Full title: Comparatively salient: Examining the influence of preceding performances on assessors’ focus and interpretations in written assessment comments.

Running title: Comparatively salient: Examining the influence of preceding performances

Andrea Gingerich, Northern Medical Program, University of Northern British Columbia, Prince George, Canada

Edward Schokking, Northern Medical Program summer student research program, Prince George, Canada

Peter Yeates, Keele University School of Medicine, Keele, Staffordshire, UK and Pennine Acute Hospitals NHS Trust, Bury, Lancashire, UK.

Correspondence: Dr. Gingerich, Northern Medical Program, University of Northern British Columbia, 3333 University Way, Prince George, British Columbia, Canada, V2N 4Z9, 1-250-960-5432, [andrea.gingerich@unbc.ca](mailto:andrea.gingerich@unbc.ca)

Keywords

Assessment of clinical performance

Assessor cognition

Comparative distinctiveness

Contrast effects

Mini-CEX

Rater-based assessment

Rater cognition

Rater error

Rater bias

Salience

**Abstract**

Recent literature places more emphasis on assessment comments rather than relying solely on scores. Both are variable, however, emanating from assessment judgements. One established source of variability is “contrast effects”: scores are shifted away from the depicted level of performance in a preceding encounter. The shift could arise from an effect on the range-frequency of assessors’ internal scales or the salience of performance aspects within assessment judgments. As these suggest different potential interventions, we investigated assessors’ cognition by using the insight provided by “clusters of consensus” to determine whether any change in the salience of performance aspects was induced by contrast effects. A dataset from a previous experiment contained scores and comments for 3 encounters: 2 with significant contrast effects and 1 without. Clusters of consensus were identified using *F*-sort and latent partition analysis both when contrast effects were significant and non-significant. The proportion of assessors making similar comments only significantly differed when contrast effects were significant with assessors more frequently commenting on aspects that were dissimilar with the standard of competence demonstrated in the preceding performance. Rather than simply influencing range-frequency of assessors’ scales, preceding performances may affect salience of performance aspects through comparative distinctiveness: when juxtaposed with the context some aspects are more distinct and selectively draw attention. Research is needed to determine whether changes in salience indicate biased or improved assessment information. The potential should be explored to augment existing benchmarking procedures in assessor training by cueing assessors’ attention through observation of reference performances immediately prior to assessment.

**Introduction**

Rater-based assessments are used for formative and summative purposes and contribute to low and high stakes decisions. However, there is often excessive variability in scores and comments with assessors identified as a major source of unwanted variability. (Alves de Lima et al., 2007; Cook et al., 2010; Hill et al., 2009; Margolis et al., 2006; Weller et al., 2009) From a psychometric perspective, the variability attributed to assessors is not informative of the trainee’s competence and limits the reliability and validity of the assessments (Downing, 2004; Downing, 2005). Flawed assessment judgements tend to be offered as an explanation for problematic assessor variability (Gingerich et al., 2014a). If we were to better understand what assessors base their assessment judgements on and what factors can contribute to variability in their formation, new strategies could be designed to support assessors and ultimately improve the resulting scores and comments.

Since assessment judgements cannot be studied directly, researchers can gain insights into what assessment judgements are based on and what factors influence those assessment judgements by studying the scores and comments that assessors provide. Research indicates that the scores assessors assign and the valence of the comments they provide can be influenced by changes in the assessment context. One well studied technique for changing the assessment context is by having assessors provide scores and comments to a preceding performance before assessing the target performance. Both experimental and field studies have shown that when there is sufficient difference in the standard of competence between the two performances the scores and the valence of comments that assessors provide for the target performance can be significantly shifted away from the standard of competence in the preceding performance (Yeates et al., 2015a; Yeates et al., 2015b; Yeates et al., 2012; Yeates et al., 2013a). For example, assessors gave lower scores and more negative comments when a borderline performance was assessed immediately after assessing a good performance (Yeates et al., 2015a).

Although the phenomenon is well known as contrast effects, at this time, there is only theory informed speculation to guide what these results reveal in terms of how assessment judgements can be variably formed. It has been proposed that contrast effects arise due to our human tendency to make all judgements of magnitude in comparative terms (Parducci and Perrett, 1971). This theory suggests that human’s make judgements of magnitude by mentally constructing a scale based on available items, with judgments of magnitude being derived from a compromise between the ends of the scale (range) and the frequency of occurrence of different items within the scale (frequency). Importantly, humans appear to draw on immediate context as often (or perhaps more often) than long term memory when mentally constructing these “range-frequency” scales (Stewart et al., 2005) making them susceptible to contextual influences. This theory would suggest that preceding performances redefined assessors’ internal range-frequency scales of performance ability, by altering the frequency of different performance levels in the distribution, and thereby affected the valence of the judgement, resulting in altered scores and comments. In summary, this would tend to suggest that assessors assign scores based on an internal range-frequency scale that can be recalibrated due to contextual factors.

Other research has studied variable assessment comments to identify potential sources of variability in scores and assessment judgements. This research has found that written comments tend to be incomplete representations of the performance with groups of assessors commenting on different subsets of aspects for a given performance (Gingerich et al., 2014b; Tavares et al., 2016; Yeates et al., 2013b). Despite the possibility for aspects of a performance to be described in an unlimited number of combinations, comments tend to group into only a few shared interpretations of a performance that align with their assigned scores (Gingerich et al., 2017; Gingerich et al., 2014b). Both findings are consistent with past psychology research that has found overall evaluative impressions of others are based on combinations of interrelated aspects (Higgins, 2000) each differently organized around one (of a few possible) salient aspects to form a coherent interpretation (Mohr and Kenny, 2006; Park et al., 1994). By assuming that assessors write about the aspects of the performance they find most salient (Fiske and Cox, 1979), as in most noticeable, important and relevant for judging the trainee’s competence (Higgins, 1996), the content of the comments reveal the basis of assessors’ judgements. This line of reasoning raises the possibility that a substantial portion of assessor variability could potentially be explained by contextual influences on what assessors find salient within a performance. Those factors have yet to be identified.

As a result, these intersecting lines of research propose two quite distinct theoretical explanations of assessor variability that point in different directions to search for interventions. The ‘differing salience’ explanation suggests the need for interventions which may direct assessor’s focus or attention to pertinent aspects of the performance; the ‘range-frequency scale’ explanation suggests the need for interventions which can calibrate assessors’ internal scale to an appropriate frame of reference. Importantly, although prior contrast effect research (Yeates et al., 2015a) did not demonstrate an influence on the content of the comments—a finding that favours a range-frequency scale explanation rather than a differing salience explanation--the methods of Yeates et al’s (2015a) study only compared the balance of comments focused on communication and comments focused on factual knowledge. In continued investigation of influences on assessment judgements, this study will use the preceding performance as a known context change and determine if that change affects what aspects of the performance assessors find salient. In particular, we use contrast effects to alter the assessment context and examine the written comments for evidence that assessor’s attention was affected whilst observing target performances. Our research question is: Do assessors’ written comments show evidence of changes in perceptual salience when assessors’ judgments are influenced by contrast effects? These findings could further our understanding of how assessment judgements are influenced and inform interventions to address assessor variability.

**Methods**

**Overview.** We used existing data from an experimental study designed to investigate contrast effects (Yeates et al., 2015a) as the stimulus material for further data collection. The dataset contained scores and written comments collected in response to video-recorded clinical encounters that a) had been exposed to conditions that induced contrast effects resulting in significant differences; b) had been exposed to conditions that did not induce contrast effects, therefore producing no significant differences; c) had not been exposed to conditions intended to induce contrast effects (i.e. an “unbiased” or control condition). Three phases were used to analyse the data. Phase 1 examined the content of the written comments to determine what aspects of the performance subgroups of assessors found salient enough to record in written comments. Similarities and differences in the content of the written comments for each of the three performances were identified using a blinded, participant-based sorting procedure (*F*-sort and Latent Partition Analysis) (Miller et al., 1986). Phase 2 quantified the link between variability in the content of the comments with variability in the assigned scores (using ANOVA). Phase 3 determined if the content of written comments provided under the conditions of significant contrast effects was different from the content provided under the conditions of insignificant contrast effects (using Chi square).

**Existing data.** Yeates and colleagues investigated the influence of contrast effects on assessors’ scores and feedback (Yeates et al., 2015a). The study’s 61 participants were consultant or registrar level assessors from the UK who had previously received training in assessment. The study was conducted via the internet. Participants were randomised to one of three groups and viewed a set of three simulated performances in varied order. Each performance featured a different foundation doctor consulting with the same patient. Performances depicted either a good, borderline or poor performance on the same task. These levels of performance had been validated by prior work (Yeates et al., 2012; Yeates et al., 2013a). The three target performances were assessed both “unpreceded” (i.e. they were the first video assessed and not preceded by another performance) and “preceded” by a performance of a different standard (i.e. they were the second video assessed and exposed to conditions intended to induce contrast effects). Assessors scored the first performance and provided up to 200 words of written feedback before completing the same tasks with the second performance. Scores were allocated using the scoring matrix from the original UK Foundation Programme mini-CEX score sheet. This asked for seven domains(history taking; physical examination; communication skills; critical judgement; professionalism; organisation and efficiency, and overall clinical care) to be scored using a 6-point scale anchored against expectations for completion of Foundation Year 1: 1 = well below expectations; 2 = below expectations; 3 = borderline; 4 = meets expectations; 5 = above expectations, and 6 = well above expectations. Feedback was sought with the instruction: ‘Please describe the strengths and weaknesses of this performance that you would like to convey to this trainee. Please describe what was done well, areas for improvement, and items they should plan to improve.’

A schematic of the performances seen by each group is shown in Table 1. Results from that study showed a clear contrast effect on the good performance, and on the borderline video, but not on the poor performance (Yeates et al., 2015a). This dataset offered an opportunity to study the above research questions in instances where contrast effects did, and did not, occur.

[Insert Table 1 about here]

**Phase 1: Identifying similarities and differences in the content of written comments**

**Overview.** We used a blinded, participant-based sorting procedure (termed an *F*-sort) to detect similarities and differences in the content of the written comments for each performance. Latent Partition Analysis was used to compile the participants’ sorting decisions to determine which comments were similar enough to be grouped together and sufficiently dissimilar from other comments to be separated from them. This identified clusters of consensus for each performance with each cluster containing a different quantity of described performance aspects or a differently evaluated performance aspect. It also provided an opportunity to replicate previous findings in a novel dataset.

**Population, sampling and recruitment.** The study population of sorters was healthcare professionals, healthcare professional students and trainees, and assessment coordinators. This population was chosen because, owing to their familiarity with giving, receiving and/or analysing assessment feedback in a healthcare context, we judged that they would be capable of meaningfully sorting feedback based on similarity and differences in construct-relevant content. Sorters were recruited via email and social media and sampled by convenience in Prince George, British Columbia, Canada. The sample size was set pragmatically at 45 participants based on prior research because formal power calculations are not possible for latent partition analysis.

**Materials and sorting procedure.** The corpus of data to be sorted comprised every free text written comment collected for the first and second videos viewed by participants in Yeates et al 2015a, and were thus either “unpreceded” or “preceded” (see earlier for details). This amounted to 122 individual portions of feedback, and approximately 8000 words. One participant did not provide any written feedback for the performances. Since this data was essential for the sorting task (described next) this assessor’s scores were also excluded from all analyses. The scores from this excluded assessor were within 1 standard deviation of the mean for each performance, and excluding them did not alter the overall pattern of results. The written comments were uniquely identified in a manner that obscured whether they were collected unpreceded or preceded but allowed researchers to trace their identity and associated scores. Each individual written comment was printed on a separate slip of paper, organised by the 3 different performances (good, borderline and poor) and then randomly ordered within performances. Sorters worked individually at the task. Each was given a stack of randomly ordered responses and asked to freely sort them into piles according to the sorting cue:

*“Take a slip of paper from the stack, read the assessment comment and place it on the table. Now read the comment on the next slip of paper. If the content of this comment is similar to the first comment, then place it in the same pile. If the content of this comment is different from the first comment, then start a new pile. Continue doing this for all the comments. You can use as many piles as you like and you can rearrange them as often as you’d like.”*

Each sorter performed this process consecutively for the three performances. To counteract learning and order effects, the order that sorters worked through the performances was rotated. Once finished, researchers recorded the composition of each pile for each sorter, and the sorter’s summary of what was being described in each pile.

**Determination of similar and dissimilar content**. The composition of the sorters’ piles was compiled and analysed using Latent Partition Analysis (LPA). LPA calculations (Miller et al., 1986; Wiley, 1967) detect comments that multiple sorters combined into the same pile and identify partitions between comments that were rarely or never grouped together. LPA has been described previously (Gingerich et al., 2014b) and was chosen because it allows empirical investigation of the content elements, number and size of latent categories as well as quantification of the relationships among the latent categories (Miller et al., 1986). LPA was conducted using RStudio version 3.3.1 (2016).

The mathematical technique of LPA has some computational and procedural similarities to factor analysis (Miller et al., 1986). For example, it requires that researchers choose the number of partitions to be tested. To ensure all possible best fitting models were tested, we tested every possible set of partitions starting with two and extending beyond the maximum number of groupings created by sorters. For each number of partitions that were tested, a separate *phi* matrix was produced. The phi matrices indicated, for each performance, how well each number of partitions fitted the data, and how strongly each comment belonged to each latent category. The analysis also supplied an *omega* matrix for each set of partitions that were tested. The omega matrices quantified how cohesive each partition was and how much it overlapped with the other partitions in a given set; they showed the probability of comments being placed into their assigned partition and the probability of them also being placed into another partition within the set (Miller et al., 1986). We used these results to inform our decisions regarding the best fitting number of partitions for each performance.

**Identifying the variable content of written comments.** LPA informs decisions but does indicate what is different about the clusters of consensus. PY and AG independently identified and listed the described aspects within each cluster of consensus, provided a summarizing label for the clusters, and chose a quote to represent each cluster before meeting to critique the analysis. The cycle was repeated until consensus on the content of each cluster was reached.

**Phase 2: Linking written comments with variable content to variable scores**

**Preparing data for analysis.** The best fitting number of partitions for each performance had been determined and the resulting groupings of similar written comments were identified as clusters of consensus. This meant each assessor’s written comment belonged to only one cluster. We then unblinded the data to unite each written comment with its associated score and to reveal if it corresponded to a preceded or unpreceded performance. As a result each portion of written feedback was associated with: a cluster; a target performance (either preceded or unpreceded); and a competence score. Competence scores were derived from the mean of the scores which each participant gave on the Mini-CEX 6-point scale to each of the 7 performance domains. This produced a scale with 36 different possible values. We chose to use this average value as a measure of overall competence as prior research on contrast effects has demonstrated their influence on overall performance scores rather than an influence on domain-level scores. We chose to use the mean (rather than the median) as prior research has demonstrated that collapsed mini-CEX scores are normally distributed (Yeates et al., 2013) and collapsed Likert scales can be treated as interval for the purposes of analysis (Stevens, 1946) without biasing parametric analyses (Carifio and Perla, 2008).

**Analysis.** To determine the link between variability in the content of the written comments and variability in the scores, we performed a one-way ANOVA separately for each performance. Since assessors and their scores were assigned to one of the clusters in the set, we used membership in the clusters of consensus as the levels of the independent variable. The dependent variable was the mean of the mini-CEX scores in the 7 assessed domains. This analysis allowed us to compare the mean competence scores for assessors grouped in one cluster (based on the similarity of their written comments) to those grouped in different clusters (due to differences in their written comments). The analysis also determined the proportion of competence score variance that could be explained (eta squared) when we accounted for assessors providing written comments that could be grouped into a set of clusters of consensus.

**Phase 3: Determining if contrast effects induced content differences in written comments**

**Analysis**: If contrast effects were influencing the content of written comments then the content would be different in the presence of significant contrast effects and unchanged when the contrast effects were insignificant. Each cluster of consensus in a set grouped a certain number of assessors’ comments that described similar content. To determine if there was a change in the content, we analysed for differences in the proportion of comments in the set of clusters of consensus for a performance when it was assessed unpreceded as compared to the proportion of comments in the set when it was assessed preceded by another performance. We used Pearson’s Chi Square test to determine if any differences were significant. Analyses were conducted using IBM SPSS Statistics 21. Note: A two-way ANOVA with contrast effects as one independent variable and clusters of consensus as another would have been a reasonable approach for identifying an interaction of the effects on the scores. Unfortunately, some of the groups in this model (as small as a single member) were too small to support trustworthy results using this analysis.

**Results**

**Phase 1: Identifying similarities and differences in the content of written comments**

**Sorters and *F-*sorts.** 45 participants completed the sorting procedure, with no partial completions. Demographically they comprised of 25 health professional students from medicine, nursing and physiotherapy; 13 physicians and 4 other health professionals; and 3 assessment coordinators. Sorters indicated that the written comments seemed plausibly authentic, and that they could discern differences in written feedback from different assessors. The groupings generated by sorters ranged from: 2 to 8 (median 4) for the good performance; 3 to 8 (median 5) for the borderline performance; and 3 to 10 (median 4) for the poor performance. Sorters grouped the written comments based on differences in the content of the feedback given as well as differences in how the feedback was written and expressed.

**Determination of similar and dissimilar content.** Latent Partition Analysis does not indicate the optimum number of clusters. Instead it gives quantitative fit parameters to guide researchers to judge the meaningfulness of each set of clusters and determine which set fits best for the given purpose. None of the sorters indicated that the written comments could be best summarized by a single group and, therefore, one cluster is not the best fit. Fit parameters were used to guide our focus on a certain subset of the sets of clusters identified through analysis. For example, sets with items with impossible fit values in the phi matrix were discarded. Eigenvalues and scree plots suggested 3-6 clusters were most reasonable; the best fitting set of clusters was certainly less than 10. Researchers (AG, PY, TS) independently reviewed the fit indices and the content of comments grouped by each of the reasonable sets of clusters. Each researcher identified and justified a set of clusters as having the best fit for each video before meeting to present their choices. After three rounds of independent deliberation and group discussion the best fitting set of clusters for each video was agreed upon and used for subsequent analyses (see Table 2 for details).

As a brief summary of those deliberations, a set of two clusters divided the comments based on valence of the feedback but the content of the comments was too varied and indistinct. As the number of clusters in the set increased, the content contained in each became more specific and/or the style of feedback delivery became more similar. At some point the additional clusters made one or more clusters in the set difficult to distinguish from the others. Compelling arguments could be made for two or three different models being the best fit for each video but a single best set of clusters was agreed upon. Due to the subjectivity of the decision, the analysis described in the following section was performed on all reasonable sets of clusters. The pattern of results was similar and would lead to the same conclusions regardless of which set was chosen. As is the convention with factor analysis, this superfluous analysis is not presented. Nonetheless, we found that the phenomenon of clusters of consensus was replicated in this dataset, producing either 3 or 4 clusters of consensus for each performance.

**Identifying the variable content of written comments.** Table 2 provides specific details regarding the sets of clusters of consensus and the main findings are summarized in the following. There was substantial similarity in the content of two or more clusters in a set. For example, all four clusters of consensus in the set identified for the ‘Good’ performance described a strong performance. All four began their written feedback with positive, complimentary comments reinforcing well-performed aspects of the encounter. Two clusters were entirely complimentary and a difference between them was the length of the comments and the number of aspects described: The cluster labelled ‘excellent communication’ had comprehensive descriptions (averaging 61 words per comment) of numerous aspects of the performance and ‘excellent interviewing’ had much briefer comments (averaging 21 words per comment) touching on only a few aspects of performance. These two clusters contained feedback with the same valence and similar content but one was less complete in its description of the performance. The other two clusters contained suggestions for improvement with each emphasizing a slightly different domain: ‘omitted questions’ focused more on aspects of the interviewing technique and ‘missed differentials’ focused more on aspects of clinical reasoning. The most noticeable difference regarding the ‘missed differentials’ cluster was an absence of comments on the Foundation doctor’s communication skills since the other three clusters were effusive about those aspects. Overall, there were many similarities in the content of the clusters of consensus for the Good performance and the identified differences included which aspects of the performance were commented on and/or emphasized, and whether or not suggestions for improvement were provided.

The content of comments could differ in other ways, for example, comments in different clusters could describe the same behaviour but express the corresponding feedback with a different tone. For the Borderline performance there were two clusters that focused on the behaviour of using a combination of open and closed questioning techniques. The comments in the ‘open questions needed’ cluster were largely positive and suggested that the interview could be improved by beginning with open questions. These suggestions were made using phrases such as, “could try”, “consider”, and “may help to”. In contrast, another cluster in the set ‘disorganized closed questions’ grouped definitive statements criticising a lack of open questions or too many closed questions. The feedback was largely negative and harshly phrased including comments from rater 23, “…all closed questions and asked in a seemingly random order” and rater 14, “Questions were often direct (allowing yes/no answers) and inappropriately worded (‘did you wet yourself’ in a patronizing manner)”. Despite emphasizing a similar judgment about the mix of open and closed questions, the valence, tone, and choice of phrasing in the written comments was noticeably different.

There were also examples where contradictory feedback was given in different clusters for ostensibly the same observed behaviour. For the Poor performance, all four clusters contained negative evaluations of skills, constructive feedback, or highlighted errors and omissions. Three of the clusters grouped comments that negatively described the Foundation doctor’s non-verbal communication (e.g. leaning back in chair, playing with pen, limited eye contact) as being too laid back and demonstrating disinterest. They also faulted her for deferring everything to the consultant physician including basic investigations that were within her scope to initiate. In contrast, the feedback in the cluster ‘recognized limitations’ included positive comments about the trainee using a calm, relaxed, and unhurried approach (presumably inferred from non-verbal communication) and commended her for knowing her limitations and waiting for the consultant’s input. This cluster was not entirely positive but its valence differed from the other three clusters due to the positive interpretation of a greater number of aspects of the performance.

In summary, there were sufficient commonalities and differences in the content, valence, or tone of the feedback comments for a set of clusters of consensus to be identified for each performance. The content grouped into a cluster was not exclusive to that cluster but it represented an identifiable uniqueness. If assessors comment on the performance aspects that grab their attention the most and feature most prominently in their assessment judgments then this finding suggests assessors found different aspects of the performances to be salient and a different number of aspects to be salient.

[insert Table 2 about here]

**Phase 2: Linking written comments with variable content to variable scores Analysis:** Variability in the content of the comments was determined by identifying a set of clusters of consensus, for each performance. As shown in Table 3, the clusters of consensus grouped together feedback comments and their corresponding competency scores. To determine if variability in the content of the comments explained variability in the scores we performed an analysis of variance (ANOVA) separately for each performance. The differences in the mean of scores for the clusters of consensus were significant for the Good (F=3.9, p=0.016) and Poor performances (F=6.3, p=0.001) and nearly significant for the Borderline performance (F=3.0, p=0.06). Accounting for variability in the content of the comments for the clusters of consensus explained 25.8% of score variance in the good performance, 13.7% of score variance in the borderline performance; and 33.9% of score variance for the poor performance. In summary, groups of assessors who provided more similar written comments also assigned more similar scores than assessors who provided dissimilar feedback.

[insert Table 3 about here]

**Phase 3: Determining if contrast effects induced content differences in written comments**

**Analysis**. The content of the comments was variable, clusters of consensus were identified, and they can be used to explain variance in scores both when contrast effects are significant (e.g. in the good performance) and when they are not (i.e. in the poor performance). Our findings therefore indicate that clusters of consensus can occur independently of contrast effects. To determine if contrast effects influenced the content of written comments, we compared the set of clusters of consensus when the performance was assessed preceded to the set that was produced when it was assessed unpreceded. More specifically, we compared how the comments were distributed within the clusters of consensus when the performance was assessed preceded versus unpreceded. Chi Square analyses indicated that the preceded clusters of consensus had a significantly different proportion of comments than the unpreceded clusters of consensus for the two performances with significant contrast effects (the good and the borderline performance) and not for the performance with non-significant contrast effects (the poor performance). As a brief summary of findings in Table 4, when the Good performance was assessed following the Poor performance there was a greater number of entirely complimentary comments (‘excellent communication’ cluster) (X2=10.37; p=.016); when it was assessed unpreceded there was a greater proportion of comments with suggestions for improvement (the ‘omitted questions’ and ‘missed differentials’ clusters). When assessment of the Borderline performance was preceded by the Good performance, there was a greater proportion of comments providing critical feedback (‘disorganized closed questions’ cluster) and when it was assessed unpreceded there was a greater proportion of comments highlighting aspects done well (‘detailed history’ cluster) (X2=9.45; p=.009). The proportions in the set of clusters of consensus did not significantly vary for the video that did not show a contrast effect: the poor performance (X2=4.88; p=.18). In summary, when scores were *not* significantly affected by contrast effects, the set of written comments produced for the preceded performance was *not* significantly different from the set produced for the unpreceded performance. When contrast effects did induce a significant change in scores, there was also a significant difference in the sets of written comments produced.

[insert Table 4 about here]

**Discussion**

Assessment of a preceding performance can result in assessors assigning scores to the target performance that are shifted away from the standard of competence in the preceding performance (Yeates et al., 2015a; Yeates et al., 2015b; Yeates et al., 2012; Yeates et al., 2013a). Called contrast effects, a proposed mechanism for the shift in scores is our human tendency to make all judgements of magnitude in comparative terms (Parducci and Perrett, 1971). More specifically, the context of the preceding performance would recalibrate the assessor’s internal scale by altering the range-frequency on which it is based (Stewart et al., 2005). It offers an explanation for variability in scores based on a change in how the scale is conceptualized and used to assign scores. Clusters of consensus are identified when there is sufficient similarity in the content of variable written comments for subsets of comments to be grouped together. A proposed mechanism for clusters of consensus is that impressions are formed by differently combining interrelated aspects of a performance (Higgins, 2000) around a central salient aspect to build a coherent interpretation (Mohr and Kenny, 2006; Park et al., 1994). It offers an explanation for variability in scores due to different overall evaluative judgments resulting from different centralized aspects and different combinations of performance aspects. The current study aimed to further investigate assessor cognition by using contrast effects to change the assessment context and examining the content of written comments through identification of clusters of consensus to reveal any change in the salience of performance aspects.

In this study, when a few dozen physicians provided written feedback comments as part of the assessment of a foundation doctor’s clinical performance, a set of clusters of consensus representing three or four different versions of feedback were identified for each performance. The clusters varied in the number of aspects of the performance that were described, how they were interpreted, and how the feedback was expressed and accounting for there being three or four versions of written feedback explained 14-34% of variance in the scores. These findings are consistent with those reported in previous health professions research (Gingerich et al., 2017; Gingerich et al., 2014b; Kogan et al., 2011; Tavares et al., 2016; Yeates et al., 2012; Yeates et al., 2013b). .

We expect elements of assessor cognition to be revealed by identifying clusters of consensus in a dataset with both significant and non-significant contrast effects. If assessors write about the aspects of the performance they find most salient, as suggested by previous psychology research (Fiske and Cox, 1979), differences in the number of aspects described and how they are described are indications of differences in what the assessors found salient in the performance. Clusters of consensus were identified both when contrast effects were significant and non-significant but the proportion of assessors making similar comments only significantly differed between groups when contrast effects were significant. For example, not only were the scores of the Good performance significantly higher when it was assessed following the Poor performance (which featured inadequate communication skills) but there were more comments grouped into the cluster with entirely complimentary statements about communication skills. When the Borderline performance was assessed unpreceded the majority of comments described the history as reasonably thorough and relevant but only a single assessor made those comments when it was assessed following the Good performance (which featured a detailed history). In other words, when contrast effects were significant, assessors more frequently commented on aspects of the current performance that were dissimilar with the standard of competence demonstrated in the preceding performance. Since the content of the comments was affected, in addition to the valence of the comments and the scores, the preceding performance had influences beyond those on assessors’ range-frequency scales. Comparative distinctiveness (Higgins, 1996; Higgins, 2000) is a mechanism for salience that could explain differences in the content of the comments when contrast effects are significant. Comparative distinctiveness refers to the properties of the performance, not the assessor, and how the surrounding context can cause some features to be perceived as more distinct and selectively draw the attention of the assessor to them (Higgins, 1996). When two different performances of a clinical encounter are assessed back-to-back, the difference in the level of demonstrated competence for a given aspect of a performance may cause it to be more noticeable, draw the assessor’s attention to that aspect, and result in it featuring more prominently in the assessment judgment. In other words, it is the comparison invoked by the preceding performance that makes the differently performed aspect more distinctive and hence changes its salience.

**Implications for assessment and future research**

There was substantial variability in the content of the written comments provided for a given performance, and in a real-life situation, a learner only receives one version of the feedback. Our findings support the recommendation of acknowledging the subjectivity of feedback (Ende, 1983). In addition to the differences in feedback content, the variability in how feedback was expressed could convey quite different messages to the student. For example, the written comments in some clusters used gentler hedging language (Ginsburg et al., 2016) of ‘might try’ and ‘could consider’ whereas others used a harsher, almost punitive tone in noting areas of improvement. It seems prudent to continue offering guidance on effective ways to exchange feedback to promote better learning opportunities (Ramani and Krackov, 2012).

Our findings also support assessment designs that accommodate variability by sampling widely: collecting many pieces of feedback from many different people in many different contexts over time and collating them before making any high stakes decisions (van der Vleuten et al., 2012). Importantly responding to assessor variability through increased sampling (i.e. increased numbers of assessments) is likely, in addition to counterbalancing assessor differences, to have the advantage of sampling broadly across clinical cases and thereby further enhancing the reliability of the resulting score. Conversely, when pragmatic constraints limit such sampling, the impact of assessor variability may be expected to be greater. An important consequence of feedback and assessment judgements that was not explored in the current study is their impact on corresponding supervisory decisions. Further research is needed to determine if the different versions of feedback are related to different supervisory behaviours with different patient safety outcomes.

Perhaps the most pertinent line for future research is to determine whether it is possible to harness the influence of the preceding performance to improve assessment judgements by directing assessor’s attention to certain aspects of the performance and shifting how they frame or interpret information. A possible intervention to combat grade inflation for Objective Structured Clinical Examinations would be to have examiners assess several reference Good performances of the specific station before examining a circuit. This would provide an opportunity to draw their attention to the key aspects of the performance and potentially cause skills that are performed at a lower level of competence to be seen as more salient in a subsequent real performance. Such an intervention would be inadequate for assessor training on its own but might form part of (or be conducted in addition to) comprehensive benchmarking exercises as part of quality assurance procedures.

Future research could help with identifying influences on assessors’ attention, judgment and other cognitive processes. Regardless of whether the influences are determined to be internal or external sources of information (St-Onge et al., 2016) or factors (Lee et al., 2017) with effects on observation, processing and/or integration mechanisms (Gauthier et al., 2016) that faculty use as lenses when observing during assessments (Kogan et al., 2011), once identified, those influences could be better accommodated within assessment designs. A better understanding of what contributes to the variability could allow for it to be better modelled within analyses.

**Limitations**

It is important to recall that the comments were written in response to a question requesting feedback as part of an assessment of a foundation doctor’s performance during experimental conditions. Since the comments were meant to capture feedback, they are an imperfect proxy of assessors’ observations and judgements. This serves as a reminder to be clear about the purpose of written comments when designing assessment forms. If the goal of having assessors record written comments is to provide effective feedback, then focusing on a select few skills to offer suggestions for improvement or comments to reinforce behaviours done well, fits with best practices (Cantillon and Sargeant, 2008). However, if the goal of recording comments is to document a justification of the assessment judgements then the missing aspects may be problematic in terms of determining if the entire construct of interest has been assessed. This is a limitation of the study which arose from the use of secondary data and prospective replication of the findings could address this issue by explicitly requesting that assessors describe all issues which they considered salient to their judgement.

A limitation of using the sorting exercise is that it was based on the similarity of comments and has the potential to have been influenced by factors that were not related to formation of assessment judgments (e.g. preferred writing style or time constraints as opposed to salience). Our sample size for the sorting exercise may have limited the sensitivity to detect more subtle differences between the proportions of membership of each cluster. As no clear method exists for performing power calculations for LPA, it is difficult to determine how well powered our analysis was. Our study population comprised individuals who are familiar with assessment feedback in medical education and a minority were assessors, but it is possible that those providing the written comments could have clustered responses differently. This could also be explored in future research.

We have acknowledged the subjectivity of LPA methodology used to analyse content differences in written comments. Although the exact number of clusters of consensus can be debated, the methods do confirm that the content of written comments is variable and that subgroups of assessors provided similar comments to each other and different from other subgroups. The methods did not allow for analysis of the accuracy or validity of the content contained in the written comments. The findings indicate contrast effects induce changes in written comment content but do not provide empirical evidence of an explanatory mechanism.

The findings which we have described do not implicitly generalise to all other situations in which contrast effects could arise, and so replication of these findings should be sought in other contexts. The use of an experimental methodology had the potential to miss effects which might arise in the complex interactions of assessment in real practice. Nonetheless, the stimulus material used in this study was derived from highly authentic simulations of mini-CEX assessments featuring qualified doctors and highly realistic case presentations. As such there is no implicit reason why these findings may not generalise to similar assessment scenarios.

**Conclusions**

In this study we replicated the finding of clusters of consensus and revealed that the content of written comments can be influenced by the assessment of a preceding performance (i.e. contrast effects). The information that assessors considered salient within a performance (i.e. the cluster into which they were grouped based on the content of their written comments) appears to be influenced by their recent assessment of another performance. This suggests that assessors’ perceptions of salient performance information may be influenced by the contrasting valence or information in the preceding performance. More research is needed to determine if the resulting changes to assessors’ comments and scores are indicative of biased or improved assessment information. The potential should be explored to augment existing benchmarking procedures by cueing assessors’ attention through observation of reference performances immediately prior to assessing trainees’ performance.

Acknowledgements: We are grateful for partial funding provided by a Northern Medical Program seed grant.

Disclosures: none

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research ethics boards.

**References**

Alves De Lima, A., Barrero, C., Baratta, S., Castillo Costa, Y., Bortman, G., Carabajales, J., Conde, D., Galli, A., Degrange, G. & Van Der Vleuten, C. (2007). Validity, reliability, feasibility and satisfaction of the mini-clinical evaluation exercise (mini-CEX) for cardiology residency training. *Medical Teacher,* 29(8): 785-790.

Cantillon, P. & Sargeant, J. (2008). Giving feedback in clinical settings. *BMJ,* 337(nov10\_2): a1961-a1961.

Carifio, J. & Perla, R. (2008). Resolving the 50‐year debate around using and misusing Likert scales. *Medical Education,* 42(12): 1150-1152.

Cook, D. A., Beckman, T. J., Mandrekar, J. N. & Pankratz, V. S. (2010). Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. *Advances in Health Sciences Education,* 15(5): 633-645.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education,* 38(9): 1006-1012.

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical Education,* 39(4): 353-355.

Ende, J. (1983). Feedback in clinical medical education. *JAMA: The Journal Of The American Medical Association,* 250(6): 777-781.

Fiske, S. T. & Cox, M. G. (1979). Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others. *Journal of Personality,* 47(1): 136-161.

Gauthier, G., St‐Onge, C. & Tavares, W. (2016). Rater cognition: Review and integration of research findings. *Medical Education,* 50(5): 511-522.

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M. & Holmboe, E. (2014). Seeing the ‘black box’ differently: Assessor cognition from three research perspectives. *Medical Education,* 48(11): 1055-1068.

Gingerich, A., Ramlo, S. E., van der Vleuten, C. P. M., Eva, K. W. & Regehr, G. (2017). Inter-rater variability as mutual disagreement: Identifying raters’ divergent points of view. *Advances in Health Sciences Education,* 22(4): 819-838.

Gingerich, A., van der Vleuten, C. P., Eva, K. W. & Regehr, G. (2014). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in mini-CEX ratings. *Academic Medicine,* 89(11): 1510-1519.

Ginsburg, S., van der Vleuten, C., Eva, K. & Lingard, L. (2016). Hedging to save face: A linguistic analysis of written comments on in-training evaluation reports. *Advances in Health Sciences Education,* 21(1): 175-188.

Higgins, E. T. (1996). Activation: Accessibility, and salience. (In E. T. Higgins and A. W. Kruglanski (Eds.). Social psychology: Handbook of basic principles(pp. 133-168). New York, NY The Guilford Press.)

Higgins, E. T. (2000). Social cognition: Learning about what matters in the social world. *European Journal of Social Psychology,* 30(1): 3-39.

Hill, F., Kendall, K., Galbraith, K. & Crossley, J. (2009). Implementing the undergraduate mini-CEX: A tailored approach at southampton university. *Medical Education,* 43(4): 326-334.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W. & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education,* 45(10): 1048-1060.

Lee, V., Brain, K. & Martin, J. (2017). Factors influencing mini-CEX rater judgments and their practical implications: A systematic literature review. *Academic Medicine,* 92(6): 880-887.

Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P. & Hawkins, R. E. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine: Journal Of The Association Of American Medical Colleges,* 81(10 Suppl): S56-S60.

Miller, D. M., Wiley, D. E. & Wolfe, R. G. (1986). Categorization methodology: An approach to the collection and analysis of certain classes of qualitative information. *Multivariate Behavioral Research,* 21(2): 135.

Mohr, C. D. & Kenny, D. A. (2006). The how and why of disagreement among perceivers: An exploration of person models. *Journal of Experimental Social Psychology,* 42(3): 337-349.

Parducci, A. & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology,* 89(2): 427.

Park, B., Dekay, M. L. & Kraus, S. (1994). Aggregating social behavior into person models: Perceiver-induced consistency. *Journal of Personality and Social Psychology,* 66(3): 437-459.

Ramani, S. & Krackov, S. K. (2012). Twelve tips for giving feedback effectively in the clinical environment. *Medical Teacher,* 34(10): 787-791.

St-Onge, C., Chamberland, M., Lévesque, A. & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: Expert assessment of examinee performance. *Advances in Health Sciences Education,* 21(3): 627-642.

Stevens, S. S. (1946). On the theory of scales of measurement.

Stewart, N., Brown, G. D. A. & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review,* 112(4): 881-911.

Tavares, W., Ginsburg, S. & Eva, K. W. (2016). Selecting and simplifying: Rater performance and behavior when considering multiple competencies. *Teaching and Learning in Medicine,* 28(1): 41-51.

van der Vleuten, C., Schuwirth, L., Driessen, E., Dijkstra, J., Tigelaar, D., Baartman, L. & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher,* 34(3): 205-214.

Weller, J., Jones, A., Merry, A., Jolly, B. & Saunders, D. (2009). Investigation of trainee and specialist reactions to the mini-clinical evaluation exercise in anaesthesia: Implications for implementation. *British journal of anaesthesia,* 103(4): 524-530.

Wiley, D. E. (1967). Latent partition analysis. *Psychometrika,* 32(2): 183-193.

Yeates, P., Cardell, J., Byrne, G. & Eva, K. W. (2015). Relatively speaking: Contrast effects influence assessors’ scores and narrative feedback. *Medical Education,* 49(9): 909-919.

Yeates, P., Moreau, M. & Eva, K. (2015). Are examiners’ judgments in osce-style assessments influenced by contrast effects? *Academic Medicine,* 90(7): 975-980.

Yeates, P., O'Neill, P., Mann, K. & Eva, K. W. (2012). Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA: Journal of the American Medical Association,* 308(21): 2226-2232.

Yeates, P., O'Neill, P., Mann, K. & W Eva, K. (2013). 'You're certainly relatively competent': Assessor bias due to recent experiences. *Medical Education,* 47(9): 910-922.

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. *Advances in Health Sciences Education,* 18(3): 325-341.

Table 1: **Summary of contrast effect findings in pre-existing dataset from Yeates et al 2015a**

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Initial Performance | Second Performance | Contrast effects |
| A (n= 19) | Good | Borderline | Significant |
| B (n= 20) | Poor | Good | Significant |
| C (n=22) | Borderline | Poor | Not significant |

Table 2. **Content of clusters of consensus for three performances**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Summary of content of feedback | Summary of how feedback was expressed | Representative quote |
| **Good performance** | | | |
| ‘excellent  communication’ | - good/excellent history  - good/excellent communication  - got collateral history from witness | - entirely positive evaluation of all skills  - comprehensive comments  - no suggestions for improvement | “Excellent history starting with open questions in a logical order (what happened, then events before and during blackout) followed by specific questions to clarify the history. Also asked about collateral history and got details of a witness. Good eye contact and body language throughout, was engaged and interested in what the patient was saying and had a very professional manner. Had obviously given a clear explanation to the patient and then checked his understanding and then explained what would happen next.” (rater 52) |
| ‘excellent  interviewing’ | - open questions asked in logical order  excellent listening skills during history-taking  - let patient tell story  - organized approach | - entirely positive description of skills  - no suggestions for improvement  - brief comments | “Importantly, allowed patient to talk without interruption, asked all the right questions and was calm and confident. Also, gave him a chance to ask her anything he might not have understood, at the end.” Rater 18 |
| ‘omitted questions’ | - good use of open questions  - structured history  - omitted history questions (e.g. associated symptoms and exacerbating factors of seizures, previous medical history, family history)  - asked for collateral history from witness | - starts with positive evaluation of skills then moves to suggestions for improvement (noting omissions). | “Good open questions to start and allowed patient to expand. Didn't move to closed questions too soon, but when did asked relevant and discriminatory ones. Good non-verbals. Generally clear question style, but try to not ask 'compound questions' eg did you feel hot and unwell - split into two separate questions. Important to get a collateral here - well done. Could investigate a little further regarding patient's specific concerns. Need to always consider dangerous pathology - ask about family history of sudden death.” Rater 1 |
| ‘missing differentials’ | - good use of open questions  - omissions in history (e.g. relevant family history)  - suggestions to improve differential  - no mention of communication skills | - positive evaluation of one skill with an omission noted and/or a suggestion for improvement.  - all but one comment is brief | “Good open questioning initially and then closed questioning, possibly could have thought what had caused this and given few differential diagnoses to patient that that’s what we are going to work on.” Rater 7 |
| **Borderline video** | | | |
| ‘disorganized closed questions’ | - notes too many closed questions  - disorganized history  - lack of reassurance to patient  - anchored to emergency department diagnosis of “fit”  - used jargon | - primarily negative evaluation of skills often followed by suggestions for improvement | “A bit hesitant with a poor structure to the history. Questions were often direct (allowing yes/no answers)and at times inappropriately worded ("did you wet yourself" in a patronising manner). Not reassuring in any way to patient and saying need to do some more tests an insufficient explanation.” Rater 14 |
| ‘open questions needed’ | - noted good manner, friendly, polite  - suggestion to start with open questions | - starts with positive statement and then suggestions for improvement | “You were very friendly and took time to answer questions but would have probably improved your encounter by starting with more open questions, that would have answered some of the issues that you asked later. You are ok to defer some issues to your senior. Recommended action: practice start with open questions and review red flags for patients with suspected epilepsy.” Rater 35 |
| ‘detailed history’ | - asked pertinent history questions to differentiate fit and faint  - more reassurance needed  - need to learn driving guidelines re: seizures | - positive evaluation of skills followed by suggestions for improvement | “Reasonable history and attempt to differentiate between syncope and seizure. Sensible questioning overall. Driving discussion could have been handled a lot better. Even if you don't know driving guidelines tell patient that you will look them up and get back to him.” Rater 36 |
| **Poor video** | | | |
| ‘rude’ | - no open questions  - asked leading questions  - unstructured interview  - disinterested body language  - appeared rude  - did not address driving concerns  - no empathy demonstrated | - primarily negative evaluation of skills with omissions noted  - suggestions for improvement rarely given  - brief comments, short sentences | “Difficult to find any positive comments! Very erratic order of questions asked. No real open questions used. Doctor appeared disinterested in the patient and asked questions that were obviously worrying to the patient with no obvious logic from the patient's point of view […]. She did not address his concerns re: driving. On the positive side, at least she was going to check with the attending before requesting a head scan but she should be able to begin basic investigations as a PGY1 doctor. She appeared rude and unapproachable.” Rater 23 |
| ‘too casual’ | - some reasonable questions  - disinterested and unprofessional body language (pen tapping)  - empathy and communication poor  - driving question could be handled better  - appeared unsure, hesitant, lacking confidence | - led with a positive statement and then negative evaluation of multiple skills  - suggestions for improvement rarely given | “Doctor tried to get out the salient points from the history - what happened, previous history and witnesses. Asked about associated symptoms and communicated the plan to the patient. However, she had a very relaxed attitude to the patient, did not appear to be interested in the presenting complaint. She dismissed his concerns and used medical jargon when explaining what would happen next (EEG). She was unsure about driving rules. Her questioning style was closed questions and she did not appear to be professional or taking an interest in the patient.” Rater 59 |
| ‘anchored to diagnosis’ | - some relevant history  - no introduction  - disinterested body language  - anchored to emergency department diagnosis  - questioning worried the patient  - didn’t provide next steps in management  - did not answer driving questions | - leads with a positive statement then comprehensive suggestions for improvement for multiple skills  - used “you” pronoun | “You asked appropriate questions relating to a seizure "shaking", "incontinence" etc, but I felt your mind was already made up that this patient had had a seizure. A lot of your questions were closed questions so you got answers quickly but potentially missed out on the whole picture.[…]Also, watch your body language- you tend to flick your arms away as if dismissing comments (you may not be aware), this looks quite bad, as if you are disinterested. I had the feeling you were going through a check list rather than getting to the bottom of things by building a rapport.[…] Rater 42 |
| ‘recognized limitations’ | - calm, relaxed approach  - asked some relevant questions  - knows limitations  - needs more open questions  - work on explaining plan of management | - leads with positive statements and then offers suggestions for improvement  - brief comments, short sentences. | “Asked some relevant questions and summarised and checked understanding. In future more use of open questions and more space for the patient to tell his story may give a better, fuller history. Some cues were ignored. A wider differential diagnosis could be considered. Appropriately recognised your limitations.” Rater 43 |

Table 3. **Linking written comments with variable content to variable scores by using clusters of consensus to explain variability in scores with one-way ANOVA**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Set of clusters of consensus | Number of assessors | Mean of ratings | Cohen’s d | Variance explained (partial eta squared; %) (p value) |
| Good performance | ‘excellent communication’ | 13 | 5.0 | 0.1 a  0.9  1.9 | 25.8  (.016) |
| ‘excellent interviewing’ | 7 | 5.0 | 0.1  0.6  1.3 |
| ‘omitted questions’ | 12 | 4.5 | -0.9  -0.6  1.1 |
| ‘missed differentials’ | 6 | 4.2 | -1.9  -1.3  -1.1 |
| Borderline performance | ‘detailed history’ | 11 | 3.8 | 0.6  0.9 | 13.7  (.06) |
| ‘open questions needed’ | 14 | 3.3 | -0.6  0.4 |
| ‘disorganized closed questions’ | 16 | 3.0 | -0.9  -0.4 |
| Poor performance | ‘recognized limitations’ | 10 | 3.1 | 0.9  1.3  1.5 | 33.9  (.001) |
| ‘anchored to diagnosis’ | 7 | 2.3 | -0.9  0.3  0.7 |
| ‘too casual’ | 12 | 2.1 | -1.3  -0.3  0.4 |
| ‘rude’ | 12 | 1.9 | -1.5  -0.7  -0.4 |

a. mean-mean comparisons presented in descending order (e.g. 1-2, 1-3, 1-4; 2-1, 2-3, 2-4 etc.)

Table 4. **Using Chi Square to determine if the proportion of membership in a set of clusters of consensus when assessment of performance was unpreceded by another performance is significantly different as compared to when it was preceded by a performance with a contrasting level of competence**

|  |  |  |  |
| --- | --- | --- | --- |
| Video with clusters of consensus | Number of assessors’ comments grouped by clusters of consensus | Number of assessors’ comments grouped by clusters of consensus | Pearson Chi square (asymptotic significance) |
| Good performance | unpreceded | preceded by  Poor video | 10.4  (.016) |
| ‘excellent communication’ | 2 | 11 |
| ‘excellent interviewing’ | 4 | 3 |
| ‘omitted questions’ | 8 | 4 |
| ‘missed differentials’ | 5 | 1 |
| Borderline performance | unpreceded | preceded by  Good video | 9.5  (.009) |
| ‘detailed history’ | 10 | 1 |
| ‘open questions needed’ | 7 | 7 |
| ‘disorganized closed questions’ | 5 | 11 |
| Poor performance | unpreceded | preceded by Borderline video | 4.9  (.181) |
| ‘recognized limitations’ | 7 | 3 |
| ‘anchored to diagnosis’ | 4 | 3 |
| ‘too casual’ | 5 | 7 |
| ‘rude’ | 3 | 9 |