National inter-rater agreement of standardised simulated-patient-based assessments.

Sam AH[1], Reid MD[1], Thakerar V[1], Gurnell M[2], Westacott R[3], Reed MWR[4], Brown CA[5],

[1] Imperial College School of Medicine, Imperial College London, UK

[2] Wellcome Trust-MRC Institute of Metabolic Science, University of Cambridge and NIHR Cambridge Biomedical Research Centre, Addenbrooke's Hospital, Cambridge, UK

[3] Birmingham Medical School, University of Birmingham, Edgbaston, Birmingham, UK

[4] Brighton and Sussex Medical School, University of Sussex, Brighton, UK

[5] Division of Health Sciences, Warwick Medical School, University of Warwick, UK

Corresponding author:

Dr Celia Brown

Warwick Medical School,

The University of Warwick,

Coventry, UK

Celia.Brown@warwick.ac.uk

ABSTRACT

Purpose

The forthcoming UK Medical Licensing Assessment will require all medical schools in the UK to ensure that their students pass an appropriately designed Clinical and Professional Skills Assessment (CPSA) prior to graduation and registration with a licence to practice medicine. The requirements for the CPSA will be set by the General Medical Council, but individual medical schools will be responsible for implementing their own assessments. It is therefore important that assessors from different medical schools across the UK agree on what standard of performance constitutes a fail, pass or good grade.

Methods

We used an experimental video-based, single-blinded, randomised, internet-based design. We created videos of simulated student performances of a clinical examination at four scripted standards: clear fail (CF), borderline (BD), clear pass (CPX) and good (GD). Assessors from ten regions across the UK were randomly assigned to watch five videos in 12 different combinations and asked to give competence domain scores and an overall global grade for each simulated candidate. The inter-rater agreement as measured by the intraclass correlation coefficient (ICC) based on a two-way random-effects model for absolute agreement was calculated for the total domain scores.

Results

One-hundred and twenty assessors enrolled in the study, with 98 eligible for analysis. The ICC was 0.93 (95% CI 0.81 to 0.99). The mean percentage agreement with the scripted global grade was 74.4% (range 40.8% to 96.9%).

Conclusions

The inter-rater agreement amongst assessors across the UK when rating simulated candidates performing at scripted levels is excellent. The level of agreement for the overall global performance level for simulated candidates is also high. These findings suggest that assessors from across the UK viewing the same simulated performances show high levels of agreement of the standards expected of students at a “clear fail”, “borderline”, “clear pass” and “good” level.

KEYWORDS

Undergraduate, assessment, inter-rater reliability,

PRACTICE POINTS

- There is little published literature comparing the standards of assessors from across the UK or that directly compares the marks assessors would give when rating the same simluated candidate performances.
- Our results show a reassuringly high level of inter-rater agreement amongst assessors nationally, supporting the concept of broad agreement in what standard of performance constitutes a pass or fail.
-

INTRODUCTION

The forthcoming UK Medical Licensing Assessment (UKMLA) will require all medical schools in the UK to ensure that their students undertake an applied knowledge test and an appropriately designed Clinical and Professional Skills Assessment (CPSA) which will act as the final high-stakes clinical examination(General Medical Council). Whilst the General

Medical Council (GMC) has set the specific requirements for the CPSA(General Medical Council), it will be designed and run by individual medical schools across the UK. The exact format of the CPSA will therefore vary between different medical schools, although all must meet the agreed specific criteria. It is anticipated that the majority of medical schools will opt to use an Objective Structured Clinical Examination (OSCE) format but the use of other standardised-patient-based (SP-based) formats are also acceptable(General Medical Council 2019). Regardless of the specific format of the CPSA, all SP-based assessment methods require assessors to provide ratings of a candidate's performance at each encounter or station. These ratings often take the form of a global grade of performance but may also include individual domain-based scoring. As the UKMLA aims to ensure that all newly qualified doctors in the UK meet a common threshold for safe practice (General Medical Council), it is important to determine whether assessors across the UK are comparable in their judgements of competence and would give equivalent marks to the same candidate performance. Quality assurance of assessor ratings is paramount given that the scores awarded in the graduation-level clinical examinations may contribute to a cut-score that determines which individuals pass or fail, which in turn dictates who may progress to provisional registration with the GMC and be granted a licence to practice medicine. Ensuring that the passing standards are fundamentally fair between medical schools is a crucial part of the educational contract that responsible institutions have with their students(Watling 2014) and also reassures the public that graduates have met the required safe threshold for practice(Wass et al. 2001), regardless of which institution they have graduated from.

The validity of OSCEs and SP-based assessments has been well explored (Swanson & van der Vleuten 2013) but their reliability is dependent on many factors including the number of stations and testing time (Van Der Vleuten & Schuwirth 2005), the number of assessors

(Swanson & van der Vleuten 2013), the number of domains assessed (Tavares et al. 2016) and the format of the scoring response (Regehr et al. 1998). To mitigate against these sources of unwanted variability and maintain standards at a systematic level, medical schools in the UK currently use external examiners: senior educators from other medical schools who ensure that the appropriate processes are followed and that standards are broadly consistent between institutions. However, the feedback that external examiners provide is often qualitative in nature which makes comparisons between institutions challenging. At an individual level, assessors are prone to variability due to cognitive biases such as leniency, inconsistency, and the halo effect (McManus et al. 2006; Iramaneerat & Yudkowsky 2007; Harasym et al. 2008), which may ultimately threaten the validity and objectivity of the assessment format (Hawkins et al. 2010). Attempts to reduce the impact of these sources of assessor variability though training have shown limited effect (Cook et al. 2009). Previous studies have established that assessors from different UK medical schools may set significantly different passing standards for identical OSCE stations (Boursicot et al. 2006; Boursicot et al. 2007), and that these standards vary when compared to a standardized control (Chesser et al. 2009). However, no study to date has explored whether assessors from different medical schools would award comparable marks to the same candidate performances, or the extent to which assessors agree when viewing candidates scripted to perform at a defined standard. This study therefore sought to establish the level of inter-rater agreement amongst assessors across the UK when rating simulated candidate performances of an examination station when other sources of variability typical of SP-based assessments were controlled for.

METHODS

Study Design

We used an experimental video-based, single-blinded, randomised, internet-based design.

Procedure

Seven 10-minute videos were created of simulated candidates completing a clinical examination typical of those assessed in graduation-level assessments (a cranial nerve examination). Volunteer Clinical Teaching Fellows affiliated with Imperial College London were recruited for this role. All simulated candidates were female, of white ethnicity, and a similar age to avoid confounding based on these factors. The simulated patient (who had experience of undertaking this role in real examinations) was the same in all of the videos. Four of the videos demonstrated the simulated candidates performing the station at one of four overall performance levels: clear fail (CF), borderline (BD), clear pass (CPX) or good (GD). The other three videos showed a candidate performing at a "clear pass" level but with various physical attributes. These were included to allow a parallel study to be undertaken exploring the impact of assessor bias, the results of which will be reported as a separate study. Each candidate followed a script created by a panel of experienced examiners to ensure they were performing at the intended level. 12 sets of five videos were then created; every set including a video of a candidate performing at each of the overall performance levels as well as one video of a candidate with a physical attribute performing at a "clear pass" level (*Appendix 1*). The ordering of the five videos differed across the 12 sets to minimise any bias associated with ordering effects. Each participant was randomly allocated to one of the 12 video sets.

Recruitment and Consent

The study was approved by the Medical Education Ethics Committee at Imperial College London (MEEC1718-105). Each medical school in the UK was contacted via the Medical

Schools Council and invited to take part in the study. Heads of assessment at each medical school were encouraged to invite a representative sample of assessors to participate in the study via the study website. Participants were informed that they were taking part in a study exploring inter-rater reliability amongst assessors but were not informed that the videos they viewed were of simulated candidates, nor that they were scripted to perform at set standards. Participants were not informed that each set of videos included candidates performing at various standards. No identifiable information was collected about the participants. Participants were required to be clinicians with at least one prior experience of formally assessing medical students in clinical examinations. Participants were informed that completion of the marksheets for all five videos and submission of the post-completion questionnaire was evidence of consent. Participants were able to withdraw from the process by closing the web browser at any time prior to completion of the study but due to the lack of collection of identifiable data, were not able to withdraw after submitting their results. Any incomplete data, where participants did not view and score all five videos, were not used in the analysis.

Measures

Participants were asked to assess the candidates at the level expected of a final-year medical student, where a pass would indicate they were competent to begin clinical practice. Participants then viewed the five videos in a randomised order and were provided with a blank mark sheet to complete alongside each video (*Figure 1*). Participants marked each candidate in four domains; “Physical examination”, “Identify physical signs and the most likely diagnosis”, “Clinical management skills”, and “Interpersonal skills”. Each domain was scored between 0 and 4, with a maximum possible total score of 16. Participants were also asked to assign each candidate a global grade of either “clear fail”, “borderline”, “clear pass”

or “good”. Participants were able to return to mark sheets for previous candidates but were not able to pause, rewind or replay the videos, to reflect the contemporaneous nature of rating competency in practice. Following completion of the mark sheets for all five videos, participants were asked to confirm their assessment experience, job role, gender, ethnicity and the geographical region where they worked.

[FIGURE 1]

Statistical analysis

Data management and analysis were conducted using Stata V16. The inter-rater reliability based on the candidates’ total scores (out of 16) as measured by the intraclass correlation coefficient (ICC) was calculated based on a two-way random-effects model for absolute agreement (Koo & Li 2016). The percentage agreement for each candidate was calculated by the number of examiners who rated the candidate at the global grade to which they were scripted divided by the total number of ratings for that candidate (n=98). The mean percentage agreement was then calculated as the total number of ratings where the global grades given matched the scripted level divided by the total number of ratings (n=392).

RESULTS

Participants

120 assessors participated in the study. Five assessors were removed from the analysis due to a self-reported lack of experience. Seventeen participants did not complete viewing and rating every candidate within the set and were therefore removed from the analysis. Table 1 shows the demographic details of all participants included in the analysis. Participants included in the analysis came from ten distinct regions across the UK. Participants were

varied in their level of experience and job role. The number of participants who viewed and rated each of the 12 sets of videos was comparable.

[TABLE 1]

Global grades

The global grades were converted to numerical values where clear fail=1, borderline=2, clear pass =3, and good=4. *Figure 2* shows the number of assessors rating each candidate at each global grade. Across the four candidates, the percentage of assessors who rated the scripted candidate at the intended global grade ranged from 40.8% (CPX) to 96.9% (CF), with a mean of 74.4%.

[FIGURE 2]

Total scores

The median score for CF was 3 (interquartile range [IQR] 2 to 4). The median score for BL was 7 (IQR 6 to 8). The median score for CPX was 14 (IQR 12 to 15). The median score for GD was 16 (IQR 15-16). *Figure 3* shows the total scores and corresponding global grades given to the candidates. Eighty assessors (81.6%) scored the candidates in the intended order, where the total scores for CF<BL<CPX<GD. Amongst those who did not score the candidates in keeping with the intended order 13 assessors (72.2%) scored CPX and GD equally, and five (27.8%) scored CPX higher than GD. Following a repeated-measures ANOVA, the partial Omega squared (ω^2) statistics for candidates and examiners were 0.946 and 0.241 respectively.

Inter-rater reliability

The inter-rater reliability for the total scores as measured by the ICC was 0.93 (95% CI 0.81 to 0.99).

[FIGURE 3]

DISCUSSION

The ICC of 0.93 for the total scores across the candidates suggests that the overal inter-rater agreement for our study was excellent (Koo & Li 2016). This reassuringly high value suggests that assessors across the UK are able to differentiate between candidates scripted to perform at set standards, and award similar marks when doing so. This is also demonstrated by the high number of assessors whose total scores reflected the intended hierarchy of candidates (n=80, 81.6%) and whose global grades agreed with the scripted standards (74.4%). All situations in which the candidates were not scored in the correct order were due to ties between CPX and GD, or when CPX was scored higher than GD. The greatest source of disagreement between the intended and given global grade was also where CPX was awarded a global grade of "good" by 57 (58.2%) of assessors. This finding may suggest whilst assessors' absolute agreement on the scores for candidates is high, the definition of what constitutes a "good" candidate compared to one who is at a "clear pass" level may vary between assessors. This premise is in keeping with previous work that has established criterion uncertainty as a source of assessor variation (Yeates et al. 2013). While of relatively low importance from a patient safety perspective, the differentiation between those candidates who are at a "clear pass" level and those who are "good" remains of particular importance to students. Further work should explore the disparity between definitions of a "good" candidate amongst assessors at a national level.

Although the overall range of total scores for each candidate varied from 6 (CF) to 10 (BL), the relatively small range in standard deviation from 1.15 (GD) to 1.86 (CPX) demonstrates that assessors whose scores were far from the group mean were rare.The simulated passmark as calculated using the borderline group method (Livingston & Zieky 1982) based on the median total domain score of the 68 candidates rated as borderline was 7/16. The percentage of assessors' scores that would have resulted in a pass were 0% (CF), 60.8% (BL), 100% (CPX) and 100% (GD). This finding is reassuring, since variation in scores between assessors had little influence on the pass/fail consequences for the simulated candidates. The less than total agreement for the borderline candidate would be expected and supports the need for multiple-station clinical assessments to ensure borderline candidates are awarded an accurate overall pass/fail decision.

The high inter-rater reliability found in our study is in contrast to the relatively low reliability typically measured in OSCEs (Brannick et al. 2011). It is important to note that our study does not reflect the distribution of performances seen in real-life cohorts of students and as such the high ICC is partly a result of the study design, where candidates were scripted to perform at distinct standards. Our scripts also included consistency of performance across domains within each standard, when in real life candidates may be relatively stronger in some domains and weaker in others, making judgements more challenging and less reliable. However, these features of the study design do not interfere with the intended objective to explore the extent to which assessors from across the UK agree when rating candidates performing at scripted levels. Instead, we highlight these points in order to avoid the misinterpretation of the high inter-reliability in isolation and caution against the transferability of this result to real world contexts where candidate cohorts are more heterogenous.

There are a number of limitations to our study. Whilst we attempted to control as many aspects of the study as possible, we were required to use different simulated candidates for each performance to ensure the stations mimicked real examination performances without revealing their scripted nature. The scripts were created and reviewed by a panel of experienced examiners but it is also possible that similarities between the scripted standards at the "Clear Pass" and "Good" levels were responsible for the comparable scores and grades. Although all assessors were volunteers and were deliberately not informed that they were viewing simulated candidates scripted to perform at set standards, it is possible that they were not a representative sample of the population of assessors as a whole. As we were assisted in recruitment by individual medical schools we are unable to calculate a response rate based on the number of examiners invited to participate. The simulated candidates in the study were all white and female and it is therefore possible that variability amongst assessors may be different when rating candidates of different genders and ethnicities. The simulated station in the study was of a specific clinical examination and the inter-rater reliability for other clinical examinations and different types of station (such as those assessing communication skills) may be different. The first video in each set was of a candidate performing at a clear pass standard which may have allowed participants to anchor a scale and improve the consistency of their marking. Finally, participants in our study were only asked to rate five videos, which may have minimized the impact of inattention and fatigue on rating. Further studies should explore these effects on a larger scale.

Implications of findings

Our findings should reassure medical schools, students and the public alike that assessors in the UK are consistent with the scores they would give to candidates performing at scripted

levels, and show high levels of agreement in their perception of competency. Whilst our study does not reflect the reality of the CPSA (where the assessment format will vary between medical schools) it does demonstrate an important and directly relevant concept; that the judgement of assessors across the UK when considering what standards constitute a pass or fail is consistent.

CONCLUSION

The inter-rater agreement amongst assessors across the UK when rating simulated candidates performing at scripted levels is excellent. The level of agreement for the overall global performance level for simulated candidates is also high. These findings suggest that assessors from across the UK viewing the same simulated performances show high levels of agreement of the standards expected of students at a “clear fail”, “borderline”, “clear pass” and “good” level.

ACKNOWLEDGEMENTS

The authors are grateful to all UK medical school assessment leads for their help in recruiting assessors. The authors are also grateful to the Medical Schools Council for administrative support with the study.

DECLARATION OF INTERESTS

MG is supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. CAB is supported by the NIHR Applied Research Collaboration (ARC) West Midlands. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

FUNDING

The Medical Schools Council funded the recruitment of the simulated candidates, simulated patient and sourcing of the recording equipment for this study.

NOTES ON CONTRIBUTORS

Professor Amir H Sam is head of Imperial College School of Medicine and consultant physician and endocrinologist at Imperial College Healthcare NHS Trust.

Dr Michael D Reid was a Clinical Education & Research Fellow at Imperial College London and is now a trainee in Geriatric Medicine at Kingston Hospital.

Dr Viral Thakerar is the lead for year 1 and 2 clinical placements at Imperial College School of Medicine and a practising general practitioner.

Professor Mark Gurnell is Clinical SubDean at the University of Cambridge School of Clinical Medicine and Professor of Clinical Endocrinology at Institute of Metabolic Science & Department of Medicine.

Dr Rachel Westacott is a Senior Lecturer in Medical Education at Birmingham Medical School and an acute medicine consultant at University Hospitals of Leicester NHS Trust (CCT in Nephrology).

Professor Malcolm Reed is a breast cancer surgeon who has been Dean of Brighton and Sussex Medical School since 2014 having moved from Sheffield University Medical School where he was head of Undergraduate Assessment for medicine. He is currently Co-Chair of Medical Schools Council and Chair of the education subcommittee.

Dr Celia Brown is an Associate Professor in Quantitative Methods at Warwick Medical School. She has research interests in selection and assessment and teaches quantitative methods at all levels in Higher Education.

REFERENCES

Boursicot KAM, Roberts TE, Pell G. 2006. Standard setting for clinical competence at graduation from medical school: A comparison of passing scores across five medical schools. Adv Heal Sci Educ. 11(2):173–183.

Boursicot KAM, Roberts TE, Pell G. 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ [Internet].

[accessed 2020 May 22] 41(11):1024–1031. http://doi.wiley.com/10.1111/j.1365-2923.2007.02857.x

Brannick MT, Erol-Korkmaz HT, Prewett M. 2011. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 45(12):1181–1189.

Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G. 2009. Sources of variation in performance on a shared OSCE station across four UK medical schools. Med Educ [Internet]. [accessed 2020 Feb 11] 43(6):526–532. http://doi.wiley.com/10.1111/j.1365-2923.2009.03370.x

Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. 2009. Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. J Gen Intern Med. 24(1):74–79.

General Medical Council. September 2019-CPSA requirements for piloting Requirements for the Medical Licensing Assessment Clinical and Professional Skills Assessment Background (https://www.gmc-uk.org/-/media/documents/mla-pilot-cpsa-requirements_pdf-80179895.pdf). [place unknown].

General Medical Council. 2019. Thematic report on Clinical and Professional Skills Assessment formative meetings [Internet]. [place unknown]; [accessed 2020 May 22]. https://www.gmc-uk.org/-/media/documents/mla-thematic-report_pdf-80181648.pdf

Harasym PH, Woloschuk W, Cunning L. 2008. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Heal Sci Educ. 13(5):617–632.

Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. 2010. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. Acad Med [Internet].

[accessed 2020 Feb 10] 85(9):1453–61. http://www.ncbi.nlm.nih.gov/pubmed/20736673

Iramaneerat C, Yudkowsky R. 2007. Rater errors in a clinical skills assessment of medical students. Eval Heal Prof [Internet]. [accessed 2020 Feb 10] 30(3):266–283. http://www.ncbi.nlm.nih.gov/pubmed/17693619

Koo TK, Li MY. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 15(2):155–163.

Livingston SA, Zieky MJ. 1982. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. [place unknown]: Princeton, NJ: Educational Testing Service.

McManus I, Thompson M, Mollon J. 2006. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ [Internet]. [accessed 2020 Feb 11] 6(1):42. https://bmcmededuc.biomedcentral.com/articles/10.1186/1472-6920-6-42

Regehr G, MacRae H, Reznick RK, Szalay D. 1998. Comparing the psychometric properties of checklists and global ratings scales for assessing performance on an OSCE-format examination. Acad Med. 73(9):993–997.

Swanson DB, van der Vleuten CPM. 2013. Assessment of Clinical Skills With Standardized Patients: State of the Art Revisited. Teach Learn Med. 25(SUPPL.1).

Tavares W, Ginsburg S, Eva KW. 2016. Selecting and Simplifying: Rater Performance and Behavior When Considering Multiple Competencies. Teach Learn Med. 28(1):41–51.

Van Der Vleuten CPM, Schuwirth LWT. 2005. Assessing professional competence: From methods to programmes. Med Educ. 39(3):309–317.

Wass V, Van Der Vleuten C, Shatzer J, Jones R. 2001. Assessment of clinical competence.

Lancet. 357(9260):945–949.

Watling CJ. 2014. Unfulfilled promise, untapped potential: Feedback at the crossroads. Med Teach. 36(8):692–697.

Yeates P, O'Neill P, Mann K, Eva K. 2013. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract [Internet]. [accessed 2020 Feb 10] 18(3):325–41. http://www.ncbi.nlm.nih.gov/pubmed/22581567

APPENDIX 1

Video Ordering

Version	**Video 1**	**Video 2**	**Video 3**	**Video 4**	**Video 5**
1	CPX	CPH	BL	CF	GD

2	CPX	CF	CPH	GD	BL
3	CPX	BL	GD	CPH	CF
4	CPX	GD	CF	BL	CPH
5	CPX	CPT	BL	CF	GD
6	CPX	CF	CPT	GD	BL
7	CPX	BL	GD	CPT	CF
8	CPX	GD	CF	BL	CPT
9	CPX	CPA	BL	CF	GD
10	CPX	CF	CPA	GD	BL
11	CPX	BL	GD	CPA	CF
12	CPX	GD	CF	BL	CPA

Key: CF – clear fail, BL – borderline, CPX – clear pass, no discernible attribute, CPH – clear pass, purple hair, CPT – clear pass, tattoo on both forearms, CPA – clear pass, regional accent, GD – good.

All participants n=98			
Demographics		n	(%)
Experience			
	None	0	0.00
	1-2 Exams	6	6.12
	3-4 Exams	13	13.27
	5+ Exams	79	80.61
Job Role			
	Consultant	40	40.82
	Primary Care Physician	33	33.67
	Specialty Training years 3+	1	1.02
	Core Training or Specialty Training years 1-2	1	1.02
	Other/please specify role & grade if appropriate	22	22.45
	Prefer not to say	1	1.02
Gender			
	Male	44	44.90
	Female	54	55.10
Region			
	East Anglia	16	16.33
	East Midlands	5	5.10
	London	15	15.31
	North West	6	6.12
	Scotland	19	19.39
	South East	7	7.14
	South West	8	8.16
	Wales	2	2.04
	West Midlands	4	4.08
	Yorkshire and the Humber	16	16.33

Table 1: Participant descriptives

Mark Sheet: Cranial Nerve Examination

Domain 1. Physical examination
Task: Examines the cranial nerves (I-XII)
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Domain 2. Identifying physical signs and the most likely diagnosis
Task: Reports abnormal findings and offers the most likely diagnosis
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Domain 3. Clinical management skills
Task: Explains management of patient
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Domain 4. Interpersonal skills
Task: Communicates appropriately with the patient and examiner
Excellent (4)
Good (3)
Adequate (2)
Fail (1)
Severe fail (0)

Global Grade
Good
Clear Pass
Borderline
Fail

Figure 1: Sample mark sheet

		Global Grade (GG) assigned by assessor				
		Clear Fail (1)	Borderline (2)	Clear Pass (3)	Good (4)	**Median GG (IQR)**
Scripted Candidate	CF	95 (96.9)	3 (3.1)	0 (0.0)	0 (0.0)	1 (1 to 1)
	BL	25 (25.5)	64 (65.3)	9 (9.2)	0 (0.0)	2 (1 to 2)
	CPX	0 (0.0)	1 (1.0)	40 (40.8)	57 (58.2)	4 (3 to 4)
	GD	0 (0.0)	0 (0.0)	4 (4.1)	94 (95.9)	4 (4 to 4)

Figure 2: The number of assessors (%) rating each candidate at each global grade, and the median global grade (interquartile range) for each candidate.

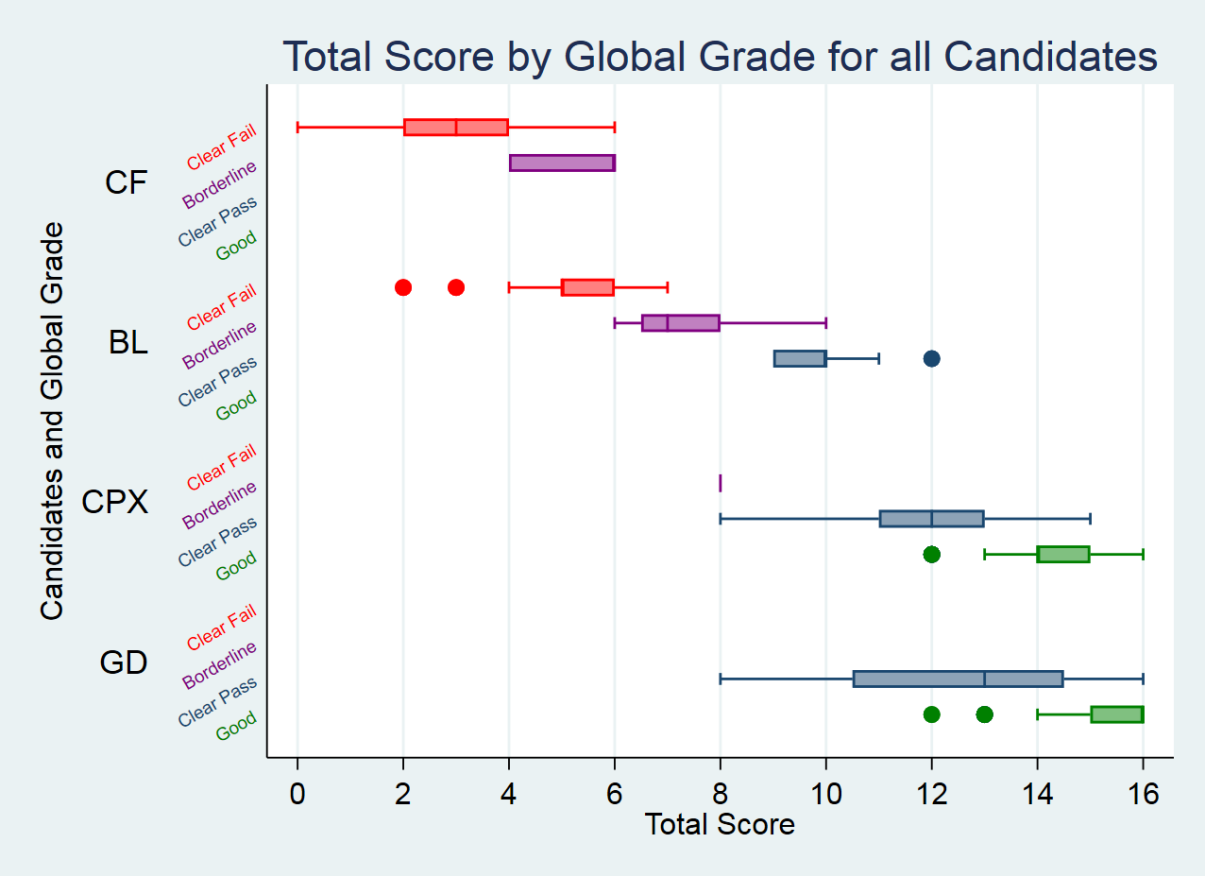

Figure 3: The total scores and corresponding global grade given to each candidate scripted to perform at each level (CF = Clear Fail, BL = Borderline, CPX = Clear Pass, GD = Good)