Objective Structured Clinical Examinations (OSCEs); Expert Clinical Examiners Decision Processes: Are they Recognition Primed and is intuition a reliable decision strategy?

Short title: Examiners OSCE decision processes

Funded by the Centre for Excellence in Professional Placement Learning (CEPPL)

A study by Peninsula Postgraduate Health Institute (PPHI) and the Faculty of Health and Social Work (FHSW), University of Plymouth (UK)

Lead participants;
- Dr Simon Cooper, Senior Lecturer, FHSW/PPHI
- Ms Gillian Blakely, Research Assistant, FHSW
- Mr Martyn Bradbury, Lecturer, FHSW
- Ms Louise Winfield, Lecturer, FHSW

The following full report will be condensed for peer review in ‘Nurse Education Today’.
Abstract

**Objective:** To identify examiners' key decision processes in Objective Clinical Examinations (OSCEs) and to measure the level of correlation between examiners' global rating and checklist scores?

**Methods:** A quantitative correlational and qualitative focus group design; including a retrospective review of 561 OSCE examination results to identify correlation levels between global rating and checklist scores. Also four focus groups with experienced clinical examiners (8–40 years of experience) to identify their OSCE experiences, opinions and decision processes.

**Results:** Global ratings were associated with a significant difference in mean station scores and we identified a strong correlation (rho 0.75) between global ratings and station scores. Findings from an ordinal regression also suggest that as the percentage station scores increase so do the global rating categories. Complex $X^2$ analysis demonstrated an overall association (p=<0.001) between global and station score categories, but results from a Wilcoxon test suggest that examiners were more likely to have allocated a higher station score than global rating (p=<0.001). In addition more experienced examiners did tend to allocate lower global ratings (p=<0.001). These findings are supported by the focus groups where examiners indicated that the marking templates tended to produce a 'higher score' than they would have allocated from 'gut reaction' and that that experience gave them confidence in allocating a range of scores including 'fails'.

The focus groups also suggest that Recognition Primed Decision (RPD) making is a good theoretical description of decision making in this situation. Examiners do claim that they make decisions without contemplation, there is some suggestion that they use their experience to pattern match against 'best practice', and that they focus on situational assessment rather than 'option' or 'optimum' assessment.

**Conclusion:** RPD making theory does appear to be a good representation of expert examiners decision processes and there is a significant association between station scores and global ratings. Global rating scales are an appropriate assessment scale and do improve the content validity, concurrent (predictive) validity, and reliability of OSCEs in this setting. Further possible developments are discussed.
Introduction and background

This study focuses on Objective Structured Clinical Examinations (OSCEs) where simulation is used to prepare and assess students for placements and practice in a simulated safe learning domain.

This work aims to develop understanding of expert clinical examiners decision processes during OSCEs. In this case short 10 minute clinical examinations for Masters level healthcare students undertaking clinical examinations. The aim is to identify (during focus groups) the key decision making processes involved in making a pass/fail decision and to review OSCE assessment documents to look for correlations between actual score and examiners ‘gut reaction’ clear pass, borderline and clear fail.

In the following background review we summarise the literature on assessment processes and decision making.

Assessment processes
To ensure safe and competent healthcare practitioners, valid and reliable methods of admission to practice are essential (Shanley 2001). Sloan et al (1995) recognise that academic written examinations assess knowledge base and that it is important to examine aspects of clinical practice such as physical examination, interpersonal and technical skills. Such skills tend to be examined through the OSCE process which may be either short (10 minute) history taking or examination skills or longer (30 minute) examinations incorporating all aspects of care, e.g. history taking, examination and handover skills; sometimes referred to as Integrated Structured Clinical Examinations (ISCEs).

The reliability and validity of OSCEs has been a primary concern of many researchers with a variety of findings. Sloan et al (1995) found that their 38 station OSCE for three levels of surgical residents had an internal reliability/consistency level of 0.91 and that it had good construct validity measured in its ability to distinguish between the differing levels of residents.

Martin and Jolly (2002) measured the predictive validity of OSCEs by rating medical students in their first year. Results showed a strong association between poor OSCE performance and future low grades in other clinical examinations, and identified positive associations between components of the examination e.g. psychiatry performance was strongly associated with the results of ‘history taking’ stations.

A-Latif (1992) found that the internal reliability/consistency of a set of OSCEs varied (0.34 to 0.78) but that it had good concurrent validity, measured as the correlation between the OSCE and a traditional clinical examination. The assessment method preferred by students varied according to the clinical component being assessed with the conclusion that there is no one superior method of assessment; multiple forms of assessment being the best form of assessment.
In an extensive review of OSCEs in the nursing profession Shanley (2001) concluded that methods of academic assessment employed in nurse training were not reliable or valid. He questioned any correlation between examination performance and standards of nursing care and argued that examinations did not replicate the ward environment.

OSCEs tend to be marked in a tick box checklist rating format for each component of the examination (see Appendix 1). However, the binary nature of checklists may overlook the more holistic components of clinical competence (Cox, 1990) with the suggestion that global ratings of performance may be an advantage (Regehr et al, 1998). Hodges and Mcllroy (2003) focussed on the employment of global rating scales in conjunction with checklists finding that the global rating scales had greater internal reliability/consistency than the checklists. Regher et al (1998) extensive evaluation of global rating scales (7 dimensions on a 5 point Likert scale), checklists, and combinations of both, found that the combination approach was most reliable and that global assessment techniques were appropriate in the OSCE setting.

Wessel et al (2003), also using a combination of scales, examined the reliability and validity of an eight station OSCE for physical therapy students. Reliability was found to be low and no significant correlations were found between OSCE scores and performance in clinical placement. They recommended a greater number of stations to improve internal reliability and predicative validity.

Hodges et al (1998) found that their OSCE had good content validity (measured as the level of ‘realism’) and found that global ratings (compared to checklists) had greater construct validity; psychiatry residents achieving higher global ratings than clinical clerks, with no difference in the checklist scores.

There is clearly evidence to suggest that global rating scales, or a combination of global and checklist scales may be a reliable and valid method of rating. In fact Hodges and McIlroy (2003) argue that the OSCE examiner is restricted by checklist marking grids.

**Decision Making**

There is no work on examiners decision processes in the OSCE setting but some comparative work by, for example, Cooper and Bond (2006), Benner (1984), Benner & Tanner (1987), Benner, Tanner & Chesla (1996), who highlight the decision processes of experts. More generally Hoffman, Donoghue & Duffield (2004) recognise that the decision process and the underlying contextual factors have to be addressed to understand the decision making process. Two distinct conceptual frameworks highlight the decision making process, one analytical and the other intuitive.

Klein (1989 a&b) remarks on the inadequacies and ambiguities of classic decision making theories in that analytical frameworks (e.g. hypothetico-deductive), which describe the generation and comparison of options based
on weighted features, fail to describe decision strategies which are time bound and where contextual flexibility is important (Klein 1989 a&b; Klein and Klinger, 1991). Consequently they refer to Naturalistic Decision Making (NDM) models to reflect how decisions are made in complex real-world settings.

One NDM model is Recognition Primed Decision Making (RPD), commonly referred to as ‘gut reaction’ (Cooper & Bond 2006; Klein 1989 a&b; Klein, Orasann & Calderwood 1991; Klein & Klinger 1991; Klein 1993; Klein 2003) which is derived from studies of the military, emergency services and aircraft cockpit crews (e.g. Klein 1989 a&b and Cannon-Bowers & Salas 1998), and is based upon evidence for pattern matching, mental simulation and mental models. Elstein (2001) argues that diagnostic problems are solved by a process of pattern matching the case to content stored in long term memory and that experience is closely tied to expert decisions. Whilst Gonzalez (2001) argues that proficient decision makers generate, evaluate and act without comparing different courses of action. The RPD making model tends to focus on situational assessment rather than ‘option’ assessment, emphasising the importance of experiences in decision making and relying on what Simon (1989) calls ‘satisficing’, or satisfactory, ‘good enough’, decisions. Such decisions are more likely when the time pressure is high, the situation is unstable and the decision maker is under pressure. All of which applies to the examination process in an OSCE where, for example, the examiner has 10 minutes to observe a student and 1 minute to rate and make a final decision about competence prior to the arrival of the next student.

From a military setting Killion (2000) argues that there is a complimentary role for analytical and recognitional models of decision making, but that the dominant operational mode is dependent on situational factors. Bond and Cooper (2006) argue that the RPD model can wrongly imply that experience equals expertise but that RPD making is clearly situationally specific. They go on to highlighting the need for situational awareness; the ability to perceive and understand the implication of environmental elements, which may be influenced by cognitive resources, intelligence level, experience and technical expertise, all moderated by the level of ‘stress’.

Hedberg and Larsson’s (2003) study of nurses’ decisions in clinical practice identified that decisions tended to be made after contemplation over comparable situations, understanding was corroborated with colleagues, and anticipated events were considered. Findings which link well with RPD theory that decisions are made through pattern matching, mental simulation and mental models. However, Hoffman, Donoghue & Duffield (2004) work on contextual factors influencing nurses’ clinical decisions found, in contrast to other studies, that education and experience had no significant effect on the variability of clinical decision making, but that, for example, level of appointment, clinical speciality and age did influence decision making variability.

In general theories such as RPD making are a good theoretical description of decision making in situationally specific time bound situations, but may be at
fault for the generalised assumption that experience equals expertise, which of course may not be the case. There may be issues of confidence in decision making without competence, and decision making based on prior poor practice, without critical analysis and reflection. However, it is likely that examiners decision making at OSCEs will be influenced by the need to make rapid time bound situationally specific ‘satisficing’ decisions, and that their performance will be influenced by their experience and ability to produce mental models from rapid pattern matching and mental simulation.

In this study we are not intending to investigate micro decision processes but to look more broadly at the key decision making processes of expert experienced clinical examiners, with reference to RPD decision making. For example, do examiners feel that they make intuitive ‘gut decisions’, what factors influence their decisions and how closely do global ratings match checklists.

Aims and Objectives

Firstly, from a set of OSCE results, we aimed to measure the correlation between a global score, clear pass, borderline and clear fail, against the checklist mark (Appendix 1). During each assessment examiners were asked to allocate a mark 0-3 to each individual assessment criteria after which they were requested to give their ‘global rating’ score before adding up the checklist score. The global rating was an additional intervention designed for this study and does not at present contribute to examination decisions; this may be reviewed based on the results of this study.

Secondly we aimed to run a number of examiner focus group interviews to identify the issues, concerns and processes involved in examination decisions, with reference to RPD making and the factors that influence decisions, for example, situational, experiential and psychosocial.

Aim: To explore expert clinical examiners ‘pass/fail’ decision processes during OSCEs.

Objectives:
- To correlate examiners global ratings of students against checklist score
- To identify examiners key decision processes

Research Questions:
- Is there a correlation between examiners ‘global ratings’ and students checklist scores?
- What are examiners key decision making processes?
Methods

Study design
A quantitative correlational and qualitative focus group design.

Methods of data collection
- Retrospective review of OSCE examination sheets to identify any correlation between global rating and checklist score.
- Demographic records of each examiner including level of experience, age, profession and place of work.
- Focus groups (Morgan 1988) with examiners; run with consenting participants at the end of OSCE examination days, with the purpose of eliciting examiners experience and identifying their decision processes. Participants were encouraged to interact with the group leader and talk to each other in addressing (focusing on) the area of interest. Interviews were audio recorded and then transcribed. Two researchers were present at each interview, one as lead and the second as a non participant observer to record group interactions.

Sample Population: All Masters level OSCE examinations and clinical examiners attending the Advanced Healthcare Practice (a Peninsula Postgraduate Health Institute Masters degree) OSCEs for first and second year modules from September 2003 to September 2006.

Recruitment: Potential participants (clinical examiners) were sent an initial introductory letter inviting them to be involved in the study at least 7 days prior to subsequent contact.

Selection: All OSCE examination stations performed within the time frame and all consenting OSCE examiners

Sample size: We aimed to draw from 9 OSCE examinations (2003-2006) with correlations drawn from a minimum of 360 OSCE assessments. From the 6 clinical examiners at each event we aimed to run 4 focus group interviews with approximately 6 participants in each group.

Exclusion criteria; OSCEs that are not completed.

Analysis
All numerical data was analysed using SPSS version 14 (Appendix 2 and 3). Descriptive and inferential statistics included; ANOVA; chi²; Spearman’s rank correlation; Wilcoxon’s T for paired tests; and ordinal regression. Where applicable the percentage checklist station scores were recoded to an ordinal format for direct comparison with the global rating scale. 95% confidence intervals are included where applicable and all tests statistical significance were two-tailed unless otherwise indicated with the level of significance at p= <0.05. (Decoster & Claypool 2004; Field 2005; Norušis 2007)
Focus group data analysis was based on Miles & Huberman’s (1984 & 1994) analysis strategy. In the first data reduction and display stage we (GB & SC) independently read and reread the transcripts (maintaining awareness of our preconceived ideas and categories). We looked for contradictions and exceptions, and differences between groups before independently identifying key codes and categories. Lastly we drew conclusions by identifying category clusters and noting relationships within the data which, following discussion, led to the development of overarching themes and sub themes.

**Ethical Approval**
The study was granted full ethical approval by the human ethics committee of the University of Plymouth based on the following criteria:

- Strict confidentiality guidelines will be adhered to throughout the study.
  - All data will be anonymised before being transcribed to an electronic medium. Written records will be destroyed within one year.
  - Participants identified within the study will be anonymised in the final published records.
- All participants will be asked to provide written informed consent to the study once their role has been defined.
- The researchers will remain open and honest to participants wherever possible.
- All participants will be informed of their right to withdraw from the study at any time.
- The researcher will endeavour to protect all participants from physical and psychological harm.
- All participants will be debriefed following the completion of the study.

**Results**
The following results are presented as quantitative findings and are followed by qualitative findings from the focus group interviews incorporating a discussion. Final conclusions are presented at the end of the report.

**Quantitative results**
We exceeded our sample size objectives with a total of 19 examiners’ results for 583 valid individual OSCE examinations based on 9 examination days attended by 107 individual students in their 1st and 2nd years of study. Retake OSCEs were also included. OSCEs included a range of systems; cardiovascular, respiratory, abdominal, central nervous system, musculo-skeletal, and mental health; including aspects such as history taking, examination and handover. Passive OSCEs (written exams) were excluded from the study. Of the 583 OSCEs we had 561 comparable global rating scores.

Of the 583 OSCEs scores ranged from 27.7% to 100.000% (mean 74.3%) (Figure 1)
A one-way Analysis of Variance (ANOVA) was performed for analysis of the difference in mean values between the global rating scores (independent variable) and the percentage station scores (dependant variable). There was a significant difference between the station score means equated to the three global rating groups (F = 431.1; p<0.001). Global ratings of ‘clear fail’ equated to a mean score of 41.6% (95% CI 0.39 – 0.44); ‘borderline’ to a mean score of 57.9% (95% CI 0.56 – 0.60); and ‘clear pass’ 81.7% (95% CI 0.81-0.83).

To enable further analysis we converted the parametric, continuous data of the percentage station scores into categories i.e. ordinal data. In the current Masters marking system 50% and above is a pass; we therefore logically assigned scores of 0-39% as ‘clear fails’; 40-59% as ‘borderline’; and 60-100% as ‘clear passes’.

Complex chi-square analysis of global scores and station score categories demonstrated that the ratings of fail, borderline and pass do differ from each other and that there is an overall association between global and station score categories (X² = 372.048, p < 0.001). However there was less of an association between ‘borderline’ and ‘fail’ ratings suggesting that examiners were more likely to have allocated a different station score to global rating in these situations.

Further analysis using Wilcoxon signed rank test for the paired sets of global and station score categories suggests that there is a difference between score
categories (z= -4.796; p= < 0.001. For (3.9%) 22 of the 561 participants their station score was less than their global rating score, for (84%) 472 participants their score was tied and for (12%) 67 out of 561 the station score was higher than the global rating score. The majority of ranks were therefore positive suggesting that examiners were more likely to have allocated a station score that was then recoded into a higher category than the global rating category, for example allocating a score of 55% (recoded into ‘borderline’) but on the global rating allocating a ‘clear fail’ mark. This fits with the above chi-square findings which indicate that although there is an overall association there is a difference in associated scores with ‘borderline’ and ‘clear fail’ students.

However for 84% of results there was concordance between the global score and the station score. This agreement is therefore a good indicator of the validity and reliability of the global rating, as there was variance in only 16%.

There was a significant correlation between the global rating scores and the ordinal station scores (rho 0.75; p=<0.001; 1 tailed), as the station scores improve there is a corresponding improvement in the global rating scale. Correlations between 0.7 and 0.9 are considered to be ‘strong’ or ‘high’ (Rowntree 1981).

Finally we ran an ordinal regression (as the dependant variable, the global rating scale, is ordinal) to identify the influence of some of the independent predictor variables. The majority of global ratings were in the highest category so the log-log link function was selected as it focuses on higher outcome categories. The predictor variables selected were ‘station scores’ and ‘examiner’s years of experience’, the latter because it was the most applicable value available for an estimate of overall ‘experience’. Both were continuous values and were entered as covariates.

Examiners years of experience did have a significant effect on the global ratings (p=<0.001). This was based on a negative coefficient suggesting that as years of experience increase the probability of being in the lower global rating category increases (i.e. more experienced examiners are more likely to allocate a lower global rating). The station score also had a significant effect on the global ratings (p=<0.001). In this case the coefficient was positive suggesting that as percentage station scores increase so does the probability of being in one of the higher categories of global ratings.
Figure 2: OSCE Performance and OSCE Decisions; findings from focus groups

- **INTRINSIC Decision Process**
  - Intuition (‘gut’) (Performance without contemplation)
  - Examiner Decision

- **EXTRINSIC Decision Processes**
  - Student Clinical competence
  - Student Communication ability
  - Student Examination approach (style, confidence)
Focus group findings and discussion

During the focus groups key questions and prompts focussed on an exploration of the decision process in OSCEs; experience in relation decisions; decision confidence and competence; and key factors effecting decision making (Appendix 4). Pre existing peer and friendship groups were apparent within each group but did not appear to overtly influence the discussion as the researchers made a conscious effort to draw in those with fewer opinions (Kevern & Webb 2001). In fact the groups flowed particularly well with opinions and attitudes expressed clearly and enthusiastically. Having completed the independent phase of analysis the researchers found that they had identified very similar themes which were discussed and adapted into the following higher level abstract analysis expressed diagrammatically in the framework illustrated in Figure 2. In the following section quotes are identified from their source, e.g. focus group 1 = FG1. (Appendix 5 includes a full list of themed quotes).

Central to our framework was the influences on examiners decision making. Findings from both the quantitative and focus group results suggest that there are both intrinsic decision processes, in the form of examiner experience and intuition; and extrinsic influences, in the form of student OSCE performance, their clinical competence, communication ability and examination approach (e.g. their style and confidence).

Examiners decisions appeared to be more challenging when candidates were considered to be borderline. They felt that it is easier to make a decision about the very good and very bad than the mediocre;
‘…very good and very bad usually stand out…it’s those mediocre, when you’ve got to sort of tick boxes’ (FG1).

A repeated theme throughout was examiners belief that students must demonstrate holistic in depth understanding of care and that it was often apparent when knowledge and skills were superficial
‘… we are looking for someone to have that flow and understanding of what they’re doing not trying to get out something that they’ve learnt late’ (FG3)

There was almost universal support for the introduction of some form of global rating scale to the degree than some felt it should play a predominant role;
‘…from a clinical perspective your global rating should be able to override what’s coming out of this objective test’ (FG3)

Or that they should be able;
‘…use the sheet less, or use it as an adjunct and not as your main’ (FG4)

and that it would be;
‘….really useful to have the global aspect there, that allows for flare…sparkling ones stand out…head and shoulders above the others’ (FG4)

However there was the expressed concern that it would be a;
‘…risk that the OSCEs loose their objectiveness if we start introducing personal feelings’ (FG4)

From the quantitative results we found that examiners were more likely to have allocated a higher station score than global rating. An issue which was raised at times in the focus groups, usually in the form of frustration that students achieved a higher than expected station score, compared to examiners ‘gut feelings’ about their performance. An issue which was usually blamed on an ‘inappropriate emphasis, or lack of weighting, for elements within the scoring template;
‘…the scoring system… may not always be correct’ (FG2)

And
‘…sometimes the OSCE marks seems unfair to your gut reaction’ (FG3)

Examiners often referred to their level of experience as being an essential element in the decision process, for example in the following analogy;
‘…like if you are sitting in a car and somebody you consider as a good driver or not, you get that feeling don’t you? But actually what you are experiencing is all the other times you have sat in the car’ (FG1).

Also that experience and age are intertwined
‘…it’s all experience …you go through life…you get older, you change your approach…you’re more tolerant of something and they are tolerant of others’ (FG2).

And that experience improves insight in the examination setting;
‘…when you have done lots of examining you get so used to knowing who is in their stride and who is not’ (FG3).

An issued that was indirectly supported in the quantitative findings which indicated that more experienced examiners tended to allocate lower global ratings.

Experience also links well with theories such as Recognition Primed Decision making, or intuition, as one respondent indicated;
‘…your experience gives you the competence and confidence…because of all the things you have integrated and learned beforehand…you do it without constantly thinking that what you are doing’ (FG1)

Quantitative results supported the notion that marking templates are highly correlated with global ratings of performance; and intuition and intuitive decision making were frequently cited as having a relevant, valid and reliable place in the OSCE examination process;
‘I ask myself, my gut - would I trust this person to see my patient and to come back to me with their findings’ (FG3)

‘…I don’t think solely objective is enough, the potential for endangering a future if we stick with solely objective methods’ (FG3)
'I have got somebody who ticks the boxes for the examination, but your gut reaction is ‘I wouldn’t really want to go there, or I wouldn’t want to take them to see my child’ (FG1)

‘…would you want this person to be your doctor? I think that’s really, a good question’ (FG2)

There was also the argument that gut feeling is ‘…an emotional intelligence’ (FG1) and that it is a valid tool as it is wrapped up in more objective measures of decision making;
‘It’s the separation of things into the subjective, the objective, you know …. it’s separating cognition from…emotion…you don’t….you know they are all integrated  and that’s why the gut feeling is there because its integrated’ (FG1).

However there was the argument that intuitive decisions can appear to be an invalid measure if they are directed at skills that can be easily measured in a quantitative measure. In other words intuitive marking scales should be used for rating softer skills such as interpersonal relationships;
‘…that lack of validity …from a gut reaction, got very much more to do with interpersonal skills than the skill you are looking at’ (FG1)

The need for students to demonstrate a level of clinical competence and the influences and ways that this should be measured were frequently cited. For example competence was described as;
‘…this person didn’t stop to think what am I doing here they just seemed to know exactly what to do, why they were doing it’ (FG3)

And measured as clinical skill performance and interpersonal skills;
‘…not competent because they didn’t interact with the patient’ (FG2)

And
‘ascertain information from the patient…they get the patient to interact…that would make me feel that they were competent’ (FG2)

‘I think you can see somebody’s clearly clinically competent, but you wouldn’t necessarily want them to care for you’ (FG2)

Whilst respondents expressed concerns that the exam process and the scoring schedules could be used to demonstrate competence;
‘…although they tick the boxes, I didn’t feel that if I was a patient …if I saw them in practice I wouldn’t feel competent’ (FG1)

‘I became very disappointed…my global score was borderline…but they appeared to have ticked all the boxes, and I thought, now this student is going to pass this station on the tick box’ (FG1)

‘…sceptical …of the degree to which competence could be assessed in 10 minutes’ (FG2).
However there was also the expressed fear that confidence can override examiners judgement of competence;
‘…if they are confident you know your sort of radar is less really, because you’re thinking they are good’ (FG4)

There was a general feeling that OSCEs did not necessarily live up to expectations for the measuring of clinical skills and that there were ways of improving the process. For example the need to question students to ascertain their clinical insight;
‘…[they] don’t score 2 points by putting a stethoscope on and just standing there looking at the ceiling’ (FG4);

and that there was a danger that an examination was seen as a demonstration of clinical competence;
‘…you don’t want this as a tick and people say oh I passed that examination, so I am competent at that…’ (FG4)

**Student communication ability** is an overlapping theme that was frequently cited as being an influence on intuitive ratings, measures of competence, and final examiners decisions. Respondents felt that ratings were dependant on the;
‘…way they relate to the patient, and I think for me, patient focus and they interact with the patient and everything’ (FG2)

And that poor communication skills would influence performance ratings;
‘…her communication skills were ones that very much push me into a position of thinking “I’m not quite sure whether I’d fully want to see her’ (FG1)

Interestingly the students’ general **examination approach** their ‘flare and clinical acumen’ (FG4) played a part in examiners decision processes;
‘…taking a professional’s approach and that’s certainly what initially I think you are looking at’ (FG2);

*How they present themselves, how they come in, how they introduce themselves and their general demeanour*’ (FG2)

‘…their confidence level, the way that they, their deportment’ (FG4)

Respondents also expressed their personal likes and dislikes, usually as a level of awareness, not as an influence. However we were left with the impression that such issues did play a part in their decision processes;
‘…brilliantly systematic…quite quickly I found myself irritated because every other word she said “my dear”, “my dear” ’(FG2)

‘…my gut reaction, I think OK, don’t like them very much but they’re doing what they should be doing’ (FG1)

‘…his shoes were a disgrace, hanging off his feet and filthy dirty’ (FG4)
There was also the view that the process was a game that should be complied with;
‘...it’s a game isn’t it, ultimately they know the rules and should be playing their part in that game’ (FG4)

Conclusions
This study aimed to identify examiners key decision processes and to see if there was a correlation between examiners global ratings and checklist scores?

Our findings are based on a sample of 19 examiners and 583 OSCEs, covering a range of skills, and with a range of scores from 28% to 100%. Global ratings were associated with a significant difference in mean station scores and importantly we identified a strong correlation ($\rho 0.75$) between global ratings and stations scores. Findings from the ordinal regression also suggest that as the percentage station scores increase so do the global rating categories.

Complex $X^2$ analysis demonstrated an overall association ($p<0.001$) between global and station score categories, but results from the Wilcoxon test suggests that examiners were more likely to have allocated a higher station score than global rating ($p<0.001$). Finally more experienced examiners did tend to allocate lower global ratings ($p<0.001$).

In conclusion, there is a significant association between station scores and global ratings. However, there was a tendency to mark stations higher than the global ratings and more experienced examiners tended to give lower global ratings. These findings are supported by the focus groups where examiners indicated that the marking templates tended to produce a ‘higher score’ than they would have allocated from ‘gut reaction’ and that experience gave them confidence in allocating a range of global scores including ‘fails’.

The focus groups also suggest that Recognition Primed Decision making is a good theoretical description of decision making in this situation. Examiners do claim that they make decisions without contemplation, there is some suggestion that they use their experience to pattern match against ‘best practice’, and that they focus on situational assessment rather than ‘option’ or ‘optimum’ assessment.

Validation of OSCEs varies with a lack of testing of some examinations (Shanley 2001) and rigorous assessment of reliability, validity and feasibility by others (Hodges et al 1998). From this study we have been able to demonstrate that global rating scales are a feasible assessment scale and that they enhance the content validity of the assessment, demonstrated by expert examiners positive views on their inclusion and with the tentative argument that validity is increased by the close correlation with check box station scores, but assuming that the check box ratings are valid. Concurrent (predictive) validity is improved as outcomes are likely to be correctly predicted when measured in a different way (Hodges et al 1998) i.e. in this
case the inclusion of a global scale is closely correlated with station scores. We would also argue that the incorporation of a global rating scale enhances the reliability of the assessment as it improves the likelihood that the ‘right’ candidate will be passed or failed. In other words it creates an additional check and balance, for example it may be used to ‘balance’ the higher station scores.

Despite the close correlation and significant association between the scales there were differences, for example there was a tendency to mark stations higher than the global ratings and more experienced examiners tended to give lower global ratings, which may not be the case in the more ‘objective’ station scoring. The suggestion therefore is that the inclusion of a global rating scale may be valid but that the accuracy or reliability of the assessment varies dependant on the scale used. Inclusion of both may therefore create a more reliable and balanced assessment a view supported by Regher et al (1998) who found that a combination approach was most reliable and that global assessment techniques were appropriate in the OSCE setting.

The study is limited however by a number of factors. We cannot be sure that examiners always made a global rating decision before adding up their station scores, which may have had an influence on their final ratings. We also studied only Masters level OSCEs of emerging advanced practice students, across a range of skills, with a limited ‘examiner’ sample. For these reasons we cannot make a judgement of how feasible, reliable or valid the use of global ratings would be in other examination settings.

Based on these results we would suggest that the additional inclusion of a global rating scale is appropriate in the OSCE setting. We intend to use it as a weighted decision tool that will influence station scores. Table 1 lays out some of the weighting possibilities, for example if the examiner in their global rating ticks the clear fail box, students will loose 10% of their marks from their station score, or if examiners tick the clear pass box, 10% will be added to the station score. Alternatively the global rating could be used as a grading tool rather than as a scoring tool. For example a station score could be capped within a grade e.g. if the examiner assigns a global borderline rating a stations score of 84% could be capped at 59%?
Table 1: Possible weightings for global assessment?

<table>
<thead>
<tr>
<th></th>
<th>Clear Fail</th>
<th>Borderline</th>
<th>Clear Pass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Option 1</td>
<td>-6</td>
<td>0</td>
<td>+6</td>
</tr>
<tr>
<td>Option 2*</td>
<td>20%</td>
<td>0%</td>
<td>+20%</td>
</tr>
<tr>
<td>Option 3*</td>
<td>If global ‘fail’ and not 0-40% station score negative 20%</td>
<td>If global ‘borderline’ and not 40-60% station score positive 20% for &lt;40% negative 20% for &gt;60%</td>
<td>If global ‘pass’ and not 60-100% station score positive 20%</td>
</tr>
</tbody>
</table>

*20% weighting included as a ‘reasonable and fair’ estimate of the value of qualitative judgements (from the results of this paper).

We are also considering respondent examiners suggestions for the inclusion of standardised prompts during OSCEs for enhanced student competence rating, for example when auscultating the chest the prompt would be ‘what are your listening for?’. Alternatively the student could be prompted to talk through everything they are doing? The balance of course is to ensure that questioning or ‘talking through’ does not become too distracting for the students.

Acknowledgements: Thanks to all the clinical examiners that participated in this study.

Funding: Centre for Excellence in Practice Placement Learning (CEPPL). University of Plymouth
References


Cox K. (1990) No Oscar for OSCE. Medical Education 24 (6) 540-545


Hodges B. and McIlroy J.H. (2003) Analytical global OSCE ratings are sensitive to level of training. Medical Education 37 (11), 1012-1016


Appendices

Appendix 1: Example OSCE showing scoring template and global rating

**Date:**

**Clinical Examiners name:**

**Academic item verifier name:**

<table>
<thead>
<tr>
<th></th>
<th>Not performed or incompetent</th>
<th>Performed but not fully proficient</th>
<th>Performed competently</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation of patient for examination</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Washes hands <em>(with verbal explanation of Ayeliffe technique)</em></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Positioning of patient</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>General Inspection</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Applies oxygen</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Palpation – radial pulse</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Blood pressure measurement</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>JVP Inspection</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Palpation – anterior chest wall</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Auscultation of heart sounds</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Auscultation of chest</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Peripheral pulses and oedema</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Closing explanation given to patient</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Global rating**

- Clear Fail
- Borderline
- Clear Pass

**Station total**

Divided by maximum possible score 24

\[ \times 100 = \% \]
APPENDIX 2 SPSS DATABASE

SPSS Final List

APPENDIX 3 SPSS OUTPUT

Total Output of Results
Appendix 4 – Topic guide OSCE assessment

<table>
<thead>
<tr>
<th>Welcome, putting people at ease</th>
</tr>
</thead>
</table>

**Short description of the interview’s purpose**

‘The purpose of these focus groups is to determine and explore examiners decision making processes’

**Explain the focus group concept**

- the interest you have in everyone’s views
- only one person speaks at a time
- interested in open and honest views
- will act to stop personal comments
- process is time-limited so may have to curtail a discussion in order to achieve the research aims

**Briefly cover consent**

**Explore decision making processes**

<table>
<thead>
<tr>
<th>Prompts as required (see below)</th>
<th>Both micro – the rating scale AND Macro – overall judgement (pass/fail) gut reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience in relation to their decisions</td>
<td>How does their experience effect their decisions?</td>
</tr>
<tr>
<td>Decision Confidence</td>
<td></td>
</tr>
<tr>
<td>Decision Competence</td>
<td></td>
</tr>
<tr>
<td>Factors influencing decisions</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 5 – Key focus group themed quotes (Focus group transcripts available on request).

Examiners decisions
‘…very good and very bad usually stand out…it’s those mediocre, when you’ve got to sort of tick boxes’ (FG1)

‘…the scoring system… may not always be correct’ (FG2)

‘…would be happy to use a global gut reaction assessment as a pass/fail decision tool? Yes? [Q: Interviewer] ‘I think that plays a part, yes’ [A: respondent] (FG3)

‘…from a clinical perspective your global rating should be able to override what’s coming out of this objective test’ (FG3)

‘… we are looking for someone to have that flow and understanding of what they’re doing not trying to get out something that they’ve learnt late’ (FG3)

‘…use the sheet less, or use it as an adjunct and not as your main’ (FG4)

‘…risk that the OSCEs lose their objectiveness if we start introducing personal feelings’ (FG4)

‘really useful to have the global aspect there, that allows for flare…sparkling ones stand out…head and shoulders above the others’ (FG4)

Examiners experience
‘…like if you are sitting in a car and somebody you consider as a good driver or not, you get that feeling don’t you? But actually what you are experiencing is all the other times you have sat in the car’ (FG1)

‘one of my feelings that globally is not purely plucked out of the air, it is actually a, a mass of all our own experiences and expertise within our area’ (FG1)

‘…your experience gives you the competence and confidence...because of all the things you have integrated and learned beforehand...you do it without constantly thinking that what you are doing’ (FG1)

‘..internalised checklist that I think I would probably have…I would have in mind’ (FG2)

‘…it’s all experience ...you go through life...you get older, you change your approach...you’re more tolerant of something and they are tolerant of others’ (FG2)

‘...when you have done lots of examining you get so used to knowing who is in their stride and who is not’ (FG3)
**Intuition ‘gut’**

‘…that lack of validity …from a gut reaction, got very much more to do with interpersonal skills than the skill you are looking at’ (FG1)

‘It’s the separation of things into the subjective, the objective, ye know it’s separating cognition from…emotion…you don’t ye know they are all integrated and that’s why the gut feeling …is there because it’ integrated’ (FG1)

‘I have got somebody who ticks the boxes for the examination, but your gut reaction is ‘I wouldn’t really want to go there, or I wouldn’t want to take them to see my child’ (FG1)

‘your gut feeling is an emotional intelligence’ (FG1)

‘… I have been constantly…watch what’s going on…I would like to think that it was intuition, but maybe it’s not, maybe it’s clinical, what you have learnt over the years’ (FG2)

‘…would you want this person to be your doctor? I think that’s really, a good question’ (FG2)

‘…sometimes the OSCE marks seems unfair to your gut reaction’ (FG3)

‘…without predisposition what you get is an instant expectation’ (FG3)

‘…you get an expectation and then they either live up to that, they exceed it or they are below it’ (FG3)

‘I ask myself, my gut would I trust this person to see my patient and to come back to me with their findings’ (FG3)

‘…I don’t think solely objective is enough, the potential for endangering a future if we stick with solely objective methods’ (FG3)

**Student clinical competence**

‘…although they tick the boxes, I didn’t feel that if I was a patient …I saw them in practice I wouldn’t feel competent’ (FG1)

‘I became very disappointed…my global score was borderline…but they appeared to have ticked all the boxes, and I thought, now this student is going to pass this station on the tick box’ (FG1)

‘…making assumptions rather than hearing the story’ (FG2)

‘…not competent because they didn’t interact with the patient’ (FG2)

‘ascertain information from the patient…they get the patient to interact…that would make me feel that they were competent’ (FG2)
‘…sceptical …of the degree to which competence …[]..could be assessed in 10 minutes’ (FG2)

‘I think you can see somebody’s clearly clinically competent, but you wouldn’t necessarily want them to care for you’ (FG2)

‘…this person didn’t stop to think what am I doing here?…just seemed to know exactly what to do, why they were doing it’ (FG3)

‘…you don’t want this as a tick and people say oh I passed that examination, so I am competent at that…’ (FG4)

‘…[they] don’t score 2 points by putting a stethoscope on and just standing there looking at the ceiling’ (FG4)

**Student communication ability**
‘…her communication skills were ones that very much push me into a position of thinking “I’m not quite sure whether I’d fully want to see her’ (FG1)

‘…way they relate to the patient, and I think for me, patient focus and they interact with the patient and everything’ (FG2)

**Student exam approach**
‘my gut reaction, I think OK, don't like them very much but they’re doing what they should be doing’ (FG1)

‘…brilliantly systematic…quite quickly I found myself irritated because every other word she said “my dear”, “my dear” ’(FG2)

‘…taking in a professional’s approach and that’s certainly what initially I think …you are looking at’ (FG2)

How they present themselves, how they come in, how they introduce themselves and their general demeanour’ (FG2)

‘…choices…practitioner makes…what they say, what they do, what they ask, how they use themselves…a repertoire’ (FG2)

‘…their confidence level, the way that they, their deportment’ (FG4)

‘…it’s a game isn’t it ultimately , they know the rules and should be playing their part in that game’ (FG4)

‘…if they are confident…[ ]...you know your sort’of radar is less, really because you’re thinking they are good’ (FG4)

‘…his shoes were a disgrace, hanging off his feet and filthy dirty’ (FG4)

‘got no flare or clinical acumen’ (FG4)