OSCEs revisited: updates on developments in evidence-based best practice

Japan Medical Education Foundation (JMEF)

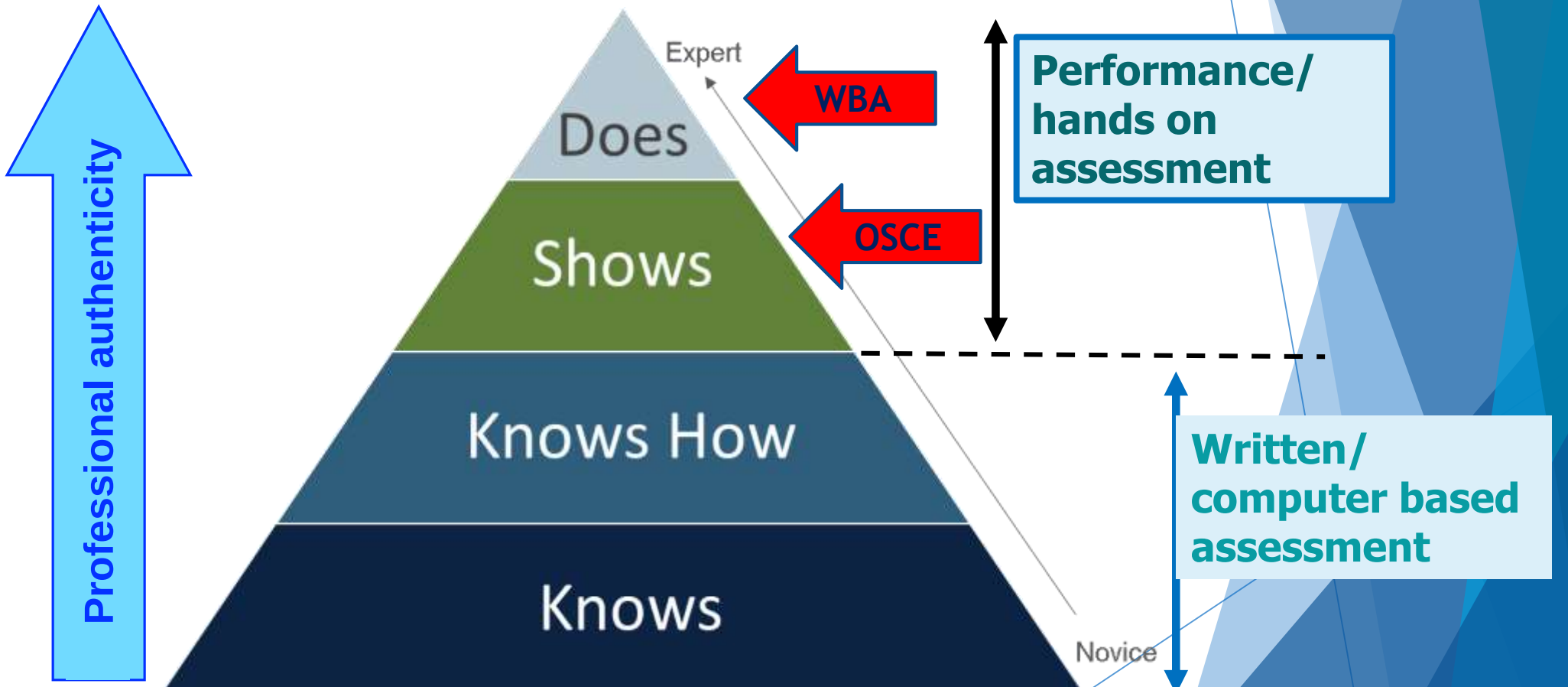
Katharine Boursicot
BSc, MBBS, MRCOG, MAHPE, NTF, SFHEA
Director
Health Professional Assessment Consultancy
Singapore

Review of OSCE pedagogy

How have OSCEs changed since 1979?

1. Revisiting fundamental **principles**
2. Widespread use of ‘OSCEs’….BUT some issues have arisen which have stained the reputation of OSCEs…**“Problematic” OSCEs**
3. Evolution of OSCEs: current **best practices** based on <u>evidence in the literature</u>

Place of OSCEs in educational model of clinical competence



Miller GE The assessment of clinical skills/competence/performance Academic Medicine (Supplement) 1990; 65: S63- S67

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

OSCEs principles revisited

HARDEN, R.M. and GLEESON, F.A. (1979), Assessment of clinical competence using an objective structured clinical examination (OSCE). Medical Education, 13: 39-54

OSCE were designed to test clinical and communication skills

Fundamental principles:

the underlying theoretical pedagogy

- All candidates presented with the same test
- Careful specification of content & standard expected
- Observation of wide sample of clinical activities
- Observation by large number of different examiners
- Structured interaction between examiner and student

A series of observations about a candidate = **fair decision**

- Designed purposively
- Valid & reliable
- Properly mapped
- Feasible
- Defensible
- Acceptable

"Problematic" OSCEs: suboptimal practices which undermine the usefulness of using the OSCE format

- Poorly blueprinted OSCE:
  - insufficient spread of skills being teste – **in ent mpling = poor validity**
  - not mapped to learning outcomes - **p**
  - too few stations (less thar **poo**
- MCQs/ SAQs/vivas disguise
- Poorly constructed s **lidity and reliability**
- Inconsistent examiner beh
- Inconsistency in co ner, SP, equipment, timing) - **poor reliability**
- No proper standard setting - **po alidity**
- No analysis of test performance - **poor validity**

OSCE pedagogy ignored / forgotten

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

OSCEs principles: developments

Fundamental principles still hold

PLUS new evidence & best practice

- Sufficient sampling across different tasks: task specificity
- Sufficient time for each task to be performed authentically

- OSCE must be mapped to course Learning Objectives: **blueprint**
- Sufficient **sampling** of LOs (12-16 stations; 2-3hrs testing time)
- Requires **demonstration** of clinical and communication skills by candidates
- **Observation** and scoring of performance by **expert** (trained) examiners

Linn, R.L. and Burton, E. (1994), Performance-Based Assessment: Implications of Task Specificity. Educational Measurement: Issues and Practice, 13:5-8 https://doi.org/10.1111/j.1745-3992.1994.tb00778.x

Example OSCE blueprint

Y axis
Balance of system/ curriculum time

X axis
Balance of types of stations across domains

	History	Explain	Exam	Procedure
CVS	Chest pain	Disch drugs	Cardiac	BP
RS	Haemoptysis	Smoking	Respiratory	Peak flow
GIS	Abdo pain	Gastros	Abdo	PR
Repro	Amenorrhea	Abnormal smear	Cx smear	
NS	Headache		Eyes	
MS		Diag RA	Hip	
Gene-ric	Pre-op assess	Consent for PM		IV cannulation Blood transfusion

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Sampling

Balance of system/ curriculum time

Balance of types of stations

	History 5	Explain 2	Exam 5	Procedure 3
Cardiovascular system 4	Chest pain	Discharge drugs	Cardiac	Blood pressure
Respiratory system 3	ough		Respiratory	Peak flow
Gastro Intestinal system 2	Abdominal ain	Gastritis	Abdomen	PR
Neurological 2	eadache	Brain tumour	Cranial nerves	
Musculoskeletal system 2	ack pain		Hip	
Generic 2	e-op assess	Consent for PM		IV cann

Total – 15 stations

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Timing	History 12 mins (5)	Explain 15 mins (2)	Exam 15 mins (5)	Procedure 7 mins (3)
Cardiovascular system (4)	Chest pain	Discharge drugs	Cardiac	Blood pressure
Respiratory system (3)	Cough		Respiratory	Peak flow
Gastro Intestinal system (2)	Abdominal pain	Gastritis	Abdomen	PR
Neurological (2)	Headache	Brain tumour	Cranial nerves	
Musculoskeletal system (2)	Back pain		Hip	
Generic (2)	Pre-op assess	Consent for PM		IV cann
	60 mins	**30 mins**	**85 mins**	**21 mins**

Total:
15 stations

Total testing time:
196 mins

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Performance Assessment: Consensus Statement and Recommendations from the 2020 Ottawa conference

- Major review of the literature on performance assessment
- Over 5,500 words
- 113 references
- Recent and most up-to-date

Boursicot K, Kemp S, Wilkinson T, Findyartini A, Canning C, Cilliers F, & Fuller R (2021)

Performance assessment: consensus statement and recommendations from the 2020 Ottawa Conference

Medical Teacher, 43(1), 58-67. https://doi.org/10.1080/0142159X.2020.1830052

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Recommendations: 1

1. Define **purpose** of the OSCE and make purpose explicit to all stakeholders
2. **Blueprint** to learning objectives/outcomes
3. Assess **clinical interactions**
4. Plan **adequate sampling** with sufficient stations and testing time
5. Design **marking schemes** to align with clinical task and clinical thinking (rating scales/checklists)
6. Use OSCE-specific criterion referenced **standard setting** (Borderline Regression Method)

Katharine Boursicot, Luci Etheridge, Zeryab Setna, Alison Sturrock, Jean Ker, Sydney Smee & Elango Sambandam (2011) Performance in assessment: Consensus statement and recommendations from the Ottawa conference, Medical Teacher, 33:5, 370-383, DOI: 10.3109/0142159X.2011.565831

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Recommendations: 2

New aspects

1. Ensure OSCEs part of a **system of assessment**
2. Adhere to **validity** framework criteria
3. Generate **metrics for OSCEs**
4. Value examiner diversity and focus **examiner training** on conduct, behaviours and bias
5. Handle **test security** through task design and circuit design to group stations
6. **Triangulate data** from OSCE performance with other assessments or outcomes, to inform decision making

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Implications for best practice

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of rating scales, use of mixed format scoring schemes
- More 'authentic' stations: longer, as appropriate to the task (not just 5 mins, up to 15 mins)
- More authentic tasks: including clinical decision-making, formulating differential diagnoses, suggesting next management steps More
- Formal training of expert clinician examiners
- Formal OSCE-specific evidence-based standard setting (Borderline Regression Method) +/= MNSP (minimum number of stations to pass)
- Post test OSCE specific metrics
- Test security
- Technology

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Developments in OSCE best practice for the 2020s

- **Scoring schemes: use of checklists, rating scales, use of mixed format scoring schemes**

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Designing scoring instruments

Assessment for:
- Learning?
- Progression?
- Numbers?
- Grades?
- Words?
- Combination?

- Background
- Familiarity with test
- Training
- Seniority
- Cognitive load

Purpose

Other considerations

- Environment
- Cognitive load
- Technology
- Weighting
- Standardisation

Assessors

- **Learners**

- Task
- Expected level
- Novice-Master-Expert?
- Feedback opportunity?

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

OSCE scoring schemes

"Thoroughness items performed without thinking do not reflect clinical reasoning ability and contribute construct-irrelevant variance to scores."
Yudkowsky et al, 2014

Checklist

- Process (task) focused or novice routine
- **Focus on Routine: Novice Candidate**
- Novice or non clinical assessor

Hybrid

- Mixed tasks – particularly if safety focus
- **Acquiring Mastery Intermediate level**
- Broad range of assessors

Domain-based rating scales

- Affective & Behavioural
- **Focus on outcome & complexity Expert candidate**
- Expert (clinical) assessor

Ilgen J, Ma, I, Hatala, R, Cook, D (2015) Are rating scales A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment, Medical Education, Vol49, Issue 2, p 161-173, https://doi.org/10.1111/medu.12621

Timothy J. Wood & Debra Pugh (2020) Are rating scales really better than checklists for measuring increasing levels of expertise?, Medical Teacher, 42:1, 46-51, DOI: 10.1080/0142159X.2019.1652260

Ilgen J, Ma, I, Hatala, R, Cook, D (2015) Are rating scales A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment, Medical Education, Vol49, Issue 2, p 161-173, https://doi.org/10.1111/medu.12621

Conclusions:

1. **Checklist** inter-rater reliability and trainee discrimination were more favourable than suggested in earlier work
   - but each task requires a separate checklist
2. **(Global Rating Scales) GRS**
   - have higher average inter-item and inter-station reliability
   - can be used across multiple tasks
   - may better capture nuanced elements of expertise

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of checklists, rating scales, use of mixed format scoring schemes
- **More 'authentic' stations: longer,** as appropriate to the task (not just 5 mins, up to 15 mins)
- **More authentic tasks:** including
  - clinical decision-making
  - formulating differential diagnoses
  - suggesting next management steps
  - Interpretation of data (related to patient)

Designing more authentic stations & tasks

- Longer more authentic stations to reflect real clinical practice
  - e.g. 10 mins to take a history + 5 mins
    - to summarise
    - suggest differential diagnosis
    - management plans
  - e.g. 10 mins examine patient + 5 mins explain findings to patient
  - e.g. 10 mins to take history + 5 mins to summarise /check with patient and explain nest steps
- tests not just history taking skills but also **clinical decision-making**, formulating **differential diagnoses**, suggesting next **management** steps

Daniels, V. J., & Pugh, D. (2018). Twelve tips for developing an OSCE that measures what you want. Medical teacher, 40(12), 1208–1213. https://doi.org/10.1080/0142159X.2017.1390214

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of rating scales, use of mixed format scoring schemes
- More 'authentic' stations: longer, as appropriate to the task (not just 5 mins, up to 15 mins)
- More authentic tasks: including clinical decision-making, formulating differential diagnoses, suggesting next management steps More
- **Formal training of expert clinician examiners**

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

OSCE examiner training

- Early paper:
- Really set the ball rolling
  - for the move to rating scales from checklists
  - formal examiner training
- "Rater cognition" literature
  - Understanding how raters make judgements
  - Recognising bias
- Numerous papers on examiner training…..latest examples
- Conclusion: **it works**
- if done properly

Hodges, B., & McIlroy, J. H. (2003, Nov).
Analytic global OSCE ratings are sensitive to level of training.
Medical Education, 37(11), 1012-1016.
https://doi.org/10.1046/j.1365-2923.2003.01674.x

Pell, G, Homer, M, Roberts, T (2008) Assessor training: its effects on criterion - based assessment in a medical context International Journal of Research & Method in Education, Vol 31, Issue 2, pages 143-154

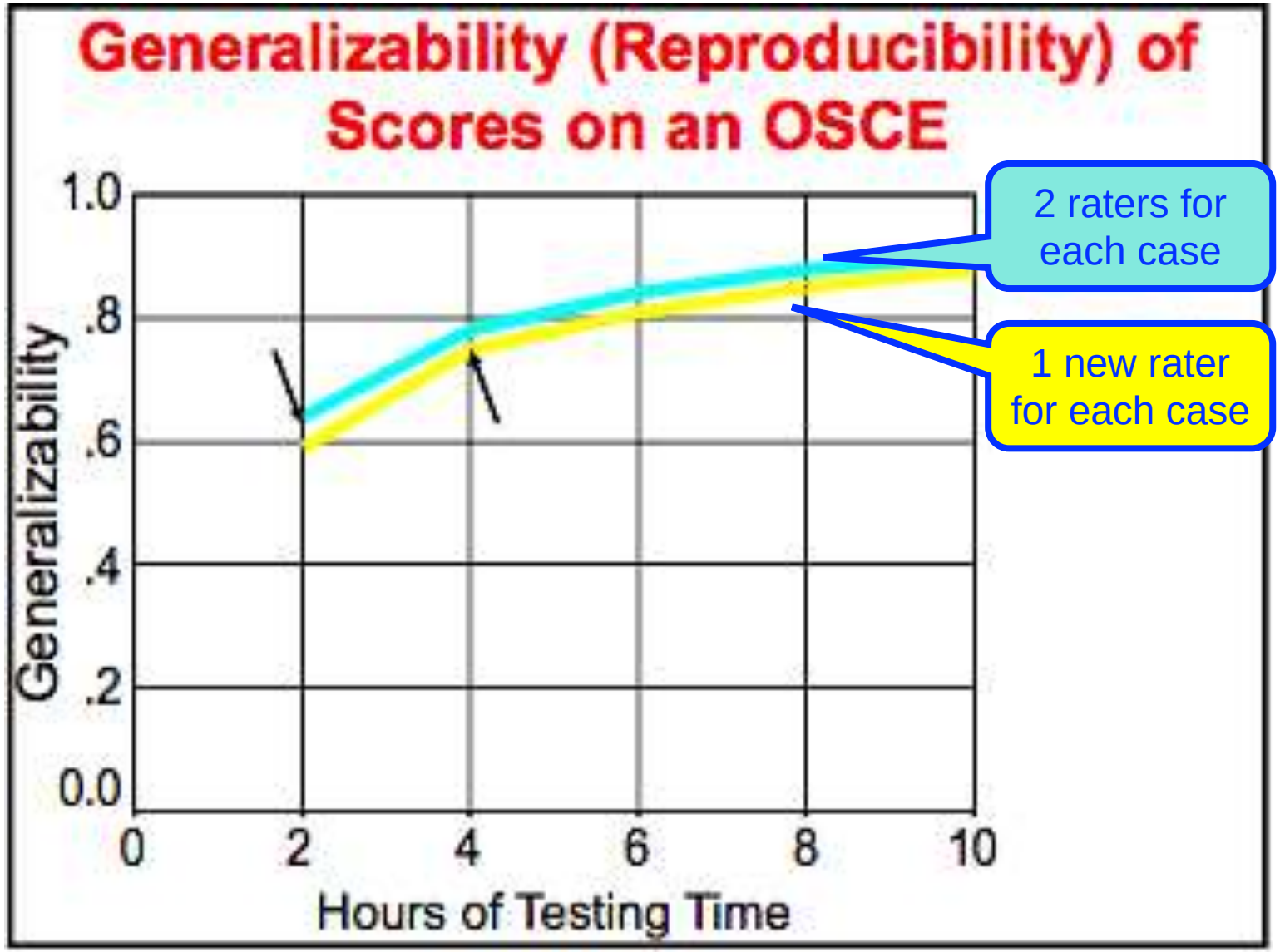
Schüttpelz-Brauns, K., Nühse, K., Strohmer, R., & Kaden, J. J. (2019). Training OSCE: minimal effort with far-reaching results. Medical Education, 53(11), 1153-1154. https://doi.org/10.1111/medu.13970

Moreno-López, R., & Sinclair, S. (2020). Evaluation of a new e-learning resource for calibrating OSCE examiners on the use of rating scales. European journal of dental education : official journal of the Association for Dental Education in Europe, 24(2), 276–281. https://doi.org/10.1111/eje.12495

Wong, A, Roberts, C, Thistlethwaite, J (2020)
Impact of Structured Feedback on Examiner Judgements in Objective Structured Clinical Examinations (OSCEs) Using Generalisability Theory
Health Professions Education, Vol 6, issue 2, pages 271-281

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Generalisability/ Reliability



Effect of more than 1 examiner: *marginal gain:* approximately 0.02

Reliability is increased by

- More testing time
- More stations

- **No** need for 2 examiners per station
- Better to have **more stations** with 1 examiner per station
- Can easily check on consistency of examiner scoring across circuits/sites with software packages/OSCE metrics

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of rating scales, use of mixed format scoring schemes
- More 'authentic' stations: longer, as appropriate to the task (not just 5 mins, up to 15 mins)
- More authentic tasks: including clinical decision-making, formulating differential diagnoses, suggesting next management steps More
- Formal training of expert clinician examiners
- **Formal OSCE-specific evidence-based standard setting (Borderline Regression Method) +/= MNSP (minimum number of stations to pass)**

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Standard setting: making pass/fail decisions

- Recognised criterion referenced methods designed specifically for OSCEs: Borderline Regression Method
  - appropriate to size of candidate cohort (more than 50)
    - Otherwise have to use Angoff
  - +/-MNSP (minimum number of stations to pass)

Naveed Yousuf, Claudio Violato & Rukhsana W. Zuberi (2015) Standard Setting Methods for Pass/Fail Decisions on High-Stakes Objective Structured Clinical Examinations: A Validity Study, Teaching and Learning in Medicine, 27:3, 280-291, DOI: 10.1080/10401334.2015.1044749

Matt Homer, Richard Fuller, Jennifer Hallam & Godfrey Pell (2020) Setting defensible standards in small cohort OSCEs: Understanding better when borderline regression can 'work', Medical Teacher, 42:3, 306-315, DOI: 10.1080/0142159X.2019.1681388

Matt Homer (2023): Setting defensible minimum-stations-passed standards in OSCE-type assessments, Medical Teacher, DOI: 10.1080/0142159X.2023.2197138

Matt Homer & Jen Russell (2021) Conjunctive standards in OSCEs: The why and the how of number of stations passed criteria, Medical Teacher, 43:4, 448-455, DOI: 10.1080/0142159X.2020.1856353

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of rating scales, use of mixed format scoring schemes
- More 'authentic' stations: longer, as appropriate to the task (not just 5 mins, up to 15 mins)
- More authentic tasks: including clinical decision-making, formulating differential diagnoses, suggesting next management steps More
- Formal training of expert clinician examiners
- Formal OSCE-specific evidence-based standard setting (Borderline Regression Method) +/= MNSP (minimum number of stations to pass)
- **Post test OSCE specific metrics**

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Post test OSCE specific metrics

Quality Assurance (QA) requires the scrutiny of the OSCE results using psychometrics specifically designed for OSCEs

- whole test
- station level

1) Coefficient of determination R2
2) Inter-grade discrimination
3) Number of failures
4) Between-group variation (assessors)
5) Between-group variance (other factors)
6) SP ratings

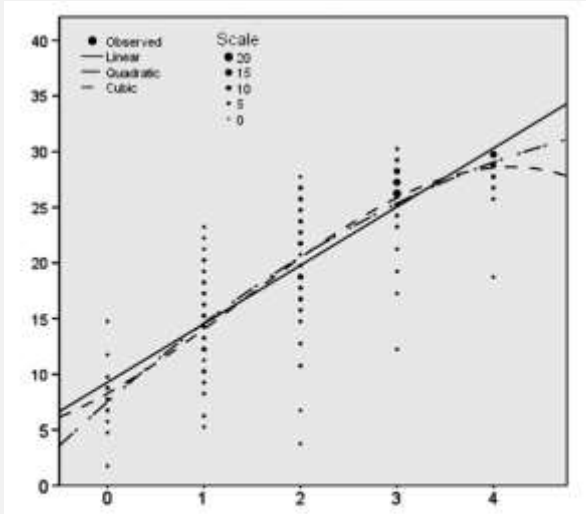


Figure 3. Curve estimation (station 14), assessor checklist score (*x*) *versus* global grade (*y*).

Station	Cronbach's alpha if item deleted	R^2	Inter-grade discrimination	Number of failures	Between-group variation (%)
1	0.745	0.465	4.21	53	31.1
2	0.742	0.590	5.23	24	30.1
3	0.738	0.555	5.14	39	33.0
4	0.742	0.598	4.38	39	28.0
5	0.732	0.511	4.14	29	20.5
6	0.750	0.452	4.74	43	40.3
7	0.739	0.579	4.51	36	19.5
8	0.749	0.487	3.45	39	33.8
9	0.744	0.540	4.06	30	36.0
10	0.747	0.582	3.91	26	29.9
11	0.744	0.512	4.68	37	37.6
12	0.744	0.556	2.80	23	32.3
13	0.746	0.678	3.99	30	22.0
14	0.746	0.697	5.27	54	27.3
15	0.739	0.594	3.49	44	25.9
16	0.737	0.596	3.46	41	34.3
17	0.753	0.573	3.58	49	46.5
18	0.745	0.592	2.42	15	25.4
19	0.749	0.404	3.22	52	39.5
20	0.754	0.565	4.50	37	34.1

Godfrey Pell, Richard Fuller, Matthew Homer & Trudie Roberts (2010) How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49, Medical Teacher, 32:10, 802-811, DOI: 10.3109/0142159X.2010.507716

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of rating scales, use of mixed format scoring schemes
- More 'authentic' stations: longer, as appropriate to the task (not just 5 mins, up to 15 mins)
- More authentic tasks: including clinical decision-making, formulating differential diagnoses, suggesting next management steps More
- Formal training of expert clinician examiners
- Formal OSCE-specific evidence-based standard setting (Borderline Regression Method) +/= MNSP (minimum number of stations to pass)
- Post test OSCE specific metrics
- **Test security**

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Test security

- Evidence is mixed
  - Some scores went up and some went down….less effects in clinical examination and practical procedure stations
  - Much of the work was done with checklists (which could be memorized)
  - Less applicable to rating scales
- Analyse your own data
  - Part of post OSCE metrics
- Sequestration can allay some anxieties among candidates

"The ability to perform a skill requires practice and experience. It is therefore questionable the extent to which knowing the task in advance offers any substantial advantage to a candidate."

Boursicot K, Kemp S, Wilkinson T, Findyartini A, Canning C, Cilliers F, & Fuller R (2021) Performance assessment: consensus statement and recommendations from the 2020 Ottawa Conference. Medical Teacher, 43(1), 58-67. https://doi.org/10.1080/0142159X.2020.1830052

Developments in OSCE best practice for the 2020s

- Scoring schemes: use of rating scales, use of mixed format scoring schemes
- More 'authentic' stations: longer, as appropriate to the task (not just 5 mins, up to 15 mins)
- More authentic tasks: including clinical decision-making, formulating differential diagnoses, suggesting next management steps More
- Formal training of expert clinician examiners
- Formal OSCE-specific evidence-based standard setting (Borderline Regression Method) +/= MNSP (minimum number of stations to pass)
- Post test OSCE specific metrics
- Test security
- **Technology**

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Harnessing technology

Specialised software programmes

- Delivery on iPads
- Electronic scoring and collation of scores
- Collection of feedback from examiners (narrative)
- Individualised score reporting for candidates
- Application of special OSCE metrics
- Monitoring of examiner performance

‘Remote’ OSCEs

- Conducted using video-conferencing software (eg Zoom)
- Acceptable for consultation skills tasks: taking history, explanation
- ‘reflection of current medical practice’
- NOT recommended for clinical examination or practical procedure tasks

Feedback for candidates

Information by station

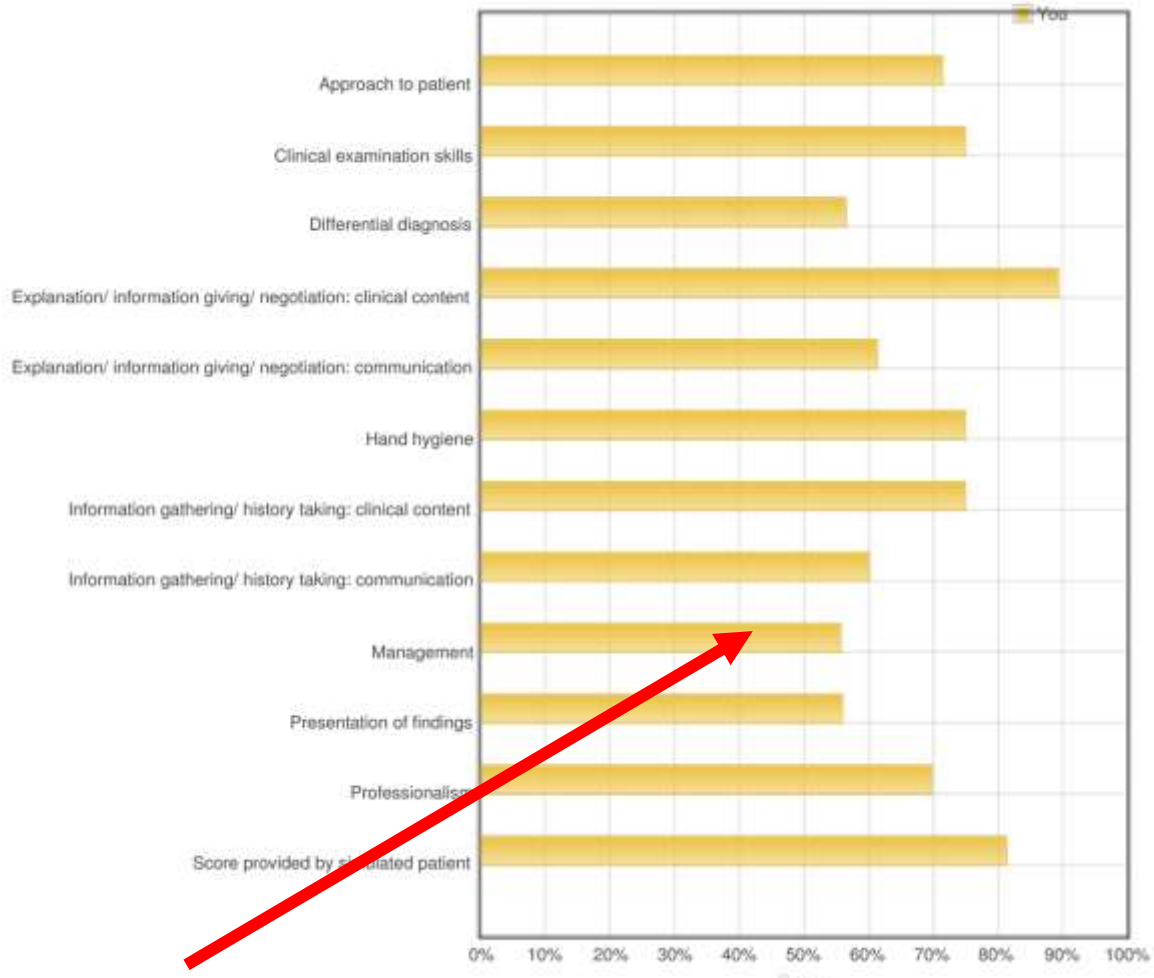
Candidates can see their scores & pass/fail status **by station**

No.	Stations	Your Score	Pass Mark
1	AY20-21 CPX-4 PE_Gastrointestinal (Summative)	37.5%	50.6%
2	AY20-21 CPX-4 PE_Respiratory_Lung_(Summative)	97.5%	51.5%
3	AY20-21 CPX-4 PE_Cardiovascular (Summative)	57.5%	55.2%
4	AY20-21 CPX-4 PE_Neuro_PNS (Summative)	76.2%	51.1%
5	AY20-21 CPX-4 PE_MSK_Back Pain Sciatica (Summative)	71.2%	49.9%
6	AY20-21 CPX-4 PE_Endocrine_Neck_Summative	53.8%	53.7%
7	AY20-21 CPX-4 Hx_SOB (Summative)	57.5%	46.6%
8	AY20-21 CPX-4 Hx_Cardiovascular (Summative)	71.2%	43.3%
9	AY20-21 CPX-4 Hx_GI(Jaundice)-Summative	68.8%	54.5%
10	AY20-21 CPX-4 Hx_MSK_Fall Fracture (Summative)	50.0%	40.1%
11	AY20-21 CPX-4 Hx_Generic_(Fever)_Summative	70.0%	45.9%
12	AY20-21 CPX-4 Pro Enc_Generic_Repeat Blood (Summative)	90.0%	47.4%
13	AY20-21 CPX-4 Pro Enc_Neuro_Epilepsy (Summative)	77.5%	55.1%
14	AY20-21 CPX-4 Pro Enc_Endocrine_Diabetes (Summative)	68.8%	51.0%
15	AY20-21 CPX-4 ProEnc_GI_Phone Consult_(Surgical Emergency)-Summative	75.0%	46.8%
	Overall results	**68.2%**	**56.8%**

Overall score

This gives them the concept of how well they have performed compared to expected standards

Information by domain



Candidates can see their scores aggregated by **domain** across all stations

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

Recent significant publications

Chan, SCC, Choa, G, Kelly, J, Maru, D, Rashid, MA. **Implementation of virtual OSCE in health professions education: A systematic review**. *Med Educ*. 2023; 1- 11. doi:10.1111/medu.15089

Boursicot, K, Kemp, S, Wilkinson, T, Findyartini, A, Canning, C, Cilliers, F, & Fuller, R (2021). **Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference**. Medical Teacher, 43:1, 58-67 https://doi.org/10.1080/0142159X.2020.1830052

Boursicot, Kemp & Fuller 2021 **Chapter 4 "Quality Assurance of OSCEs"**
in 'Understanding Assessment in Medical Education through QA' Eds Malau-Aduli, Hays, van Der Vleuten, McGraw Hill, UK

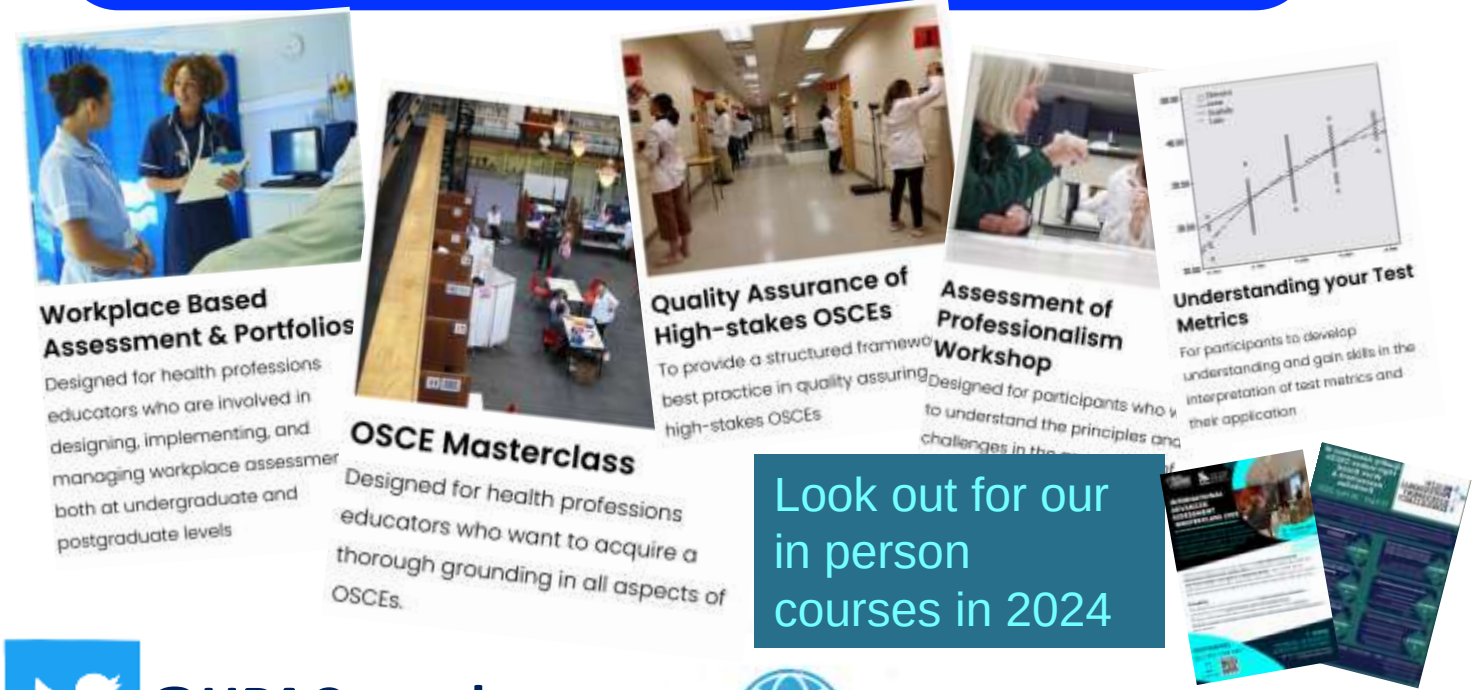
Thank you for your attention

katharineboursicot@hpac.sg

HEALTH PROFESSIONAL ASSESSMENT CONSULTANCY

EXPERTS IN HEALTH PROFESSIONS EDUCATION

- HPAC is a **consortium of experts** who undertake consultancies in a number of areas, especially assessment and examinations at undergraduate and postgraduate level.
- We provide **high-quality courses** on all aspects of assessment in health professions education, as well as **consultancy, tailored faculty development** and **research services**.



Look out for our in person courses in 2024

@HPAConsultancy

https://www.hpac.sg

https://www.facebook.com/hpac.sg/