Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-institutional OSCE Quality Assurance in Europe using a mutual assessment strategy; are we equipped for it?

Thomas JB Kropmans (thomas.kropmans@nuigalway.ie)
National University of Ireland Galway College of Medicine Nursing and Health Sciences
https://orcid.org/0000-0001-7071-3266
Eirik Søfteland
Universitetet i Bergen
Anneke Wijnalda
Universiteit Utrecht
Marie Thoresen
Norges Teknisk-Naturvitenskapelige Universitet Fakultet for Medisin og Helsevitenskap
Magnus Hultin
Umea Universitet Medicinska fakulteten
Rosemary Geoghegan
National University of Ireland Galway College of Medicine Nursing and Health Sciences
Angela Marie Kubacki
St Georges University London, School of Medicine
Nick Stassens
Katholieke Universiteit Leuven Universitair Psychiatrisch Centrum KU Leuven - Campus Kortenberg
Suzanne McCarthy
University College Cork College of Medicine and Health
Stephen Hynes
National University of Ireland Galway College of Medicine Nursing and Health Sciences

---

Research article

Keywords: OSCE, quality assurance, cronbach's alpha, SEM, psychometric theory, Standard Error of Measurement

Posted Date: February 4th, 2020

DOI: https://doi.org/10.21203/rs.2.22576/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background

Until 2008, Objective Structured Clinical Examinations (OSCE) were well researched, laborious and costly paper based method of exam delivery restricting international comparison. Cross-institutional comparison of OSCE Quality Assurance in Europe has never been done and due to wide spread electronic assessment analysis is now available.

Methods

Twenty educational institutions across Europe using an electronic OSCE Management Information System where invited of which 8 confirmed to join a mutual comparison of Quality Assurance outcome. Two theories evaluate the quality of the observed test scores, the Classical psychometric theory (*Cronbach's alpha*) and the Generalizability theory. Outcomes for both were compared for all universities including the Standard Error of Measurement (SEM) as well as cut-scores, Pass/Fail score and Global Rating Scores, Cronbach's Alpha and related SEM (68% and 95% CI) and G-theory Coefficients with related absolute and relative SEM (68% and 95% CI).

Results

Outcomes differ between participating universities and observed marks contradict global rating of fail, borderline and excellent performance. G-theory coefficients and Standard Error of Measurement were lower and smaller compared to the classical approach using Cronbach's Alpha as measure of reliability. The Classical psychometric based SEM varies from 2.8% to 11.2% respectively whereas the 95% CI equivalent varied from 9.2% up to 22% (on a 0 - 100% scale). The relative SEM from G-theory analysis varied from 3.1% to 7.0% for criterion-referenced marks, and the absolute SEM for norm-referenced marks varied from 3.8% to 7.8% respectively. The 95% CI around the relative and absolute SEMs values varied from 7.3% to 15.3%. More students failed the examination if the 95% CI is applied to the observed scores.

Conclusion

To protect society and to improve educational decision making, the Standard Error of Measurement and associated confidence intervals needs to be embedded in EU assessment strategies to rule out 'false positive Pass decisions'.

Background

The Objective Structured Clinical Examination (OSCE) is a well-established and researched assessment solution in Medical Education. Ronald Harden first described this type of observational assessment in 1975 [1]. A wide variety of educational research has since assessed the validity and reliability of the OSCE and its role in evaluating students' ability to perform certain clinical/critical tasks competently [2-10]. Since the establishment of the Association of Medical Education in Europe (AMEE) in 1973, generic guidelines

or ‘AMEE guides’ have been published to cover topical issues and support healthcare educators. AMEE guides 81 and 49, part I and II both provide historical background, theoretical framework, organisational and administrative OSCE advice as well as guidance on reviewing exam metrics [11,12].

Up until 2008, OSCEs were administered on paper, a laborious and costly method of exam delivery [13-17] and one which restricted the use of the data captured for research purposes. In 2000, hand-held electronic devices were introduced to capture assessment data and provide automated feedback to institutions and students. In 2012 Kropmans c.s. inaugurally published a paper on an institution-wide web-based OSCE management information system retrieving, storing and analysing OSCE data[18]. This saved 70% of administration cost with a 30% error reduction compared to the previous paper-based process and allowed intra-institutional Quality Assurance (QA) at the School of Medicine in the National University of Ireland, Galway[18]. We are no able to look into cross-institutional QA. In order to build a sustainable electronic solution to retrieve, store and analyse OSCE data in a continuous manner, commercialisation was needed to further develop this advanced assessment solution[19]. Between 2008 and 2018 numerous universities and professional bodies worldwide adopted this software solution to administer their OSCEs electronically [20].

A series of publications has utilized electronic OSCE scoring systems for exam practice and for prospective research[19 ,21,22]. Cross-institutional comparison of student achievements in a progress test have been published for benchmarking purposes [23-26]. Studies evaluating the context between education and assessments largely reported comparisons between different methods of teaching [27,28]. To our knowledge, cross-institutional comparison in medicine, veterinary medicine and pharmacy of OSCE Quality Assurance as part of an European (EU) assessment strategy has never been done and is now possible with the widespread adoption of electronic results recording.

Our main purpose of this study is to share quality assurance (QA) data on clinical skills assessments from OSCEs across European institutions, including using comparisons of Classical and Generalizability theory psychometric measures to evaluate these assessments in a robust and objective manner. Additionally the study aims to exchange best practices with the intention of developing cross-institutional assessments and quality assurance strategies within EU universities.

Methods

We invited twenty educational institutions across Europe already using Qpercom’s electronic OSCE Management Information System (OMIS/Qpercom Observe) to administer, execute and analyse their OSCEs to participate in this retrospective multi-centre comparative study organised by the School of Medicine, domain of Medical Informatics & Education of the National University of Ireland Galway, Qpercom’s home base. We obtained written informed and signed consent from each participating institution, thereby respecting existing mutual Non-Disclosure Agreements, signed by both Qpercom and participating institutions. Eight institutions agreed to participate in this study. We respectfully excluded

those who declined to participate, twelve in all. Eco-authors of each participating university selected their OSCE exams, preferably penultimate/final year OSCEs from term 2017 – 2018 for analysis (Observe 2018, version 1.9.105).

During an OSCE students go through a consecutive series of stations performing a clinical task or scenario in each of the stations within a certain time frame. An examiner is observing each student and marks performance on a list of station items or a more global score sheet. At the end of completion of the assessment sheet in each station, the examiner completes a 3 to 5 item Likert scale ranking observed performance from fail, borderline to pass, good or excellent based on her/his professional opinion of the performance observed. At one of the institutions, in addition to a cut-score based on the sum of items, students that have more than one fail on the Global Rating Score (GRS) fail the OSCE. The latter GRS scores are not incorporated in the item scores but used for retrospective standard setting, utilizing Borderline Regression Analysis or Borderline Group Average. Retrieval, storage and analysis of data can be done on paper and manually or by utilizing an OSCE management information system.

Mainly, two theories exist to evaluate the quality of the observed test scores, the Classical psychometric theory (*Cronbach's alpha*) and the Generalizability theory [29-31]. The primary goal of both is to evaluate the quality of the observed test scores by estimating reliability coefficients (reproducability coefficient for repeated test results) and standard errors of performance measurement. Once the reliability coefficient has been estimated via one of these methods the magnitude of measurement error, expressed in the form of the standard error of measurement (SEM), can be deduced from

$$Standard\ Error\ of\ Measurement = SD\ * \sqrt{(1 - Rxx)}$$

Rxx is the reliability coefficient used

We applied the Classical psychometric approach to calculate the following Quality Assurance (QA) indicators - Cronbach's alpha, 1 SEM (68% Confidence Interval (CI)) and 2 SEM (1SEM*1.96) (95% CI) and at the same time applied the G-theory analysis (a two facet Sx(E:S) design; Students (S) x Examiners embedded in Stations (E:S) design)) to calculate the absolute G-coefficient, the relative and absolute SEM (G-theory) and 68 and 95% CI utilising the statistical software package EDU-G 6.0 for Windows [12,19,32-35].

To review the 95% CI around the relative and absolute SEMs, values need to be multiplied by 1.96 (t-value). The relative G-coefficient is generalizable to other assessments with Criterion-Referenced SEM (relative) and Norm-Referenced SEM (absolute). Criterion-referenced OSCEs compare a person's skills against a predetermined standard, learning goal, performance level, cut-score or other criterion. Norm-referenced measures compare a person's knowledge or skills to the knowledge or skills of the norm group which determines whether a relative or absolute SEM should be applied, see Table 2[(1)].

Statistics

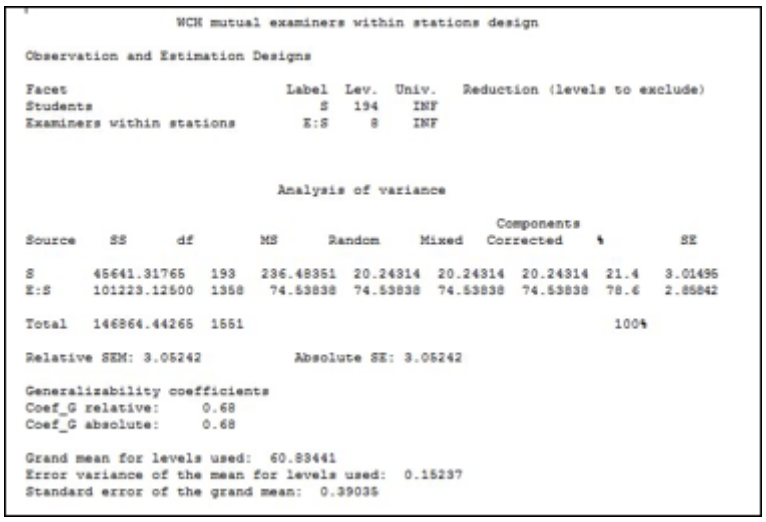
Algorithms for Cronbach's alpha and SEM are embedded in Qpercom Observe (version 2019). Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group (per assessment score sheet and per exam). Though not entirely accurate, it is perceived as 'reliability', whereby in a good assessment the better students should do relatively well across the board (i.e. on the checklist and global scores at each station). The (overall) Cronbach's alpha value that is usually regarded as acceptable in this type of high stakes assessment, whereby standardised and real patients are used and the individual station metrics are not standardised, should be 0.7 or above [12,33] (Figure 1).

The ± 68% CI boundaries around the 'observed score' are supposed to contain the potential 'true score' [29-31]. The SEM*1.96. reflects the boundaries of 95% CI within which the 'true score' can be expected. The generalizability-coefficient as well as the absolute and relative SEM, 95% CI were also calculated utilising the statistical software program EDU-G 6.0 for Windows (Swiss Society for Research in Education Working Group Edumetrics - Quality of measurement in education) and were added as a comparison to the Classical psychometric analysis, with the Cronbach's alpha and alpha derived SEM [32].

Results

Table 1 shows the simplyfied output of statistical software EDU-G for Windows addressing the G-theory analysis utilizing a two facet design with the facets 'Students' and 'Examiners embedded in stations'. The output varies per university in this is summarized in Table 2. Furthermore, table 2 shows the overall outcome of each participating university presenting one of the penultimate or final year OSCEs during term 2017 - 2018. The School of Medicine of the National University of Ireland in Galway examined 192 students in a 2 component Advanced Clinical Skills module comprising 16 stations. The Women and Child Health (WCH) module with 8 stations was included for analysis. The KU Leuven Faculty of Medicine OSCE Ma3 GNK examined 308 students over 51 separate randomly assigned OSCE scenarios. The University of Utrecht held their Veterinary Medicine OSCE, Gebouw Landbouwhuisdieren, over the course of 5 days assessing 24 random scenarios in 7 stations for 208 students.\

Table 1: Simplyfied output of statistical software package EDU-G 6.0 for Windows WCH module NUI Galway 2018 Students x Examiners within Stations design



```
                    WCN mutual examiners within stations design

Observation and Estimation Designs

Facet                          Label  Lev.  Univ.   Reduction (levels to exclude)
Students                           S   194    INF
Examiners within stations        E:S     8    INF


                              Analysis of variance

                                                          Components
Source    SS        df        MS       Random     Mixed   Corrected     %         SE

S       45641.31765   193   236.48351  20.24814  20.24814  20.24814  21.4   3.01495
E:S     101223.12500  1358   74.53838  74.53838  74.53838  74.53838  78.6   2.85842

Total   146864.44265  1551                                             100%

Relative SEM: 3.05242           Absolute SE: 3.05242

Generalizability coefficients
Coef_G relative:     0.68
Coef_G absolute:     0.68

Grand mean for levels used:  60.83441
Error variance of the mean for levels used:  0.15237
Standard error of the grand mean:  0.39035
```

Table 2. Cross-institutional comparison of Quality Assurance data in order of the start of using an electronic OSCE Management Information System. Cut-scores in percentages, Failed station/Global Rating Scale index and (number of stations); Cronbach's alpha; Standard Error of Measurement (SEM) based on alpha with 68% (1*SEM) and 95% Confidence Interval (CI) (1.96*SEM); relative G-coefficient generalizable to other assessments with Criterion-Referenced SEM (relative) and Norm-Referenced SEM (absolute)[(1)]. Umeå University use a combination of a cut score and the Global Rating Scale (2 Fail = fail overall). NTNU don't use an overall cut-score and refer to the computed pass mark per station (at least 6 out 8 stations need to be passed), respectively.

Electronic since	Participating institutions using Qpercom Observe	Overall Cut-score (not applied per station)	Failed Station/GRS index (nr of stations)	Cronbach's alpha	SEM based on Cronbach's alpha 68% - (95%CI)	G-coeff. (relative)	G-theory (relative) SEM 68% (68%CI)	G-theory (absolute) SEM 95% (95%CI)
2009-10	NUI Galway WCH 2018	50%	168/33 5.0 (8)	0.74	2.8 (5.4)%	0.68	3.1(6.0)%	3.1(6.0)%
2010-11	KU Leuven Faculty of Medicine, Ma3 GNK 2018	50%	509/412 1.2 (51)	0.64	9.0 (18.0)%	0.73	4.0(7.8)%	4.4(8.6)%
2010-11	Utrecht University, Vet.Med. LDH 2018	68%	388/294 1.3 (25)	0.54	11.2(22.0)%	0.42	4.5(8.8)%	5.0(9.8)%
2011-12	Umeå University, SKARP OSCE 2 VT18	70%	0/32 0 (14)	0.51	10.4 (20.5)%	0.76	3.7(7.3)%	4.8(9.4)%
2015-16	NTNU MD4042 IIC 2018	–	42/6 7 (8)	0.72	10.6(20.8)%	0.62	3.9(7.6)%	4.5(8.8)%
2015-16	St George University London Final YR 2018	65%	697/33 21.1 (15)	0.86	6.3(12.3)%	0.36	7.0(13.7%)	7.2(14.1)%
2016-17	UCC School of Pharmacy PF4003	50%	5/10 0.2 (4)	0.65	7.6(15.0)%	0.36	6.4(12.5)%	7.8(15.3)%
2017-18	University Bergen, Ordinaer 2018	60%	100/15 6.6 (11)	0.60	7.8(15.4)%	0.56	3.7(7.3)%	3.9(7.6)%

Umeå University's OSCE had 95 candidates passing through 14 stations at the end of 4 years. Norwegian University of Science and Technology Trondheim (NTNU), OSKE MD4042 IIC is 1 out of a 2 module OSCE administering 52 students. St George's, University of London's final year OSCE consists of 15 stations and a total of 222 students were examined this OSCE. Students who deferred their first attempt were excluded from this analysis. University College Cork's School of Pharmacy (pilot) exam contained 4 stations and examined 16 students in total. The newest OSCE of the University of Bergen examined 146 students over 11 separate OSCE stations, at the end of three years medical study (half-way).

Overall cross-institutional reliability expressed as Cronbach's alpha (Classical test theory) of the OSCE and calculated as an average derived from stations internal consistency measurements (as listed in Table 2) varies from poor, to acceptable to good. When expressed using the overall G-coefficient (Generalizability theory) the reliability coefficient varies from unacceptable to acceptable. There is a small but not statistically significant correlation of -0.4 between Cronbach's alpha and G-coefficient. The Classical psychometric based SEM varies from 4.8% to 11.2% respectively whereas the 95% CI equivalent varies from 9.2% up to 22% (on a 0 - 100% scale). The relative SEM from G-theory analysis varies from 3.7% to 7.0% for criterion-referenced marks, and the absolute SEM for norm-referenced marks varies from 3.9% to 7.8% respectively. The 95% CI around the relative and absolute SEMs values varies from 7.3 to 15.3%.

In addition, the differences in cut-off scores between universities are noticeable, varying from 50% to 70%. Different rules also apply between universities where the number of stations to be passed is applied. The NUI-Galway applies a 50% rule if you pass 50% of the stations with an average mark of 50% plus 1 SEM, you pass the exam. Whereas, Umeå University in addition to an overall cut-score applies the rule where each failed station is considered as being a 'red-flag', two red-flags means you have failed the OSCE. St George's, University of London students must achieve the numerical pass mark overall (set using a borderline regression method) and pass at least 65% of the stations. Students who fail 6 or more stations, regardless of their numerical pass mark, will fail the OSCE.

Considering a small confidence interval of 3.7% for 1 SEM (Table 2, University of Bergen) and a cut-score of 60%, all students pass the OSCE on their raw scores. If 1 SEM is applied for this group of students in this exam, everyone will still pass the exam, and only 1 (0.6%) student would fail if 2 SEM is applied. The latter is probably due to OSCEs being new in Bergen and as we have seen elsewhere, high scores will drop over the years. Taking into account an SEM of 7.0% CI (Table 2, St George's University), 11 students would have failed (5%) on a cut-score of 50%, 16 (7.2%) students would have failed if applying a 68% CI, whereas 51 (23%) would have failed the exam applying a 95% CI around the observed raw score. In a smaller university like NUI Galway applying an SEM of 3.1% (the lowest of SEM's presented here) 12 (6.2%) students failed applying a 68% CI and 35 (18.2%) students would fail if a 95% CI was applied around the observed raw score. The SEM and confidence intervals applied to the observed raw scores, specifically those calculated using the G-Theory analysis (relative or absolute SEM) seem to be a more accurate way than any other measure of reliability and are expressed in the same measurement unit as the examination itself.

Discussion

Generally, we provide unique insight for schools of medicine, veterinary medicine and pharmacy in eight European universities. Our main purpose was to compare between Classical and Generalizability theory psychometric measures for Quality Assurance to assess clinical skills in a robust and objective manner. Here the results clearly show which of these measures should be preferred for such QA. Relative or absolute SEM rather than Cronbach' Alpha's SEM enables a more fair discrimination between good

(pass), bad (fail) whereas false positive 'good' performances being the grey 'borderline' zone in between pass and fail decisions should be prevented.

In complex measurement situations such as performance assessment in an OSCE, random and systematic measurement errors may arise from many different possible sources. Which standard error of measurement (SEM) should be preferred to build confidence intervals around the observed scores?[31,35] Since the observed score is partly the product of all sources of measurement errors and their interactions, all standard errors of measurement should be considered when building a confidence interval. Confidence intervals around an observed score indicate the range of scores within which an individual's true score might fall. The larger the reliability coefficient, the lower the variation around the observed score and the narrower the interval will be if examinees were retested. Papers and book chapters of Suen, H.K. (Хои Суен), & Lei, P.W. (Пуи ВА Лей) provide an excellent overview on how item response theory, Classical test theory and generalizability theory evolved over time and more importantly, where they match and differ. Cronbach's alpha and G-coefficients in our study vary between poor and good, and unacceptable and acceptable respectively. Cronbach's alpha, is an overestimation of reliability and therefore obsolete and should no longer be used [31].

More striking than the use of an obsolete measure like Cronbach's Alpha in the software, in our opinion, is the substantial discrepancy throughout most participating Universities comparing the high number of students failing a station on the station item score versus the number of students failing on the Global Rating Scale (GRS), illustrated in the high station Fail/GRS index. Umeå University calculates cut score from the sum of the items in the full OSCE and not at a station level, while only allowing one GRS fail. Thus, students may fail on the GRS only and NOT based on the initial station score e.g. the OSCE score. St Geroge's, sets the standard using borderline group average, utilising the global rating given by examiners for each student. This sets the cut score against which a student's raw score is compared to, to initiate the pass/fail decision. The raw score for each station is made up of 5 domain scores rather than a checklist score. The universities of Leuven and Utrecht are most consistent in their station item score and Fail/GRS index as being 1.2 and 1.3 respectively. Whereas other universities show wide divergence between these two scales and as such this should affect generalisability of the examination results more than internal consistency on its own.

For adjustment of cut-scores from retrospective standard setting, stations item scores and GRS scores are used in a formula for Borderline Regression Analysis [36,37]. Discrepancies between station item scores and GRS results affect the outcome of Borderline Regression Analysis and R^2, and Intergrade discriminators as the slope of the regression curve are affected due to high variation in scores on both scales [37,38]. We think that the latter has serious implications for the validity of our OSCEs. For Umeå Universities, no students failed on the station item scores as the software feature for standard settings of cut-scores setting as such was not utilised (e.g. it is set to be zero instead of 50%, 60% or higher). Umeå University allows one Global Fail and therefore the Global Rating is more important than the actual item score. Pass/fail decisions, then, seem to differ between universities, and it would be recommended to

start a consortium wide discussion based on the available evidence and to determine EU wide pass/fail decision rules as part of an EU assessment strategy.

Applying the SEM reduces the amount of false positive decisions on passing. Applying either a 68% or 95% CI compared to using the raw observed item scores should therefore seriously be considered to improve reliability of decision making. Despite previous recommendations in the literature relating to examiner training[39], of most participants observed marks contradict the global rating of Fail, Borderline or Excellent performance on a 5 item Global Rating Score, for station(s) using either specific items or more global items score sheets. The latter is about their professional judgement of student's performance and apart from Leuven, Utrecht and Umeå, there seems to be an issue with this aspect of rating where examiners are reluctant to fail students in a clinical skills setting using a Global Rating Scale which might affect validity of the decision made. As commonly known, reliability is a pre-requisite of validity and therefore needs to be taken seriously.

Measurement errors need to be taken into account in case of pass/fail decisions[40]. Nevertheless, it is optional to apply the 1 or 2 SEM Confidence Intervals, but this is not often done as far as we are aware. The true scores observed for a particular student vary between upper and lower bound of the 68% e.g. 95% Confidence Interval of the SEM. A 68% confidence interval leaves a 32% chance of letting a student pass and needs to be considered as a significant opportunity for false positive pass decisions, and most certainly if the lower bound of the CI crosses the standard set cut-score with a considerable chance of pass decisions being doubtful. One aspect of the safety net to prevent 'doubtful decisions' could be to consider using the 95% Confidence Interval around the observed raw e.g adjusted scores, reducing the chance of letting students through who should have failed down to 5% instead of up to 32%. Secondly, the number of stations that need to be passed applying the SEM might be considered. Depending on existing local strategies this rule should be applied either at station level or at OSCE level depending on PASS/FAIL regulations in place using the number of stations that need to be passed and/or the overall OSCE final score.

As Suen and Lei addressed, Reckase (1995) elegantly demonstrated rater reliability is not the same as score reliability [41]. Yet, the Classical test theory provides neither a conceptual basis nor a statistical mechanism to assess this reliability from combined sources of measurement error. It is not so much that Classical test theory denies the existence of multiple sources of measurement errors, rather it is that Classical test theory cannot accommodate this fact either conceptually or statistically [31,42]. Moreover, G-theory analysis calculates the G-coefficient as an overall reliability coefficient and estimate relative and absolute SEM directly. Cronbach's alpha very much depends on number of items, stations, subjects and moreover the Standard Deviation in a sample of students going through an assessment design and is still widely used[31].

We compared Cronbach's alpha and G-theory analysis for this study, and the reliability of the G-coefficient 0.53 is substantially lower than the average Cronbach's Alpha coefficient 0.69, as addressed in Table 2 and confirmed in the literature. While no Classical reliability coefficient can consider all defined sources of

measurement errors simultaneously, the G-coefficient does exactly that but is as such not embedded in the software solution used by this consortium. Each of the reliability coefficients, except for Leuven, over-represent the reliability of the scores. The SEM associated with Cronbach's alpha is probably more due to the large variation in the sample (sample SD), number of items in the form and number of students being assessed, rather than due to the amount of error in the assessment design. The G-coefficient in general is a better but lower estimate of reliability, whereas relative and absolute SEM can be applied for criterion and norm-referenced OSCEs to improve educational decision making, respectively.

It is interesting to see Umeå university's 'reliability measures' varying between 1. the lowest Cronbach's alpha and 2. the highest G-coefficient (within this consortium) and that the second lowests SEM of 3.2% of NUI Galway, as 'suggested preferred measure of reliability' is not as such being applied. Utilizing the GRS only as 'cut-off score for PASS or FAIL' entails that the benefits of applying the SEM to rule out 'false positive decisions' is apparently overlooked.

Maybe we medical educators should ask ourselves whether we should implement the SEM to protect society while ruling out 'false positive' PASS decisions in a new era of automated QA analysis? We within this newly established consortium suggest this measure as 'current best practice' realising that this best practice should be embedded in European assessment strategies.

Limitations Of This Study

Qpercom Observe, formerly OMIS, is one of multiple OSCE management information solutions available on the market. In the outline of this study, it was not possible to compare Quality Assurance analysis between different e-assessment solutions. It is most common that only one advanced OSCE solution is used per university. It is most certainly worthwhile to compare the outcome of multiple OSCE management systems being used worldwide[43]. This study is not a random selection of universities or assessment tools. Using shared QA analysis to enhance worldwide applicable assessment strategies appeared to be a politically sensitive operation in the EU. Only eight out of twenty universities chose to participate in this study. Some of the participating universities combine different OSCE settings due to multiple sites, locations or combined clinical skills modules. We only investigated a few of those in this consortium and not all of them throughout pre and clinical years. We hope this paper will open borders and that more universities will be willing to share QA of their clinical examinations and software solutions used.

Conclusions

Despite a large amount of research being available about QA of observational assessments, our comparison between universities is unique in its kind. Measures of QA differ as much as names, number of participants, number of stations and/or scenarios and learning outcomes differ. Cronbach's alpha, for most universities in this paper, overestimates reliability compared to G-theory analysis. The G-theory relative SEM and 68% (1 SEM) and 95% CI (2 SEM) could well be added to QA measures, raising the bar

for criterion-based assessment and probably decreasing the amount of false positive decisions on students receiving a pass on a clinical skills exam. More comparable educational research is required analysing different years of the curriculum, learning outcomes per year of study, and stations properly designed to assess these outcomes between universities. Examiners, in most but not all univerisities, seem to be confused if station raw scores are compared with the Global Rating Scores. Electronic OSCE management information systems are required utilising a full spectrum of instant psychometric analysis. Electronically achieved QA data enable higher and lower level institutions to retrieve, store, analyse and finally, to share observational assessment data to enable and enhance EU clinical skills assessment strategies.

Considering the title of this paper, the question may not be whether we are equipped for it but whether we are ready and willing to share and improve assessment strategies? Confusion between station item scores and Global Rating Scales need to be addressed, and Standard Error of Measurement (95% CI) using relative SEM (G-Theory) * 1.96 to rule out false positive decisions on pass and fail scores needs to be embedded in educational decision making. Prior to sharing stations within and between disciplines, it is crucial to share and improve Quality Assurance strategies within systems and universities.

Abbreviations

BSc = Bachelors in Science

MSc = Master of Science in Education, Sociology and Psychology

PhD = Doctor in Phylosofy

Dr = Doctor

MD = Medical Doctorate

MBBS = Bachelor of Medicine and Surgery

OSCE = Objective Structured Clinical Examination

AMEE = Association of Medical Education in Europe

QA = Quality Assurane

EU = Europe

OMIS = OSCE Management Information System

GRS = Global Rating Scale

SD = Standard Deviation

Rxx = Reliability coefficient

SEM = Standard Error of Measurement

CI = Confedence Interval

S = Students (main facet in generalizability theory analysis)

E:S = Examiners embedded in Stations (main facet in generalizability theory analysis)

EDU-G = Swiss Society for Research in Education Working Group Edumetrics

68% CI = 68 percent confidence interval based on the SEM around the observed score

95% CI = 95 percent confidence interval based on the SEM around the observed score

WCH = Module name Women and Child Health

KU Leuven = Katholieke Universiteit Leuven (Belgium)

NUI Galway = National University of Ireland Galway

Ma3 GNK = Module name Masters year 3 Geneeskunde (Medicine)

VetMed LDH = Module name veterinary medicine Landbouw Huis Dieren (small farm animals)

SKARP OSCE 2 VT18 = Module name School of Medicine Umea University

NTNU MD4042 IIC = Module name year 4 Norwegian University of Science & Technology

YR = year

UCC PF 4003 = Module name University College Cork Pharmacy

G-coefficient = Generalisability coefficient

G-theory = Generalizability theory analysis

GRS = Global Rating Scale

Fail/GRS = Fail/Global Rating Scale Index

R^2 = Reliability x reliability = explaining variability around reliability coefficient

Declarations

Ethics approval and consent to participate

As educational assessment data are analysed completely anonymously and written informed consent was achieved for Quality Assurance analysis from participating institutions, ethical approvals are not required.

Consent for publication

All co-authors approved the current version of the paper being submitted to BMC Medical Education

Availability of data and material

All data and materials were retrieved, stored and suggested for analysis by each co-author of the participating Universities

Competing interests

Dr. Thomas Kropmans, MSc PhD is a senior lecturer in Medical Informatics & Education at the College of Medicine, Nursing & Health Sciences of the National University of Ireland Galway, and co-founder of Qpercom, a spin off of the School of Medicine. He coordinated the cross-institutional research, wrote various drafts of the paper and coordinated the final drafts. Prior to the start of the research, he performed the literature search & review and designed the flow chart for cross-institutional comparison based on the analysis done for each participating institution. He is employed by the National University of Ireland, Galway and is not paid for his activities within Qpercom.

Funding

The research contribution of Stephen Hynes work was funded through the utilisation of an Innovation Voucher from Enterprise Ireland and a research grant of the Department of Medical Informatics and Medical Education of the National University of Ireland, Galway.

Authors' contributions

Dr. Eirik Søfteland MD, PhD, Professor in Anaesthesiology, University of Bergen, Norway and team retrieved, stored and analysed the Norway data and organised internal approval for the mutual paper and suggested Bergen's OSCE for further analysis. Eirik participated during various initial drafts and critically appraised and adjusted the final draft.

Drs Anneke Wijnalda, Utrecht University, Faculty of Veterinary Medicine and team retrieved, stored and analysed the Dutch veterinary medicine data, arranged internal approval for the mutual paper and suggested the LDH OSCE for further analysis. Anneke wrote substantial parts of the initial drafts and critically appraised and adjusted the final draft.

Dr Marie Thoresen, MD, specialist in Internal Medicine and Respiratory Diseases, MMSc Medical Education, Harvard Medical School. Center of Assessment in Medical Education (CAME), Norwegian University of Science & Technology (NTNU) and team retrieved, stored and analysed the Norway NTNU

data, arranged internal approval for the mutual paper and suggested their OSCE for further analysis. Marie critically appraised various draft and reviewed the final draft.

Dr Magnus Hultin, PhD, the program director of the Medical Program Umeå University. Chair of the Swedish Proficiency test for Medical Doctors educated outside EU and EEA. Associate professor of Anesthesiology and Critical Care Medicine and Senior Lecturer at Umeå University Hospital. Magnus was awarded as an Excellent teacher at Umeå University 2018 and selected the Swedish data, arranged internal approval for the mutual paper, adjusted previous and reviewed the final draft.

Dr Rosemary Geoghegan, PhD National University of Ireland Galway, College of Medicine, Nursing and Health Sciences, School of Medicine. Vice Dean for Teaching Learning and Assessment and leads the Teaching, Learning and Assessment unit, in collaboration with central university, developing and delivering the National, University and College Strategic Plan for Teaching, Learning and Assessment. Rosemary and team organised, retrieved, stored and analysed the Galway Ireland year 4 data and suggested the Advanced Clinical Assessment OSCE for further analysis. Rosemary discussed various QA approaches, arranged internal approval for the mutual paper and due course of drafting the paper implemented the 68% CI interval in the NUIG analysis.

Ms Angela Marie Kubacki, Associate Dean for Admissions and Senior Lecturer in Clinical Communication at St George's, University of London, UK and team retrieved, stored and analysed the London data and suggested St George's OSCE for further analysis. Angela organised internal approval and critically appraised and adjusted the final draft.

Nick Stassens, KU Leuven, educational expert and OSCE coordinator at the Skills Center of the Faculty of Medicine. Nick and team retrieved, stored and analysed the Belgian data, organised internal approval for the mutual paper and suggested the Leuven OSCE for further analysis. Nick participated during various drafts, suggested the lay-out of Table 2 to avoid competitional comparison and critically appraised and adjusted the final draft.

Dr Suzanne McCarthy, PhD is a lecturer in clinical pharmacy practice at the School of Pharmacy, University College Cork, pilotted Qpercom Observe in 2017 and as such retrieved, stored and analysed the Cork data. She organised approval for participation and critically appraised the final draft.

Stephen Hynes BSc in Medicine, research student (employed by the university), was paid through an innovation voucher of Enterprise Ireland of the Domian of Medical Informatics & Medical Education of the School of Medicine, National University of Ireland Galway and performed additional analysis and adjusted various first drafts of the paper.

Acknowledgements

The authors wish to acknowledge the approval and written informed consent provided by each participating university. We therefore highly appreciate the effort of all co-authors and their Faculty and Head of School in having gone that extra mile.

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975; **1**(5955): 447-51.
2. Al-Osail AM, Al-Sheikh MH, Al-Osail EM, et al. Is Cronbach's alpha sufficient for assessing the reliability of the OSCE for an internal medicine course? *BMC Res Notes* 2015; **8**: 582.
3. Brown G, Manogue M, Martin M. The validity and reliability of an OSCE in dentistry. *Eur J Dent Educ* 1999; **3**(3): 117-25.
4. Coovadia HM, Moosa A. A comparison of traditional assessment with the objective structured clinical examination (OSCE). *S Afr Med J* 1985; **67**(20): 810-2.
5. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Adv Health Sci Educ Theory Pract* 1996; **1**(3): 227-33.
6. Gledhill RF, Capatos D. Factors affecting the reliability of an objective structured clinical examination (OSCE) test in neurology. *S Afr Med J* 1985; **67**(12): 463-7.
7. Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the OSCE format: reliability and generalizability. *Med Educ* 1996; **30**(1): 38-43.
8. Hodges B. Validity and the OSCE. *Med Teach* 2003; **25**(3): 250-4.
9. Malau-Aduli BS, Teague PA, D'Souza K, et al. A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools. *Med Teach* 2017; **39**(12): 1261-7.
10. Setyonugroho W, Kennedy KM, Kropmans TJ. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Educ Couns* 2015.
11. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: organisation & administration. *Med Teach* 2013; **35**(9): e1447-63.
12. Pell G, Fuller R, Homer M, Roberts T, International Association for Medical E. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Med Teach* 2010; **32**(10): 802-11.
13. Brown C, Ross S, Cleland J, Walsh K. Money makes the (medical assessment) world go round: The cost of components of a summative final year Objective Structured Clinical Examination (OSCE). *Med Teach* 2015: 1-7.
14. Lee CB, Madrazo L, Khan U, Thangarasa T, McConnell M, Khamisa K. A student-initiated objective structured clinical examination as a sustainable cost-effective learning experience. *Med Educ Online* 2018; **23**(1): 1440111.

15. Palese A, Bulfone G, Venturato E, et al. The cost of the objective structured clinical examination on an Italian nursing bachelor's degree course. *Nurse Educ Today* 2012; **32**(4): 422-6.
16. Reznick RK, Smee S, Baumber JS, et al. Guidelines for estimating the real cost of an objective structured clinical examination. *Acad Med* 1993; **68**(7): 513-7.
17. Sudan R, Clark P, Henry B. Cost and logistics for implementing the American College of Surgeons objective structured clinical examination. *Am J Surg* 2015; **209**(1): 140-4.
18. Kropmans T, Donovan BO, Cunningham D, et al. An Online Management Information System for Objective Structured Clinical Examinations. *Computer and Information Science* 2012; **5**(1): 10.
19. Meskell P, Burke E, Kropmans TJ, Byrne E, Setyonugroho W, Kennedy KM. Back to the future: An online OSCE Management Information System for nursing OSCEs. *Nurse Educ Today* 2015; **35**(11): 1091-6.
20. Patricio RHPLM. The Objective Structured Clinical Examination as a performance assessment.: Elzeviers; 2015.
21. Fischer MA, Kennedy KM, Durning S, et al. Situational awareness within objective structured clinical examination stations in undergraduate medical training - a literature search. *BMC Med Educ* 2017; **17**(1): 262.
22. Fernandez C, Guerandel A. An independent appraisal of an electronic OSCEmanagement software system in a Psychiatry examination. MedEdPublish. 27/02/2019 ed; 2019.
23. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, Thoben AJ, van der Vleuten CP. Benchmarking by cross-institutional comparison of student achievement in a progress test. *Med Educ* 2008; **42**(1): 82-8.
24. Pugh D, Bhanji F, Cole G, et al. Do OSCE progress test scores predict performance in a national high-stakes examination? *Med Educ* 2016; **50**(3): 351-8.
25. Pugh D, Hamstra SJ, Wood TJ, et al. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Adv Health Sci Educ Theory Pract* 2015; **20**(1): 85-100.
26. Sattar K, Ahmad T, Khalil MS, Soliman MM, Punnamperuma GG, Abdulghani HM. Medical students' perception of the progress test as a quality-controlled assessment tool for improving learning and teaching, at a public sector medical college in Saudi Arabia. *J Pak Med Assoc* 2017; **67**(2): 300-4.
27. Todres M, Stephenson A, Jones R. Medical education research remains the poor relation. *BMJ* 2007; **335**(7615): 333-5.
28. Archer J, McManus C, Woolf K, et al. Without proper research funding, how can medical education be evidence based? *BMJ* 2015; **350**: h3445.
29. Lee J. Cronbach GCG, Harinder Nanda, and Nageswari Rajaratnam. The Dependability of Behavioral Measurements. Theory of Generalizability for Scores and Profiles.: Wiley, New York,; 1972.
30. Lawson DM. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. *J Manipulative Physiol Ther* 2006; **29**(6): 463-7.

31. Lei KSP-W. Classical versus Generalizability theory of measurement. *Educational Measurement: Issues and Practice* 2007.
32. Cardinet Jcs. EDUG USER GUIDE In: Group SSfRiEW, editor. Edumetrics - Quality of measurement in education. NEUCHATEL – SWITZERLAND IRDP; 2010.
33. 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): 181-8.
34. Pell G, Fuller R, Homer M, Roberts T. Advancing the objective structured clinical examination: sequential testing in theory and practice. *Med Educ* 2013; **47**(6): 569-77.
35. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012; **34**(11): 960-92.
36. Dwyer T, Wright S, Kulasegaram KM, et al. How to set the bar in competency-based medical education: standard setting after an Objective Structured Clinical Examination (OSCE). *BMC Med Educ* 2016; **16**: 1.
37. Yousuf N, Violato C, Zuberi RW. Standard Setting Methods for Pass/Fail Decisions on High-Stakes Objective Structured Clinical Examinations: A Validity Study. *Teach Learn Med* 2015; **27**(3): 280-91.
38. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *Eur J Dent Educ* 2009; **13**(3): 162-71.
39. Schleicher I, Leitner K, Juenger J, et al. Examiner effect on the objective structured clinical exam - a study at five medical schools. *BMC Med Educ* 2017; **17**(1): 71.
40. Severo M, Silva-Pereira F, Ferreira MA. Decision making for borderline cases in pass/fail clinical anatomy courses: the practical value of the standard error of measurement and likelihood ratio in a diagnostic test. *Anat Sci Educ* 2013; **6**(3): 157-62.
41. Reckase MD. The reliability of ratings versus the reliability of scores. *Educational Measurement: Issues and Practice* 1995; **14**(31).
42. Lei PW, Warcholak ND, Suen HK, Williams BL, Magsumbol MS. Empirical Bayes estimation of the prevalence of uninsured individuals by county in the state of Tennessee and analyses of predictive factors. *Eval Health Prof* 2007; **30**(1): 47-63.
43. Harden R, Lilley P, Patricio M. The Definitive Guide to the OSCE, 1st Edition; The Objective Structured Clinical Examination as a performance assessment.: Churchill Livingstone; 2015.

Figures

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Figure 1

Levels of internal consistency using Cronbach's Alpha