1.1 Protocol of a SR of the reliability of the mini-CEX

> “A test, broadly defined, is a set of tasks or stimuli designed to elicit responses that provide a sample of an examinee’s behaviour or performance in a specified domain. Coupled with the test is a scoring procedure that enables the scorer to evaluate the behaviour or work samples and generate a score. In interpreting and using test scores, it is important to have an indication of their reliability.” The *Standards* (1)

1.1.1 Introduction

A sufficient sample size for the evaluation of the reliability of assessment for any measurement method is needed to minimise idiosyncratic random error and allow appropriate application of statistical tests (2)$^{p159-60}$. In 2009 an unpublished analytical review completed to give direction for this thesis identified a problem with the reliability of the mini-CEX (see *Supplement – 2010 Analytical review of the reliability of the mini-CEX*). The reliability of this assessment process was in doubt because the main variance comes from differences in the assessors’ opinions and from residual error. That is, influences that are unrelated to the trainee’s clinical skills and competence have a major effect on the score given for each competency item and overall assessment. The most consistent observation is that the trainee differences contributed least to the variance of the scores across a wide variety of assessment contexts, types of trainees and types of assessors. The series of studies identified at the time (3-17) indicated that rater variation needs to be a focused area for future research to enable the development of reliable and eventually valid clinical competence assessment methods. They also highlighted the need for better research design, methodology and analysis, including “reliable” reporting of evaluation studies for methods of assessment. Finally, the review demonstrated the potential utility of systematic reviews for guiding public policy decisions in medical education.

Systematic reviews (SRs) are important as a way of identifying the existence of a lack of evidence to answer questions and sort out potential problems with methodologies and available evidence[1]. SRs can also provide information for more specific questions and narrow areas of research (18), especially ones that have a utility related answer and/or contribute to theory development (19). Pragmatic and

[1] “Perhaps the major difference is that when scientific standards are applied to the conduct of a research synthesis the weaknesses in evidence become more transparent; simply because weaknesses are less evident in traditional reviews does not mean they are not there.” H Cooper 2010 (18)

implementable reliability information would be useful for the evaluation of any assessment program using assessment results based on the opinion of one person about the competence of another. For example the optimal number needed to provide a reliable assessment score in any one circumstance for any judgement-based assessment often remains uncertain. A published literature review using systematic review methods in 2010 identified and reported the need for at least 10 mini-CEX assessments to achieve adequate reliability (20). In principle, from the evidence at the time it would be reasonable to accept the need for a minimum of 10 assessments to achieve a minimally adequate reliability coefficient in any assessment program.

The accuracy of this recommendation has not been updated in subsequent systematic reviews. Also because of the small number of included studies by the review at the time (20), an effect size with confidence intervals was unable to be calculated.

1.1.2 Primary outcomes

1. A systematic review (20) identified the need for at least 10 assessments to achieve adequate reliability. Therefore the reliability coefficient for 10 assessments will be used as an "effect size".
2. Meta-analysis will be performed using the random-effects model for each selection strategy and all outcomes without pooling. Sensitivity analysis was by repeating the statistics with each study removed. Heterogeneity analysis was performed using tau-squared, the Q-statistic, and I-squared.
3. The pre-specified most important outcomes that add to the validity evidence of reliability for the specific assessment result are described below and include details of how the outcome is defined and measured. All the measurements are performed after the assessment data has been collected at the time of the assessment.
4. The primary outcome measure is the reliability coefficient for 10 assessments. This number was chosen since a literature review using systematic review methodology (20) identified the need for at least 10 assessments to achieve adequate reliability. In addition this outcome measure was chosen with the intent of achieving the review aim of identifying information that is of practical utility to any assessment program and that will allow a measure for quality evaluation, feedback to program directors about comparable reliability and an ability to provide simple benchmarking measures.
5. Secondary outcome measures will be any measured reliability indices as indicated in the Standards for Educational and Psychological Testing (1) will be considered as secondary outcomes.

1.1.3 Domain being studied

The domain being studied is the evidence for the reliability/precision of the mini-CEX, a judgement-related assessment format used in the workplace. This assessment format is used to provide structured feedback based on observed performance for an individual's clinical and professional competencies. The domains of “*Reliability/precision and errors of measurement*” (1) will be evaluated in this review.

1.1.4 Participants/population

A full search was performed with the specific aim of identifying systematic reviews in which the authors evaluated validity evidence for workplace based assessments which include the mini-CEX. The population, “intervention”, comparator and outcome (PICO) format were: (1) the population are medical trainees in any circumstance of medical education and training; (2) the “intervention” is the evaluation process of the reliability and/or validity evidence for the use of the mini-CEX as an assessment method; (3) the areas for analysis and synthesis (“comparators”) are the pre-specified classical reliability and validity measures; and (4) outcomes are the reliability indices and validity measures as indicated in the *Standards* (1) are acceptable for inclusion (Table 1).

Table 1 Inclusion and exclusion criteria

PICO Format	*Study Inclusion Criteria*	*Study Exclusion criteria*
Populations	• Medical trainees in any circumstance of medical education and training	• Non-medical trainees • Studies in medicinal areas not related to humans (e.g. Veterinary) • Studies not involving medical practitioners (eg Dentistry) • Studies not involving humans
The “intervention”	• Evaluation process of the reliability and/or validity evidence for the use of the mini-CEX as an assessment method • Evaluation of multiple types of WBAs but includes the mini-CEX	• Does not include the mini-CEX as one of the main evaluation targets • Description without evaluation • Implementation studies • Protocols for planned studies
The “comparators” are the areas for analysis and synthesis	• Pre-specified reliability and validity measures	• No primary comparative data or not a systematic review of validity measures
The “outcomes”	• Measured reliability and validity measures as in the *Standards* (1)	• No primary data or not a systematic review of validity “outcomes”
Study designs	• Measured reliability indices and validity measures as recommended in the Standards (1)	• Studies without primary data such as commentaries, editorials, non-systematic reviews, letters, editorials, abstracts and dissertation

The “intervention” is an assessment of the clinical competency of a medical trainee by another individual who gives judgement-related scores about the presence of the competency. The mini-Clinical Evaluation Exercise (mini-CEX) is currently a widely used judgement-based assessment used in the workplace for post-graduate medical education and training, as well as for medical students in their learning environments. The mini-CEX assessment method uses the opinion of an assessor to quantify the clinical competence of a trainee using simple specific descriptive items for a number of complex qualities of competence during a single short and specific assessment encounter, usually in the clinical workplace. For the mini-CEX that which is intended to be measured should be demonstrated by a trainee during the assessment and also accurately observed by an observer (assessor). The assessment is applied to an individual in a local medical working environment. The accuracy and reliability of the assessment is dependent not only on the individual trainee but also the context and the assessor.

Therefore the primary research studies to be included for the evaluation process will be examining the reliability/precision evidence for the use of the mini-CEX as an assessment method. This will include studies evaluating multiple types of WBAs and includes the mini-CEX in that evaluation. Studies not included are those that do not include the mini-CEX as one of the main evaluation targets, description studies without new primary data, implementation studies without new data and protocols for planned studies.

1.1.4.1 Types of study to be included

All primary research studies related to or involving evaluating any form of reliability/precision evidence about the mini-CEX interpretation and use are to be appraised with initially no restriction on the type of design. The study designs needed for reliability/precision evidence for the review will be identified during the data-extraction and appraisal. All designs irrespective of the presence of nesting and crossing will be included for the initial full text appraisal.

1.1.4.2 Context

Medical education at all levels of training, including both undergraduate and postgraduate training. All medical educational settings will be included, for example undergraduate, postgraduate, all types of Hospitals, as well as outpatient settings and general practice. All population settings and countries will be included. Context in any form is not an inclusion or exclusion criteria. Cross-sectional and context dependent studies generally are the norm for validity studies involving workplace-based assessments.

1.1.4.3 Reliability measures for inclusion

Inclusion criteria were studies with reliability coefficient (G) and standard error of measurement (SEM) for varying numbers of workplace-based assessments using G and D Studies. Reliability coefficients as ratios and their SEM to be used as effect outcome for each item of the mini-CEX.

The “gold standard” design and analysis to obtain a reliability coefficient is the application of Generalisability Theory to identify and separate the undifferentiated error of classical test theory into the main sources of systematic error in measurement processes. A simplified understanding of this design is that all data is obtained for all possible combinations between examinees, examiners, cases and items in cells of a data matrix without missing variables. Various forms of analysis of variance (ANOVA) can then be performed on the individual data (represented in crossed design matrix cells) without any inseparable overlap between variable data from each individual cell within the matrix. This allows the design of a study to obtain a reliability coefficient with the numerator term only related to the variance due to the subject, with all other variance in the denominator added to the subject variance. Hence the reliability coefficient for that assessment score will indicate the reliability of the assessment to identify only variance due to the subject. The main “outcome measure” was a reliability coefficient suitable for the question and design of the study, and all associated variance components. The other reliability measures described above were considered acceptable if appropriate for the question, study design and data-set. The “gold standard” for the reliability coefficient is that obtained using a Generalisability study from a fully crossed study design. This methodology is more likely to provide a reliability coefficient that measures the variance due to differences in the clinical competence of the junior doctor as the only variable in the numerator of the reliability coefficient, and all other sources of variance in the denominator plus the variance associated with the junior doctors.

1.1.5 Study selection

1.1.5.1 Titles and abstracts selection

Titles and abstracts were reviewed with broad inclusion and exclusion criteria applied for full article retrieval: (1) any study undertaking any form of research or evaluation of the mini-CEX; and (2) all papers with any form of reliability measure and a population of any type of medical trainee. Exclusion criteria were non-systematic review articles, editorials, qualitative studies, and general or implementation discussions. The full papers for non-systematic reviews were obtained to view the bibliographies for relevant articles. The full text of all the remaining papers were obtained and read for a second round of inclusion and exclusion criteria application in view of the initial broad inclusion criteria. Any relevant study, in any language, and from any time period (dating from the beginning of each data-base) were to be included. The inclusion criteria in the second round selection process were to limit articles to primary research, those having any form of reliability estimates of the mini-CEX and to any type of medical practitioner or trainee. The reliability measures that were considered acceptable included: the Intraclass correlation coefficient (ICC)(21), Pearson correlation(21), Kappa coefficient in one of its recognised forms(22)^{p434,}(23)^{p147,}(21), Generalisability Theory analysis(21;24-28), internal consistency for the appropriate question and design(21;29;30), inter-rater reliability(31)^{p142,}(21)^{p137}, intra-rater reliability(21), test-retest reliability(30)^{p201,}(21), stability over time(32)^{p117,}(21), and/or the method of measuring agreement of Bland and Altman for the right study question(33).

1.1.5.2 First full text study selection

Titles and/or abstracts of studies retrieved using the search strategy and those from additional sources will be screened independently by two review authors to identify studies that potentially meet the inclusion criteria outlined above. The full text of these potentially eligible studies will be retrieved and independently assessed for eligibility by two review team members. Selection of studies will be performed independently by two coders (DM and JS). Disagreement will be resolved by discussion after independently stating the reasons for the choice. If disagreement persists then a third reviewer will arbitrate. Study selection and data extraction will not be blinded to authors and journal. Standardised, pre-piloted digital tables will be used to extract data from the included studies for assessment of study quality and evidence synthesis.

1.1.5.3 Specific question related full text study selection

All full-text studies selected will be evaluated for the presence of indicators of measures reliability and precision. Studies having one or more of the expected measurements will be included for full study appraisal and analysis.

1.1.6 Study Quality and bias potential

Although standards for Educational and Psychological Testing have been established and continually reviewed for many decades (34), an equivalent process for assessment in medical education has not been well promulgated. Nevertheless, an optimal and substantive reliability study has been identified as needing a number of measures with support information (35). The set of requirements identified by Crossley and colleagues are to ensure that(35): (1) all relevant factors are sampled, and that the sampling meets Generalisability theory's assumption that the conditions represent a random and representative sample of the factor's 'universe'; (2) large samples of each factor to allow adequate generalization; (3) a Generalisability study and a Decision study using a Generalisability Study design; (4) for unbalanced designs and randomly missing data, either urGENOVA(26), or one of minimum norm quadratic unbiased estimator, (MINQUE), maximum likelihood or restricted maximum likelihood(36); (5) the mathematical procedure used and the degrees of freedom of the effect estimates are stated; (6) the regression model used is described and justified; (7) raw variance components are presented; (8) standard error of measurement (SEM) and/or reliability coefficients are presented, with the equations used to calculate them(30)p202,(21)p142; (9) the method of reporting reliability (precision or discrimination) is appropriate to the purpose of the assessment with a precision indicator such as SEM and the associated confidence intervals (CI)(30),(35); and (10) confidence intervals (CIs) for the reliability coefficient are reported(37). In addition statistical issues related to test reliability evaluation including consideration for the assumptions made for statistical purposes when applying ANOVA or correlation(38). A sufficient sample size for the evaluation of the reliability of assessment for any measurement method is preferable to minimise idiosyncratic random error and allow appropriate application of statistical tests(2)$^{p159-60}$. Variations on the fully crossed design should be interpreted in the light of the potential confounding and bias, with subsequent inferences confined to the imitations due to the design. Because of the potential for bias in workplace evaluations, all variance components should be fully reported to substantiate any inferential claims about reliability of an assessment.

1.1.6.1 Study Quality and potential for bias for G and D Generalisability Theory studies

A generalizability (G) study is used to separate and estimate variation in a trainee's assessment result due to the trainee. The design also separates out as many other causes of variation of the measurement as is feasible within the limitation of sample size and method design. An unbalanced nested study design is often the only feasible design for workplace-based assessments, so the total variance measured is that due to the trainees plus variance not due to the trainees, that is, measurement error (24;39). Each characteristic of the measurement situation, for example the test form, test item, rater, test occasion, and so on, is called a "facet" in Generalisability Theory (39). Facets can be a cause of variation for the assessment result on each occasion. In an unbalanced nested design the variance from the facets are all included in the error measurement and cannot be separated, including their interactions (26;40).

A decision (D) study uses the information provided by the G-study to design the best possible application of the measurement for a particular purpose, including the number of assessments needed to maximise the variance of the assessment result due to the trainee and minimise the error variance due to the facets. The increase in score variance due to the trainee from an increase in sampling is expressed by a calculated G-coefficient from the original data. When planning a D-study for an unbalanced and nested design, the population to which the results refer is defined, the facets causing error variance from the measurement conditions are defined, and specifies of the proposed interpretation of the measurement is made clear (39). The decision maker uses the information from the G-study to evaluate the effectiveness of alternative designs for minimizing error and maximizing reliability of the assessment result for any object of measurement (the trainee) (39). D-studies can use nested designs for determining sample size which will reduce the estimated error variance and hence increases the estimated generalizability (39). So from the original data an initial G-Study the G-coefficient as the reliability coefficient will be determined. The D-Study will provide the method to calculate the number of assessments that would provide a G-coefficient of an acceptable level. For this review the acceptable reliability coefficient will be ≥ 0.80. For other designs, additional types of information could also be obtained about other facets such as the type of rater, the test form and type of trainee (24;26;39;40).

Although standards for Educational and Psychological Testing have been established and continually reviewed for many decades (34), an equivalent process for assessment in medical education has not been well promulgated. Nevertheless, an optimal and substantive reliability study has been identified as needing a number of measures with support information (35). The set of requirements identified by Crossley and colleagues are to ensure that(35): (1) all relevant factors are sampled, and that the sampling meets Generalisability theory's assumption that the conditions represent a random and representative sample of the factor's 'universe'; (2) large samples of each factor to allow adequate generalization; (3) a Generalisability study and a Decision study using a Generalisability Study design; (4) for unbalanced designs and randomly missing data, either urGENOVA(26), or one of minimum norm quadratic unbiased estimator, (MINQUE), maximum likelihood or restricted maximum likelihood (36); (5) the mathematical procedure used and the degrees of freedom of the effect estimates are stated; (6) the regression model used is described and justified; (7) raw variance components are presented; (8) standard error of measurement (SEM) and/or reliability coefficients are presented, with the equations used to calculate them (30)[p202,](21)[p142]; (9) the method of reporting reliability (precision or discrimination) is appropriate to the purpose of the assessment with a precision indicator such as SEM and the associated confidence intervals (CI) (30;35); and (10) confidence intervals (CIs) for the reliability coefficient are reported (37). In addition statistical issues related to test reliability evaluation including consideration for the assumptions made for statistical purposes when applying ANOVA or correlation (38).

1.1.6.2 Specific question related full text study selection

Inclusion criteria were studies with a reliability coefficient (G) and standard error of measurement (SEM) for varying numbers of workplace-based assessments using G and D Studies. Reliability coefficients as ratios and their SEM are to be used as effect outcome for each item of the mini-CEX. A systematic review (20) identified the need for at least 10 assessments to achieve adequate reliability. Therefore the reliability coefficient for 10 assessments will be used as the primary outcome measure. This number was chosen since a literature review using systematic review methodology (20) identified the need for at least 10 assessments to achieve adequate reliability. However, in an unpublished review of studies up until the end of 2009, the number of mini-CEX assessments needed to achieve an $R \geq 0.80$ ranged from 5 to 50 assessments (see *Supplement - 2010 Narrative review of the reliability of the mini-CEX*).

The specific outcome measure was chosen with the intent of achieving the review aim of identifying information that is of practical utility to any assessment program and that will provide a measure for quality evaluation, feedback to program directors about comparable reliability and an ability to provide simple benchmarking measures. Studies having one or more of the expected measurements will be included for full study appraisal and meta-analysis.

1.1.7 Data extraction (selection and coding)

General extracted information will include: the variables and characteristics detailed in Tables. Two review authors will extract data independently (DM and JS). Discrepancies will be identified and resolved through discussion (with a third author where necessary). Authors of eligible studies will be contacted to provide missing or additional data if it is relevant to an important synthesis whether narrative or meta-analytic.

1.1.8 Strategy for data synthesis

A quantitative and descriptive synthesis is planned.

1.1.8.1 Narrative synthesis

A narrative synthesis of the findings from the included studies will be structured around the context of the assessment, population characteristics of assessor and assessee, and the content of the outcomes as detailed in the data-extraction details and tables.

1.1.8.2 Meta-analytic assessment information

We anticipate that there may be limited scope for meta-analysis because the examination of internal structure of judgement-based assessments such as the mini-CEX has not been commonly employed (41). However if more recent studies have examined this aspect of validity evidence then a random-effects model will be used given the potential for methodological variability across studies (42).

Reliability coefficient (G) and standard error of measurement (SEM) from Generalisability Theory studies will be the effect size evaluated for a meta-analysis. The reliability coefficient (a ratio) and SEM to be used as effect outcome for each item of the mini-CEX. As described above the arbitrary number chosen was the reliability coefficient for 10 assessments for the outcome. However because the ideal number remains unknown, the primary studies were also evaluated for other numbers of assessments needed for which sufficient data is available. The intent was to identify the number of assessments needed for the summed mean score, the overall performance score and scores for each individual competency.

Aggregate participant data was to be used. A meta-analysis would require appropriate data described above and the measures from the available studies be sufficiently homogeneous. The software for the data-synthesis will be *Comprehensive Meta-analysis Version 2.2.064*.

The meta-analysis was to be performed using the random-effects model for each selection strategy and all outcomes without pooling. Sensitivity analysis was by repeating the statistics with each study removed.

Heterogeneity analysis was to be performed using tau-squared, the Q-statistic, and I-squared. Heterogeneity for pooled measures for the meta-analysis was to be assessed by computing the χ^2 statistic (Cochrane's Q), and with p values < 0.10 as indicative of statistically significant heterogeneity (43). Any observed heterogeneity was also evaluated by computing the I^2 statistic and will consider values of I^2 ≤25%, ≤50%, and ≤75%, as indicative of low, moderate, and high degrees of heterogeneity, respectively (44). The potential for significant heterogeneity was anticipated as shown in a previous systematic review of validity evidence for criterion validity (45).

Sensitivity analyses based on study quality and bias potential was planned to be based on stratified meta-analyses to explore heterogeneity in effect estimates according to study quality, study populations and the contexts of assessments. Evaluation of evidence for the possibility of publication bias used the traditional funnel-plot method (43). However, since the funnel plot is a simple scatter plot of the intervention effect estimates from individual studies against some measure of each study's size or precision (43), it is anticipated that it may not be useful if small numbers of studies are only identified. It will be used to identify the presence of outliers.

1.1.8.3 The number of assessments for an acceptable minimum reliability

The number of assessments for a minimum acceptable reliability is the number of assessments for an acceptable minimum adequate reliability level of R = 0.80 (NAAMAR)[2] was to be calculated for each study for which either variance components or a *G*-coefficient for one study is available.

1.1.8.4 Analysis of subgroups or subsets

If sufficient studies are available analysis of subgroups including the professional type, the level of training, and the gender of the assessor and trainee will be performed. Similarly different settings such as country, acute or primary care sector, professional or family care and different types of study (e.g. randomised or non-randomised) will be evaluated if possible.

[2] NAAMAR = $= \sigma^2_{\text{error}} / \{(\sigma^2_{\text{subjects}}/0.8) - \sigma^2_{\text{subjects}}\}$

Reference List

(1) Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association tAPAatNCoMiE. Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association; 2014.

(2) Wiswanathan M. Measurement Error and Research Design. Thousand Oaks: Sage Publications; 2005.

(3) Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med* 1995;**123**(10):795-9.

(4) Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine* 2002;**77**(9):900-4.

(5) Boulet JR, McKinley DW, Norcini JJ, Whelan GP. Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Advances in Health Sciences Education* 2002;**7**(2):85-97.

(6) Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: A method for assessing clinical skills. *Ann Intern Med* 2003;**138**(6):476-81.

(7) Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine* 2003;**78**(10 Suppl):S33-S35.

(8) Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Medical Education* 2006;**40**:950-6.

(9) Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P, Hawkins RE. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: a validity study. *Academic Medicine* 2006;**81**(10 Suppl):S56-S60.

(10) Alves de Lima A, Barrero C, Baratta S, Costa YC, Bortman G, Carabajales J, Condez D, Galli A, DeGrange G, van der Vleuten CPM'. Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. *Med Teach* 2007;**29**:785-90.

(11) Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education* 2008;**42**(4):364-73.

(12) Nair BR, Alexander HG, McGrath BP, Parvathy MS, Kilsby EC, Wenzel J, Frank IB, Pachev GS, Page GG. The mini clinical evaluation exercise (mini-CEX) for assessing clinical performance of international medical graduates. *Medical Journal of Australia* 2008;**189**:159-61.

(13) Cook DA, Beckman TJ. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Advances in Health Sciences Education* 2008;**DOI 10.1007/s10459-008-9147-x**(Published on line 26th November 2008).

(14) Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *Journal of General Internal Medicine* 2008;**24**(1):74-9.

(15) Davies H, Archer J, Southgate L, Norcini J. Initial evaluation of the first year of the Foundation Assessment Programme. *Medical Education* 2009;**43**(1):74-81.

(16) Weller JM, Jolly B, Misur MP, Merry AF, Jones A, Crossley JG, Pedersen K, Smith K. Mini-clinical evaluation exercise in anaesthesia training. *British Journal of Anaesthesia* 2009 May;**102**(5):633-41.

(17) Hill F, Kendall K, Galbraith K, Crossley J. Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. *Medical Education* 2009;**43**:326-34.

(18) Cooper H. Reseaerch Synthesis and Meta-Analysis. A Step-by-Step Approach. 5th ed. Los Angeles: Sage; 2010.

(19) Schmidt FL, Hunter JE. Methods of Meta-Analysis. Correcting Error and Bias in Research Findings. Third ed. Thousand Oaks, California: Sage Publications Inc; 2015.

(20) Pelgrim EA, Kramer AW, Mokkink HG, van den Elsen L, Grol RP, van der Vleuten CP. In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education Theory and Practice* 2009;**10.1007/s10459-010-9235-6.**

(21) Streiner DL, Norman GR. Health Measurement Scales. 3rd ed. Oxford: Oxford University Press; 2003.

(22) Kirkwood BR, Sterne JAC. Essential Medical Statistics. 2nd ed. Oxford: Blackwell Science; 2003.

(23) Bowling A. Research Methods in Health. Investigating Health and Health Services. 2nd ed. Maidenhead Berkshire: Open University Press; 2002.

(24) Gleser GC, Cronbach LJ, Rajaratnam N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika* 1965;**30**(4):395-418.

(25) Shavelson RJ, Webb NM, Rowley GL. Generalizability Theory. *American Psychologist* 1989;**44**:922-32.

(26) Brennan RL. Generalizability Theory. New York: Springer-Verlag; 2001.

(27) Shavelson RJ, Webb NM. Reliability Coefficients and Generalizability Theory. In: Green JL, Camilli G, Elmore PB, `, editors. Handbook of Complementaty Methods in Education Research.Washington DC: American Education Research Association; 2006. p. 309-22.

(28) Crossley J, Davies H, Humphries G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;**36**:972-8.

(29) Downing SM. Reliability: on the reproducibility of assessment data. *Medical Education* 2004 Sep;**38**(9):1006-12.

(30) McIntire SA, Miller LA. Foundations of Psychological Testing. A Practical Approach. Second ed. Thousands Oaks: Sage Publications; 2007.

(31) Westgard JO, Darcy T. The truth about quality: medical usefulness and analytical reliability of laboratory tests. *Clinica Chimica Acta* 2004;**346**(1):3-11.

(32) Cohen L, Manion L, Morrison K. Research Methods in Education. 5th ed. London: RoutledgeFalmer; 2000.

(33) Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**(8476):307-10.

(34) Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association tAPAatNCoMiE. Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association; 1999.

(35) Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, Norcini J. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Medical Education* 2007;**41**:926-34.

(36) Baltagi BH, Song SH, Jung BC. A comparative study of alternative estimators for the unbalanced 2-way error component regression model. *Econometrics Journal* 2002;**5**(2):480-93.

(37) Ebel RL. Estimation of the reliability of ratings. *Psychometrika* 1951;**16**:407-24.

(38) Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall/CRC; 2004.

(39) Webb NM, Shavelson RJ, Haertel EH. Reliability Coefficients and Generalizability Theory. In: Rao and CR, editor. Handbook of Statistics Psychometrics. Volume 26 ed. Elsevier; 2006. p. 81-124.

(40) G String IV User Manual [computer program]. Version Version 6.1.1. Hamilton, Ontario, Canada: Ralph Bloch & Geoff Norman; 2015.

(41) Sandilands D, Zumbo BD. (Mis)alignment of medical education validation research with contemporary validity theory: The Mini-CEX as an example. In: Zumbo BD, Chan EKH, editors. Validity and validation in social, behavioral, and health sciences.Cham, Switzerland: Springer International Publishing; 2014. p. 289-310.

(42) Kogan JR, Holmboe ES, Hauer KS. Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review. *JAMA* 2009;**302**(12):1316-26.

(43) Cochrane Handbook for Systematic Reviews of Interventions 5.1.0 (Updated March 2011). Available from www cochrane-handbook org 2015

(44) Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003 Sep 6;**327**(7414):557-60.

(45) Al Ansari A., Ali SK, Donnon T. The construct and criterion validity of the mini-CEX: a meta-analysis of the published research. *Academic Medicine* 2013 Mar;**88**(3):413-20.