EDUCATION
RESEARCH SERVICE

AWARD WINNING ORIGINAL ARTICLE

Measuring the quality of an objective structured clinical examination in a chiropractic program: A review of metrics and recommendations

Alice E. Cade, MHSc, BSc (Chiro), PhD and Nimrod Mueller, MChiro

ABSTRACT

Objective: The objective structured clinical examination (OSCE) is a commonly used assessment of clinical skill. Ensuring the quality and reliability of OSCEs is a complex and ongoing process. This paper discusses scoring schemas and reviews checklists and global rating scales (GRS) for marking. Also detailed are post-examination quality assurance metrics tailored to smaller cohorts, with an illustrative dataset.
Methods: A deidentified OSCE dataset, from stations with a checklist and GRS, of 24 examinees from a 2021 cohort was assessed using the following metrics: Cut scores or pass rates, number of failures, R^2, intergrade discrimination, and between-group variation. The results were used to inform a set of implementable recommendations to improve future OSCEs.
Results: For most stations, the cut score calculated was higher than the traditional pass of 50% (58.9.8%–68.4%). The number of failures was low for traditional pass rates and cut scores (0.00–16.7%), except lab analysis where number of failures was 50.0%. R^2 values ranged from 0.67–0.97, but the proportion of total variance was high (67.3–95.9). These data suggest there were potential missed teaching concepts, that station marking was open to examiner interpretation, and there were inconsistencies in examiner marking. Recommendations included increasing examiner training, using GRSs specific to each station, and reviewing all future OSCEs with the metrics described to guide refinements.
Conclusion: The analysis used revealed several potential issues with the OSCE assessment. These findings informed recommendations to improve the quality of our future examinations.

Key Indexing Terms: Clinical Competence; Checklist; Reproducibility of Results; Benchmarking

J Chiropr Educ 2023;38(1):9–16 DOI 10.7899/JCE-22-29

INTRODUCTION

The Objective Structured Clinical Examination (OSCE) is a high-stakes, performance-based summative assessment of clinical skills.[1] Since the OSCE format was first used by Harden in the 1970s[2] it has been thoroughly studied and widely adopted by medical and complementary and alternative medicine educational institutions, including chiropractic programs.[3–5]

With assessments such as the OSCE, it is important to ensure the quality and rigor of examinations.[6] But how the quality of an OSCE is measured, and what mechanisms are available to ensure improvements in the quality of assessments over time, is not always clear. Moreover, given that chiropractic programs often have class sizes under 100,[7,8] statistical analyses appropriate for smaller cohorts are needed, as many analyses require large samples sizes, which chiropractic programs cannot always provide. While many analyses currently exist, there is no recommended battery of tests for small samples of OSCE scores. This paper provides an evidence-based pathway for educators to analyze, review, and improve small-scale OSCEs.

The purpose of this paper was to review the scoring of OCSEs and discuss post-exam statistical analyses. Its aim was to demonstrate a battery of analyses appropriate for small samples sizes using an actual OSCE data set from a European

This paper is the first prize winner of the Chiropractic Educators Research Forum/World Federation of Chiropractic Alan Adams Education Research Award award-winning paper presented at the World Federation of Chiropractic/Association of Chiropractic Colleges Global Education Conference, November 2-5, 2022. The award is funded in part by sponsorships from NCMIC, ChiroHealth USA, Activator Methods, Clinical Compass, World Federation of Chiropractic, and Brighthall. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, these sponsors.

First Published Online December 5 2023

Table 1 - An Example of a Nondiscrete Checklist

Neurological Examination	Done Well	Done Poorly	Not Performed
Stated 3 neurological related working hypotheses (WH)			
Demonstrates clinical rationale: WH are relevant to the case			
Peripheral nerves			
Cerebellar			
Cortical			
Cranial nerves			

chiropractic program. A secondary aim was to illustrate how these analyses could be used to inform quality improvement for futures OSCEs.

Scoring Issues in OSCE-Style Examinations

The OSCE assesses specific healthcare competencies in a mock environment, as a substitute for clinical competence, using a checklist and/or global rating scale.[9] Many areas of student performance can be assessed with an OSCE,[9] and assessments should aim to maximize the validity, reliability, objectivity, and feasibility of the OSCE.[10]

Scoring Checklists

Checklists are often used to score student performance, but assessing multiple areas within a single examination can lead to increasingly intricate checklists, trivializing the task.[9,11] Complex checklists can also lead to observer overload, negatively impact scoring behavior, reduce interobserver reliability, and increase the risk of inaccurate assessment, decreasing the exam validity.[9] The number of checklist items depends on the station and time allotted—about 8–25 customized checklist items are acceptable.[12,13] When used correctly, checklists can improve interexaminer reliability and support novice examiners.[14]

Each item on the checklist should be discrete, objective, and represent only 1 concept. If several points are combined in 1 item, specific instructions should be provided regarding scoring said item.[9] Table 1 presents an example of a checklist with nondiscrete concepts that may be too open to examiner interpretation.

Dichotomous checklist items (ie, competent or not yet competent) are easier to score, but may be narrow and outdated.[12] Table 2 presents some examples. A more modern "key features" approach focuses on essential elements of the task, such as eliciting history, seeking critical physical findings, or planning investigations that confirm or refute differential diagnoses.[15] A review of the marking checklist in Table 1 suggests it may be too open to examiner interpretation, as each point encompasses many concepts. This contrasts with the checklist in Table 2, which may be too prescriptive.

Weighting of checklist marking is favored (such as 0, 1, 4 for not done/done/done well) as it increases rubric accuracy[12] but more complex scoring does not enhance validity.[12,15] Negative scoring, where points are removed for incorrect answers, does not increase validity, instead it measures risk-taking behavior.[16]

Global Rating Scales

In cases where students have memorized the marking guide or used a nonspecific approach, the student may pass but the examiner may not be confident of their competence or clinical decision-making skills. Global rating scales (GRS) can address these issues as they are sensitive to different levels of expertise in examinees.[14] A GRS can increase interitem and interstation reliability, can be used in multiple tasks, and may better describe many facets of student expertise.[14]

Because a GRS requires examiners to use their judgment, it is crucial to minimize its subjectivity, so clear examiner instructions, adequate training, and behavioral anchors are needed. Behavioral anchors are descriptions of the range of performance for each station and can improve inter-rater reliability.[17] An example is shown in Table 3. GRSs should be recorded directly at the end of the station, after the checklist is scored.[12,18]

Statistical Analyses for Post-Examination Quality Metrics

Once an examination is complete and student scores have been acquired, there are several ways to objectively review the exam quality. The following section describes possible analyses that investigate OSCE quality, and their usefulness is discussed in relation to what they show, when they can be used, and when to exclude them from an analysis package.

Table 2 - A Dichotomous Checklist for a Neurology Station

Cerebellar Testing (C = competent, NYC = not yet competent)			
Verbalizes Test Name	**Verbalized Patient Instructions**	**Performs Test Bilaterally**	**Total**
❑ Romberg's part I & II	❑ C ❑ NYC	❑ C ❑ NYC	**/3**
Examiner to give feedback STRAIGHT AWAY: "The patient is unable to perform part II without falling over"			
❑ Student DOES NOT perform Tandem Romberg's after Romberg's			**/3**
❑ Heel to shin	❑ C ❑ NYC	❑ C ❑ NYC	**/3**
❑ Finger to nose	❑ C ❑ NYC	❑ C ❑ NYC	**/3**
❑ Rapid alternating hand movements	❑ C ❑ NYC		**/2**
Examiner Feedback: "Bilateral heel to shin, finger to nose and rapid alt hand movements are slow and uncoordinated"			**/14**

Table 3 - An Example of a Potential GRS

GRS - Cerebellar Testing					
	1 - Poor	2 - Fail	3 - Pass	4 - Good Pass	5 - Excellent Pass
Behavioral Anchors	-Does not attempt all tests -Instructions confused/ not understandable -Student performs tests that are dangerous to the patient welfare	-Most tests are attempted, but some are incorrectly performed -Instructions are confusing but comprehensible -Examination is unstructured and confusing	-Most tests are performed correctly, but may not give full clinical picture -Instructions are adequate but potentially confusing. -Examination is reasonably structured	-All tests are performed correctly -Instructions are comprehensible -Examination is well structured -Student is aware of potential patient welfare issues	-Student performs all tests perfectly -Instructions are clear, concise, and comprehensible to a lay-person -Examination is well structured and performed smoothly -Student recognizes which tests to NOT perform to preserve patient welfare

Number of Failures

This is a quick overview of the examination. A high failure rate does not necessarily mean a poorly designed station; expert judgment should be used to determine if the station was inappropriate for examinee skill level, and if a review of the course content is needed.[19]

Cut Scores

Traditional testing has a predetermined pass mark (50%–70%) that a student must attain to pass their examination, but this approach may not be useful when determining how to handle a borderline student or a particularly lenient or stringent examiner.[20] A borderline regression (BLR) analysis allows for a defensible and feasible method to identifying the cut score (or pass mark) and is reliable in small samples.[21] In BLR, checklist scores (Y-axis) for each GRS level (X-axis) are plotted and fitted with a regression line. Where the regression line for the borderline group intersects with the Y-axis is taken as the cut score.[21,22] This is shown in Figure 1.

Borderline regression does more statistical skills to compute and can be potentially sensitive to outliers. Such outliers may be a badly performing student who gets a near zero checklist score, or where an examiner gives the wrong overall grade.[21] Because of these potential limitations, other metrics should also be used.

G Coefficient

Generalizability theory, or G-theory, is an alternative assessment of reliability that assumes the reliability of a score or observation depends on the population that is observed and the environment where the testing is performed.[23] G-Theory is most robust and unbiased with sample sizes over 300,[24] making it an inappropriate statistical choice for smaller OSCE cohorts. Thus, it will not be discussed further in this paper.

Cronbach's Alpha

Cronbach's alpha is a measure of internal consistency. Meaning, it is a measure of how well a test actually measures what one wants it to measure.[25] The higher the internal consistency, the more confidence one can have that the examination is reliable. A Cronbach's alpha $\geq$ 0.70 is considered acceptable.[26] The degree to which multiple measures (or sections within the station) agree with each other is usually presented as "*alpha if item deleted.*"[6] The "*alpha if item deleted*" scores should be lower than the overall alpha score. If this is not the case, it suggests the station was not measuring what it was supposed to, the station was poorly designed, the topic was poorly taught, or the assessors were inconsistent.[23] Interestingly, an alpha of over 0.90 can be instructive as well, meaning the station might be too easy or redundant in nature. Cronbach's alpha is most reliable when used with high sample sizes and scales with a higher number of items,[26] so this metric should not be viewed in isolation when assessing an OSCE for quality.

R Squared

The R^2 coefficient is the proportional change in the dependent variable (the checklist score) due to change in the independent variable (the GRS). Generally speaking, a higher R^2

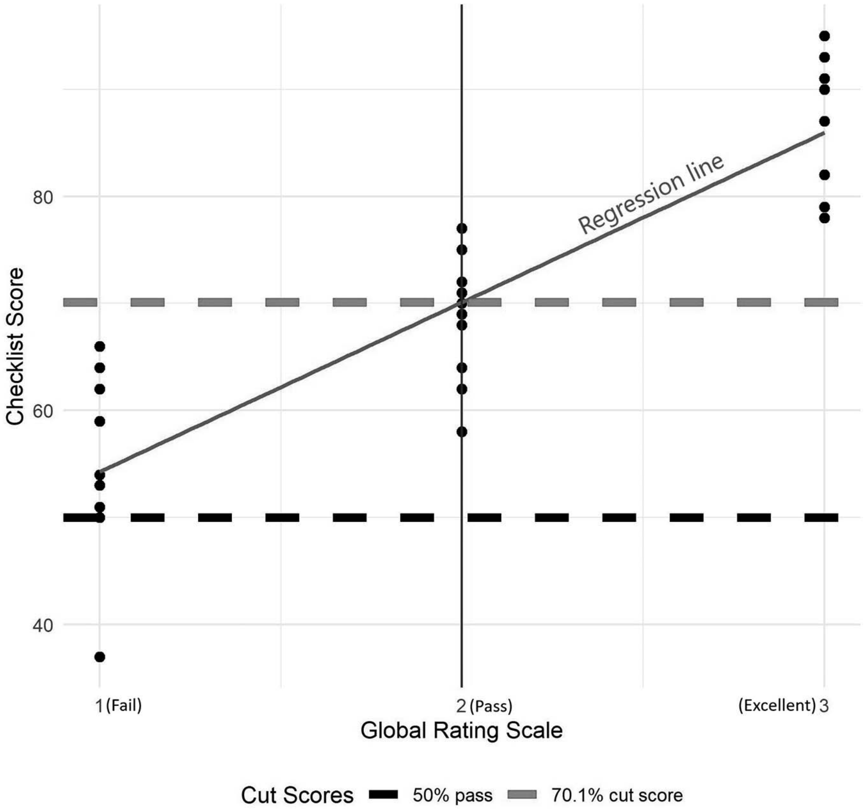

Figure 1 - Plot of mock OSCE scores against their associated mock GRS (1 = Fail, 2 = Pass, 3 = Excellent pass) indicating the difference between a 50% pass (black dashed line) and the cut score (red dashed line), or passing level suggested by a borderline regression analysis.

is preferable.[27] An adjusted R^2 of 0.759 implies that 75.9% of the variation in the students' GRS is accounted for by variation in their checklist scores. If the R^2 is low, it suggests a review of the station checklist or station design is necessary.[6]

Intergrade Discrimination

This statistic gives the average increase in checklist mark with each step up on the GRS, or the slope of the regression line (Fig 1).[28] If the slope is low, it indicates either high examiner variance of marking or a number of badly failing students that affected the regression line.[6]

Between-Group Variation

The proportion of total variance is an estimation of variance in checklist scores due to student performance[29] and indicates the consistency of the examination process. This is a reflection of other factors, such as differences in room setup, environment, or examiners.[20] This metric should be under 30%, with values over 40% being problematic.[6] This metric should be viewed in concert with R^2. A high proportion of variance and low R^2 suggests a poorly designed checklist, whereas a high proportion of variance and high R^2 suggests inconsistent marking.[6]

Using the information presented above as review, we conducted an analysis of OSCE scores from an actual exam. The aim of this project was to demonstrate a battery of analyses for OSCE quality that are appropriate for smaller samples. A secondary aim was to illustrate how these analyses could be used to inform post-examination changes to improve the quality of the OSCE in future iterations.

METHODS

A deidentified OSCE data set for the 2021 cohort (n = 24) from a European chiropractic program and their marking guides were supplied for review. OSCE stations were included in this analysis if they used a checklist score (numerical) and a global rating scale. This study was reviewed and approved by the Barcelona College of Chiropractic Research Committee and the Health and Disabilities Ethics Committee (HDEC) of New Zealand (2021 EXP 11582).

The OSCE data set of 7 stations and 24 examinees was evaluated. Descriptive statistics (unadjusted means, standard deviations, and counts) were used to describe the characteristics of the study sample. Adjusted R^2, regression slopes, and proportion of total variance were created using the stats package from base R software (R Foundation for Statistical Computing, Vienna, Austria). Borderline regression was performed via simple linear regression of OSCE checklist (dependent variable) and GRS scores (independent variable) also using the stats package. Model assumptions were visually assessed via quantile–quantile plots, fitted value and residual plots, and scatter plots. Statistical tests for OSCE quality were included if they were appropriate for use in small cohorts. Cronbach's alpha

Table 4 - Summary Statistics for the 2021 OSCE Exit Examination

Station/GRS	Min	Max	Median	Mean	SD
Care Plan Checklist	40.0	92.0	74.0	72.2	10.9
Care Plan GRS	2.0	5.0	3.0	3.3	0.7
Lab Analysis Checklist	30.0	85.0	50.6	56.4	17.7
Lab Analysis GRS	1.0	5.0	2.0	2.4	1.3
Neurology Checklist	47.0	91.0	78.5	75.5	9.5
Neurology GRS	1.0	5.0	4.0	3.8	1.0
Professionalism Checklist	30.0	100.0	77.5	76.9	19.8
Professionalism GRS	1.0	5.0	3.5	3.6	1.2
Systems Checklist	52	98	79.5	77.125	14.874
Systems GRS	2	5	4	3.917	1.139
Technique 1 Checklist	34	91	72	69.625	15.05
Technique 1 GRS	1	5	3.5	3.25	1.26
Technique 2 Checklist	14	92	78	70.667	22.563
Technique 2 GRS	1	5	4	3.792	1.444

n = 24 examinees.

was not calculated for all stations as this required the individual section marks from each stations' marking checklist, which was not provided in the data set.[26] Statistical significance was defined as *p* values less than .05. All data were presented at 1 or 2 significant digits for ease of reading, but all calculations were performed with unrounded data.

RESULTS

The 7 stations analyzed, each with an identical GRS, were Technique 1 and 2, Professionalism, Neurological examination, Laboratory analysis, Systems review, and Care plans. The highest and lowest scoring stations were Systems review (mean: 77.1 ± 14.9) and Laboratory analysis (mean: 56.4 ± 17.7), and the percentage of fails based on a traditional 50% pass rate ranged from 0 (Systems review) to 50% (Laboratory analysis). Summary statistics are shown in Table 4. The raw data for the stations can be found in Appendix 1.

An overview of the data, including cut scores and 50% pass scores, is given in Figure 2. Both R^2 (0.67–0.96) and between-group variation was high (67.31–95.85), with intergrade discrimination falling between 10.5 and 16.5.

For all stations, the cut score suggested by the borderline regression analysis was higher than the traditional pass rate of 50% (Table 5).

DISCUSSION

This study reviewed statistical analyses appropriate for assessing smaller OSCE cohorts and suggests the use of multiple metrics to analyze, review, and improve future exams. These metrics included the number of fails that traditional and borderline regression informed, Cronbach's alpha, R^2, intergrade discrimination, and between-group variation. Overall, these metrics suggest that the examination processes that generated the dataset may not be performing optimally, reducing the OSCE reliability and quality. In terms of which of the metrics seem to be most informative, the R^2 in combination with the proportion of variance suggested the need for a review of the scoring scales and that inconsistent marking was a problem. At this point in time, it is challenging to compare these metrics to other quality assurance processes, as no analyses specific to small cohorts have been found.

The Laboratory analysis station also stands out as it has the largest number of fails of all the stations, with 50% of the examinees failing the station—using the 50% pass rate or cut scores. This suggests the Laboratory analysis station may be beyond the current capabilities of the examinees or there were missed concepts in teaching its content. A visual analysis of the plots in Figure 2 suggested that factors such as no clear fails in the Care-plan or Systems stations may have affected the BLR analysis. Additionally, while these analyses have recommended for smaller sample sizes there is a caveat, especially for BLR analysis—stations must be of high quality with an R^2 of 0.50 or higher and have an even spread of candidates over the GRS.[30,31] If both of these criteria are not met, as in the Care plan and Systems stations, then a previously identified pass rate be used instead of BLR cut scores.[31]

Overall, the R^2 values were high, implying that most of the variation in the students' global ratings are accounted for by variation in their checklist scores.[27] There also were no apparent outliers in the intergrade discrimination scores, which are expected to be around 10% of the total checklist grade.[6] The neurology station intergrade discrimination was low suggesting a large variance in marking, which may have been due to differences in examiners, as the broad nature of the marking guide (Table 1) required a great deal of examiner interpretation.

The most concerning trend noted is the proportion of total variance (should be under 30%[6]), which estimates how much variance in checklist scores is due to student performance alone.[29] The proportion of variance in this dataset was much higher, ranging from 67.31 to 95.85. Furthermore, a high proportion of variance with a high R^2, like in the Systems station, suggests inconsistency in examiner marking rather than a problem with the scoring checklist.[6]

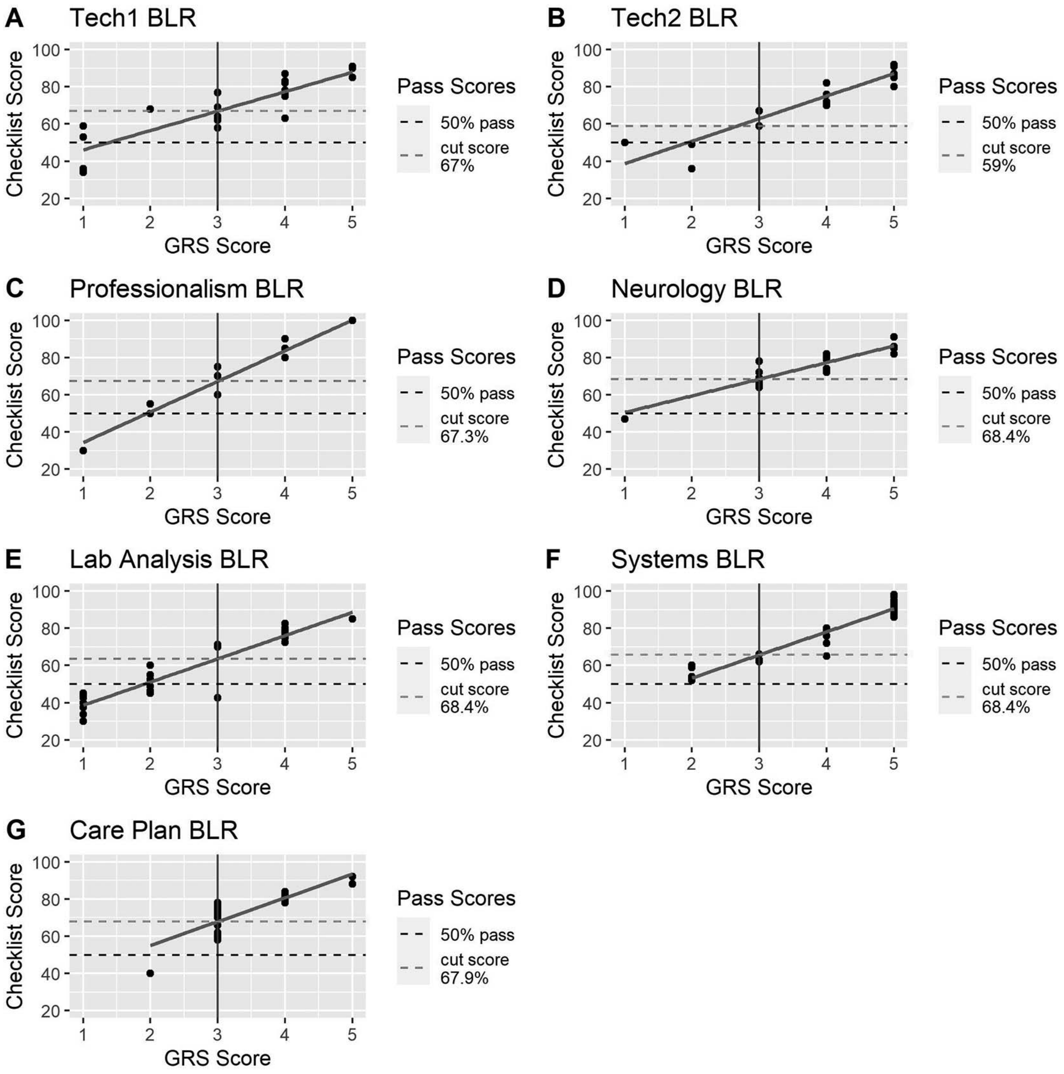

Figure 2 - Plots of all OSCE stations with borderline regression analysis. Panel A shows the Technique 1 (Tech 1) station, panel B the Technique 2 (Tech 2), panel C the Professionalism station, panel D shows Neurology, panel E the Lab analysis station, panel F shows the Systems review (Systems) station, and panel G shows the Care plan station. BLR = borderline regression.

Based on a review of these data, several recommendations are suggested to increase the quality and reliability of OSCE processes for this specific chiropractic program.

- Review the stations with few or many fails for missed concepts or checklist utility.
- Create a GRS specific to each station.
- Use a borderline regression analysis to decide the cut score for each station.
- Increase examiner training to reduce inconsistent marking.

Table 5 - Statistical Metrics for OSCE Stations

Station	Cut Score	R^2	% Fails (at 50%)	% Fails (at Cut)	Intergrade Discrimination	Proportion Variance
Care Plan	67.91	0.67	4.17	33.33	12.76	67.31
Lab Analysis	63.64	0.86	50.00	62.50	12.49	86.34
Neurology	68.42	0.84	4.17	20.83	8.95	84.78
Professionalism	67.28	0.97	16.67	29.17	16.45	95.85
Systems	65.75	0.90	0.00	33.33	12.41	90.30
Technique 1	67.01	0.75	8.33	41.67	10.45	76.44
Technique 2	58.97	0.89	16.67	20.83	14.77	89.35

- Using all metrics discussed above to perform an overall check of station quality and reliability.

In terms of limitations to this study, while the sample size was small it was congruent with the study's aims of statistical assessment valid for small sample sizes,[30,31] but a larger sample may provide more robust findings. Additionally, it should be noted that these data and recommendations are specific to the program that provided the data and should not be generalized to all programs without further research.

Taking a step back from the specifics of these data, what this study has done is to provide a novel method for assessing OSCE quality in smaller programs—describing a number of statistical tests, why they would be used, and a description of their interpretation to show how an analysis could be used to improve small-cohort OSCEs. These metrics provide an objective, evidence-based method to uncover potential problems, whether they be missed concepts, biased examiners, or examinee performance. This battery of tests may provide the first step in serially improving clinical examinations in successive years as each set of metrics could be compared to previous years. Future studies could also compare and contrast OSCE results from other chiropractic programs to determine if the issues identified in this study were unique or universal throughout other programs. Further studies could also review quality changes before and after the recommendations that were informed by this study to illustrate the test batteries use and detail any improvements over time. Additionally, the software (R) used for analysis is free and relatively simple to use, further reducing barriers to faculty seeking to improve their own examinations.

CONCLUSION

This study identified statistical analyses useful for measuring the quality of small-scale OSCEs and used real-life data to illustrate how these analyses could be used to identify examination issues. It also created recommendations to correct problems specific to the dataset and may delineate the pathway to help anticipate future challenges and improve the quality of future examinations.

FUNDING SOURCES AND CONFLICTS OF INTEREST

There were no funding sources or identified conflicts of interest in this study.

About the Authors

Alice Cade (corresponding author) is a senior lecturer and research fellow in the Department of Basic Science at the New Zealand College of Chiropractic (6 Harrison Road, Mt Wellington, Auckland, New Zealand; dralicecade@gmail.com). Nimrod Mueller is the co-head of the Clinic Unit at Barcelona College of Chiropractic (Carrer dels Caponata, 13, 08034 Barcelona, Spain; nimrodmueller@gmail.com). This article was received December 7, 2022; revised February 2, 2023, and May 2023; and accepted July 3, 2023.

Author Contributions

Concept development: AC, NM. Design: AC, NM. Supervision: N/A. Data collection/processing: AC. Analysis/interpretation: AC. Literature search: AC. Writing: AC. Critical review: AC, NM.

REFERENCES

1. Rushforth HE. Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. *Nurse Educ Today*. 2007;27(5):481–490. doi:10.1016/J.NEDT.2006.08.009
2. Cuschieri A, Gleeson FA, Harden RM, Wood RAB. A new approach to a final examination in surgery. Use of the objective clinical examinations. *Ann R Coll Surg Engl*. 1979;61(5):400–405.
3. Kobrossi T, Schut B. The use of the objective structured clinic examination (O.S.C.E.) at the Canadian Memorial Chiropractic College outpatient Clinics. *J Can Chiropr Assoc*. 1987;31(1):21.
4. Russell BS, Hoiriis KT, Guagliardo J. Correlation between student performances on course level integrated clinical skills examinations and objective structured clinical examinations in a chiropractic college program. *J Chiropr Educ*. 2012;26(2):138–145. doi:10.7899/JCE-10-026
5. Ouzts NE, Himelfarb I, Shotts BL, Gow AR. Current state and future directions of the National Board of Chiropractic Examiners. *J Chiropr Educ*. 2020;34(1):31–34. doi:10.7899/JCE-19-24
6. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics AMEE guide no. 49. *Med Teach*. 2010;32(10):802–811. doi:10.3109/0142159X.2010.507716
7. Hawk C, Rupert RL, Hyland JK, Odhwani A. Implementation of a course on wellness concepts into a chiropractic college curriculum. *J Manipulative Physiol Ther*. 2005;28(6):423–428. doi:10.1016/j.jmpt.2005.06.015
8. Harvey P, Goodell K. Development and evolution of an information literacy course for a doctor of chiropractic program. *Commun Inf Lit*. 2008;2(1):6. doi:10.15760/comminfolit.2008.2.1.56
9. Hurley KF, Giffin NA, Stewart SA, Bullock GB. Probing the effect of OSCE checklist length on inter-observer reliability and observer accuracy. *Med Educ Online*. 2015;20(1):1–5. doi:10.3402/meo.v20.29242
10. Barman A. Critiques on the objective structured clinical examination. *Ann Acad Med*. 2005;1(34):478–482.
11. Gupta P, Dewan P, Singh T. Objective structured clinical examination (OSCE) revisited. *Indian Pediatr*. 2010;47(11):911–920. doi:10.1007/S13312-010-0155-6
12. Homer M, Fuller R, Hallam J, Pell G. Shining a spotlight on scoring in the OSCE: Checklists and item weighting. *Med Teach*. 2020;42(9):1037–1042. doi:10.1080/0142159X.2020.1781072
13. Homer M, Pell G, Fuller R, Patterson J. Quantifying error in OSCE standard setting for varying cohort sizes: a resampling approach to measuring assessment quality. *Med Teach*. 2016;38(2):181–188. doi:10.3109/0142159X.2015.1029898
14. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales

in simulation-based assessment. *Med Educ*. 2015;49(2):161–173. doi:10.1111/MEDU.12621

15. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ*. 2005;39(12):1188–1194. doi:10.1111/J.1365-2929.2005.02339.X
16. Lesage E, Valcke M, Sabbe E. Scoring methods for multiple choice assessment in higher education - Is it still a matter of number right scoring or negative marking? *Stud Educ Eval*. 2013;39(3):188–193. doi:10.1016/j.stueduc.2013.07.001
17. Bernardin HJ, Smith PC. A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *J Appl Psychol*. 1981;66(4):458–463. doi:10.1037/0021-9010.66.4.458
18. Pugh D, Smee S. *Guidelines for the Development of Objective Structured Clinical Examination (OSCE) Cases*. Ottawa: Medical Council of Canada; 2013.
19. Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. *Med Teach*. 2015;37(12):1106–1113. doi:10.3109/0142159X.2015.1009425
20. Preusche I, Schmidts M, Wagner-menghin M. Twelve tips for designing and implementing a structured rater training in OSCEs. *Med Teach*. 2012;34(5):368–372. doi:10.3109/0142159X.2012.652705
21. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: A comparison of the modified borderline-group method and the borderline regression method. *Adv Heal Sci Educ*. 2006;11(2):115–122. doi:10.1007/s10459-005-7853-1
22. Hejri SM, Jalili M, Muijtjens AMM, Vleuten CPM Van Der. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *J Res Med Sci*. 2013;18(10):887.
23. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Med Teach*. 2012;34(3):161–175. doi:10.3109/0142159X.2012.651178
24. Atilgan H. Sample size for estimation of G and phi coefficients in generalizability theory. Eurasian J Educ Res, 2013;51:215–227.
25. Tavakol M, Dennick R. Post-examination analysis of objective tests. *Med Teach*. 2011;33(6):447–458. doi:10.3109/0142159X.2011.564682
26. Bujang MA, Omar ED, Baharum NA. A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. *Malays J Med Sci*. 2018;25(6):85. doi:10.21315/MJMS2018.25.6.9
27. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global ratings scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73(9):993–997. doi:10.1097/00001888-199809000-00020
28. Steinhorst RK, Myers RH. Classical and modern regression with applications. *J Am Stat Assoc*. 1988;83(401):271. doi:10.2307/2288958
29. O'Grady KE. Measures of explained variance: cautions and limitations. *Psychol Bull*. 1982;92(3):766–777. doi:10.1037/0033-2909.92.3.766
30. Homer M, Fuller R, Hallam J, Pell G. Setting defensible standards in small cohort OSCEs: understanding better when borderline regression can 'work.' *Med Teach*. 2020;42(3):306–315. doi:10.1080/0142159X.2019.1681388
31. Moreno-López R, Hope D. Can borderline regression method be used to standard set OSCEs in small cohorts? *Eur J Dent Educ*. 2022;26(4):686–691. doi:10.1111/eje.12747