg S, Eva KW. 2016. Selecting and Simplifying: Rater Performance and Behavior When Considering Multiple Competencies. Teach Learn Med. 28:41–51.

Tulving E. 2002. Episodic Memory: From Mind to Brain. Annu Rev Psychol. 53:1–25.

Volkan K, Simon SR, Baker H, Todres ID. 2004. Psychometric structure of a comprehensive objective structured clinical examination: a factor analytic approach. Adv Health Sci Educ. 9:83–92.

Walter SR, Li L, Dunsmuir WTM, Westbrook JI. 2014. Managing competing demands through task-switching and multitasking: a multi-setting observational study of 200 clinicians over 1000 hours. BMJ Qual Saf. 23:231–241.

Wang B. 2014. Effect of time delay on recognition memory for pictures: The modulatory role of emotion. PLoS One. 9 (6): e100238.

Yeates P, O’Neill P, Mann K, Eva K. 2013. Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Heal Sci. 18:325–41.

Yeates P, Woolf K, Benbow E, Davies B, Boohan M, Eva K. 2017. A randomised trial of the influence of racial stereotype bias on examiners’ scores, feedback and recollections in undergraduate clinical exams. BMC Med. 15:179. Available from: http://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-017-0943-0

**Tables and Figures:**

**Table 1: Regression coefficients between overall score variance and listed predictor variables.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **β** | **St Err** | **Pearson’s r** | **Relative R2** | **t** | **p** |
| Intercept | 1.09 | 0.12 |  |  | 9.0 | <0.001\*\* |
| Delay (Short) | - | - | - | - | - | - |
| Delay (Long) | 0.04 | 0.06 | 0.06 | 0.004 | 0.66 | 0.51 |
| Performance (K+/C-) | - | - | - | - | - | - |
| Performance (K-C+) | 0.17 | 0.07 |  |  | 2.33 | 0.020\* |
| Performance (mixed) | -0.14 | 0.07 |  | 0.038 | -2.00 | 0.046\* |
| Duration | 0.004 | 0.006 | 0.05 | 0.003 | 0.64 | 0.52 |
| Recollection | -0.007 | 0.002 | -0.17 | 0.028 | -3.44 | <0.001\*\* |

\* = significant at p<0.05

\*\* = significant at p<0.001

**Table 2: Regression coefficients between domain separation and listed predictor variables.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **β** | **St Err** | **Pearson’s r** | **Relative R2** | **t** | **p** |
| Intercept | 2.61 | 0.308 |  |  | 8.5 | <0.001\*\* |
| Delay (Short) | - | - | - | - | - | - |
| Delay (Long) | -0.18 | 0.16 |  | 0.011  | -1.10 | 0.27 |
| Performance (K+/C-) | - | - | - | - | - | - |
| Performance (K-C+) | -0.52 | 0.15 |  | 0.032 | -3.37 | <0.001\*\* |
| Duration | -0.03 | 0.017 | -0.13 | 0.017  | -1.85 | 0.07 |
| Recollection | 0.02 | 0.005 | 0.25 | 0.062 | 4.12 | <0.001\*\* |

\* = significant at p<0.05

\*\* = significant at p<0.001