APPENDIX D
STANDARD SETTING PAPER MATRIX

Author (ID)	**Description**
Bandaranayake RC. (2008) Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37 *Medical Teacher,* 30(9-10): 836-845.	Not so much a full review as an introduction and simple practical guide. See de Champlain's (2010) trenchant comments.
Barman A. (2008) Standard setting in student assessment: is a defensible method yet to come? *Annals of the Academy of Medicine, Singapore,* 37(11): 957-963.	A basic review of standard setting practices.
Beard JD. (2005) Setting standards for the assessment of operative competence. *European Journal of Vascular and Endovascular Surgery,* 30: 215-218.	Blinded videos of saphenofemoral disconnection by an experienced surgeon (ie competent) and by inexperienced trainees (ie not yet competent) were scored using a checklist by 14 experienced vascular surgeons and 14 vascular trainees. Observers asked to decide whether surgeons competent, borderline or not competent. 13 vascular operating room nurses (OR) also observed and asked to judge. The contrasting group method was used to compare the cut point scores. Contrasting group method Clear separation between the surgeons and the trainees scores for the inexperienced trainee (median 16, range 13-18, the inexperienced surgeons median 6.5, range 2-12), p=0.001. Judgements for competent 14-18, borderline 15-7, not competent 8-2, p=0.0001. OR nurses rated similar to the surgeons. Experienced= Median 18, range 14-16, compared to inexperienced surgeons = median 7, range 2-14. Highly significant difference p=0.0001. Judgements on competence similar.
Beard JD, Educ, et al. (2005) Setting standards for the assessment of operative competence. *European Journal of Vascular and Endovascular Surgery,* 30(2): 215-218.	Contrasting groups used to define competent, borderline, and not competent for ratings of video of experienced trainee versus inexperienced trainee. Checklist plus global ratings. High degree of discrimination whether by experienced surgeons, trainees, or OR nurses. Almost no overlap between groups. Possible evidence of construct validity of test, and possibly of positive relationship between scores and performance.
Boursicot KAM, Roberts TE, Pell G. (2007) Using Borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education,* 41: 1024-1031.	Quant. 3 UK medical schools graduation exam. 6 station OSCE using Borderline group method (BGM) or Borderline Regression Method (BRM) to generate pass marks. Compared pass rates across the 3 schools and the pass rates within the schools with previous pass marks which had used the Angoff method. Half day workshop for examiners on BGM/BRM. Borderline group method/Borderline Regression method Angoff BGM used a 3 point rating scale (pass, borderline, fail) BRM used a 5 point rating scale (fail, borderline, pass, v good pass, excellent). Z scores were used to compare student score means between schools for each station. To compare school

	station pass rates used binominal distribution of pass/fail to compute z scores. 2 tailed test to convert z scores into p-values. Findings: Two schools (A and C) consistent in their pass marks but the third school (B) pass marks lower. Previous Angoff scores when compared with BGM scores did not show significant difference across A and B schools but showed intra-institutional consistency re: standard of borderline candidates. However the actual differences varied between 0.56 and 8.97 – practically may affect number of students passing. In school C significant differences between Angoff and BRM pass marks across all stations – BRM pass marks higher than Angoff pass marks. May suggest differences between Angoff and BRM examiners conceptualisation of borderline students. Conclusion: Even when using different standard setting methods standards set at different schools can be different. Could be that school B the examiners accepted lower level of minimum competence or some differences in collective standards of examiners at different schools and differences in what is borderline or minimally competent.
Boursicot KAM, Roberts TE, Pell G. (2006) Standard Setting for Clinical Competence at Graduation from Medical School: A Comparison of Passing Scores Across five Medical Schools. *Advances in Health Sciences Education,* 11: 173-183.	Quant. 6 OSCE station. Each of the 5 medical schools set their own pass score using Angoff method. (Descriptors were decided upon by the examiners in the training-workshop to help examiners in the Angoff process). Descriptors used: best student, worst student, a strong student who was likely to pass, a weak student who was likely to fail, minimally competent student who might just pass or just fail. Analysis asking 'to what extent do raters from different schools agree on the relative difficulty of different stations?' using one-way ANOVA Analysis also looked at inter-station reliability of stations passing scores by looking at agreements of examiners for all stations in all med schools. Angoff Method Probability score for each station varied: 10% to 75%. Inter-quartile range: 20%-40%. Overall aggregated passing score varies by 13% between highest and lowest scoring med school. Inter-medical school correlations: between 0.7 and 0.9. High reliability score Intra-class coefficient for the 6 stations across the 6 schools was 0.316 indicating poor inter school reliability – the low value being largely due to differences in mean school scores. Conclusions: Passing scores varied greatly across the 5 schools despite the use of the same standard setting method and stations. Students with the same level of competency could pass or fail depending upon which medical school they attend.

	Only 6 OSCE stations used so reliability low. High correlation scores of 0.9 gives high measure of reliability. Previous experience of using the Angoff method doesn't mean there would be uniformity across medical schools. Limitation of using Angoff method: training may not be sufficient in a short time, items should be independent of each other not dependent upon the topic of the individual stations.
Burch VC, Nash RC, Zabow T, Gibbs T, Aubin L, Jacobs B, Hift RJ. (2005) A structured assessment of newly qualified medical graduates. *Medical Education,* 39: 723–731.	Study of 58 graduating doctors in South Africa. Administered 7 station OSCE, Angoff cut score of 85%, all failed. The examiners were not invited to attempt the OSCE (see Brian Jolly Commentary).
Chesser AMS, Laing MR, Miedzybrodzka, Brittenden J, Heys SD. (2004) Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. *Medical Education,* 38: 825-831.	Quant. 8 examiners (who underwent training) across 12 station OSCE to 192 undergraduate medical students. Students who failed standard on 3 or more stations failed exam. Modified Angoff Method SPSS (V10.) Reliability estimated by calculating Cronbachs' alpha coefficient (0.55) Used factor analysis to analyse scores to ensure that skills are assessed across all OSCE stations equally so that one station doesn't have undue influence i.e. the compensatory approach, so if failed 3 of the 12 stations irrespective of whether they did well on one station they failed outright). Eigenvalue set at >1 Cut off score: pass/fail Findings: Non-compensatory approach 20 (10.4%) failed the exam. Conclusions: the use of factor scores has the potential to combine the strengths and weaknesses of the compensatory and non compensatory approaches to standard setting.
Clauser BE, Nungester RJ. (2001) Classification accuracy for tests that allow retakes. *Academic Medicine*, 76: S108-110.	False positive "errors may put the public at risk by allowing unqualified candidates to become licensed or certified". The false positive rate increases as reliability decreases, and decreases as cut score increases, and hence fail rate increases. "One important strategy is to limit the number of retakes". Another is to increase the initial cut score (as in the GMC approach), especially where the cost of a false positive is higher than that of a false negative. Finally, the paper explores the consequence of raising the cut score for resits, as suggested by Millman (1989). The effect of increasing the cut score by 0.25 SD per administration is considered, along with possible resistance to this approach.
Clauser BE, Clyman SG. (1994) A contrasting-groups Approach to standard setting for Performance Assessments of Clinical Skills. *Academic Medicine,* 69: 10, Oct suppl. S42-S44.	Evaluation of the resulting standard. 280 3rd yr med students who completed a 7 stage CBX test in medicine. Standard setting used committees of clinicians (non associates with the students). Individual committee members were asked to make transaction lists for pass or fail performances- based on what they thought the appropriate standard should be for end of 3rd yr students. These were discussed and committee members could change their score but did not have to. If

	no unanimous pass rate was reached a pass rate was given if 4 out of the 6 judges passed it. The process was repeated with a second committee 6 months later was repeated with 5 raters' (3 out of 5 judges had to agree on pass mark). Variation on Ebel, norm-referencing contrasting-groups method. Kappa coefficients representing agreement between first and second committee meetings. Kappa Co efficents representing agreement between decisions made by the committee and those by the cut off point score are .95, .92, .81 for the 3 cases examined. Failure rates across full sample (280) 24%, 21%, 22% for the 3 cases respectively. When the cut-off scores (using 40 of the transactions lists) were re-estimated with the second committee the rates remained unchanged for case 1 and 2 – 280 examinees received identical classifications across occasions. Case 3 – failure rate of 10% = 88% examiners being identically classified across occasions. When placement cut-off scores was re estimated based on all 140 transaction lists used at the second committee failure rates for the 3 cases were 24%, 28% and 18%. Percentage of examinees identical classifications across occasions were 100%, 93% and 96%. High level of agreement within judges across occasions and across both committees. Contrasting-group approach to standard setting has potential for use with case based performance.
Clauser BE, Harik P, Margolis MJ, McManus IC, Mollon J, Chis L, Williams S. (2009) An empirical examination of the impact of group discussion and examinee performance information on judgements made in the Angoff standard-setting procedure. *Applied Meas Educ*, 22: 1-21.	Discussion reduced discrepancies between Judges in an Angoff procedure but did not increase strength of relationship between Judgements and conditional item difficulties. Use of performance data increased this relationship. Analysed against item difficulty, Judges tend to overestimate success on difficult items and underestimate difficulty on easy items. There is an informative discussion on this point.
Cohen-Schotanus J, CPM van der Vleuten. (2010) A standard setting method with the best performing students as point of reference: practical and affordable. *Medical teacher,* 32(2): 154-160.	Standard setting on 60% of the 95th percentile score student. Reduces variation in cut off scores and fail rates. Assumes this is desirable. Essentially a normative approach, even though the authors describe it as a compromise method.
Cusimano MD. (1996) Standard setting in medical education. *Academic medicine: journal of the Association of American Medical Colleges,* 71(10 Suppl): S112-120.	High quality review article covering methods, outcomes and variability.
Cusimano MD, AI Rothman. (2003) The effect of incorporating normative data into a criterion-referenced standard setting in medical education. *Academic medicine: journal of the*	Hofstee compared to Angoff and Ebel. Hofstee gave 'most realistic' fail rate (highest, at 29%!) and highest precision (modified Jaeger measure: 2.9 compared to 22.6 and 6.0 respectively).

Association of American Medical Colleges, 78(10 Suppl): S88-90.	
Cusimano MD, AI Rothman. (2004) Consistency of standards and stability of pass/fail decisions with examinee-based standard-setting methods in a small-scale objective structured clinical examination. *Academic medicine: journal of the Association of American Medical Colleges,* 79(10 Suppl): S25-27.	Comparison of Borderline groups and Contrasting groups in OSCE (also 'traditional' 60% cut score). Borderline groups gave lower cut score (55.2% vs 61.7%) and higher pass rate (23.6% vs 32.9%) but better Jaeger indices of consistency and better stability.
Dauphinee WD, Blackmore DE, et al. (1997) Using the Judgments of Physician Examiners in Setting the Standards for a National Multi-center High Stakes OSCE. *Advances in Health Sciences Education*, 2(3): 201-211.	Practical account of using Borderline groups in the MCC Qualifying Examination part 2 OSCE over 2 administrations. Borderline Pass and Borderline Fail categories were aggregated, and then one SEM was added. Cronbach's alpha was 0.73 and 0.74. SEM was 2.8 and 2.81 over the administrations. Various practical details are discussed.
De Champlain AF (2004) Ensuring that the competent are truly competent: an overview of common methods and procedures used to set standards on high-stakes examinations. *Journal of Veterinary Medical Education*, 31:61-5	Excellent review of standard setting procedures.
De Champlain AF. (2010) Setting and maintaining standards in multiple-choice examinations: guide supplement 37.2 - viewpoint. *Medical Teacher,* 32(5): 436-437.	Review of the BEME guide on standard setting.
Downing SM*,* Tekian A*,* Yudkowsky R. (2006) Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine;* 18*:* 50–7	Compared 5 different standard setting methods (Angoff, Ebel, Hofstee, Borderline Group, and Contrasting Groups) for OSCEs. Noting that the different methods produced different cut scores, they commented that there is no "gold standard" in standard setting, and the best that can be obtained are defensible methods and outcomes. Different standard-setting methods produce different passing scores; there is no "gold standard." The key to defensible standards lies in the choice of credible judges and in the use of a systematic approach to collecting their judgments. Ultimately, he commented, all standards are policy decisions.
Elder A, McManus IC, McAlpine L, Dacre J. (2011) What skills are tested in the new PACES examination? *Annals Academic Medicine Singapore*, 40: 119-125.	Practical description of PACES, including comment on compensation.
Fowell SL, Fewtrell R, et al. (2008) Estimating the minimum number of judges required for test-centred standard setting on written assessments. Do discussion and iteration have an influence? *Advances in Health Sciences Education: Theory and Practice,* 13(1): 11-24.	Minimum number of judges required estimated by generalizability theory, for an MCQ and SAQ, with and without discussion. With a set desired RMSE of 2%, 10 judges are required without discussion and 6 are required after discussion, for both methods.
George S, Haque MS, Oyebode F. (2006). Standard setting: comparison of two methods. *BMC Medical Education,* 6: 46.	This paper compares norm referencing with modified Angoff standard settings for 4th year medical student MCQ based paper. The authors observed that 15% of the students failed with the norm referencing set at mean

	minus 1 SD, while 0% of students failed with Angoff. The test-retest reliability was 0.59-0.74.
Gruijter DNM. (1985) Compromise models for establishing examination standards. *J Educ Measure,* 22: 263-269.	The judges are asked to estimate the means and standard deviations of the pass rates and cut scores, and then give their estimates of the uncertainty of their estimates. As in Beuk's method, a graph was then drawn of scores versus pass rate, plot the estimated cut score and estimated pass rate, and call this point M. Then plot the relationship between cut scores and pass rates for the candidates as a decreasing curvilinear function. The uncertainty estimates are then used to draw the ellipse of all possible values around M, and where this ellipse touches the plot of student performance is the optimal compromise between cut score and pass rate.
Hambleton RK. (1978) On the use of cut off scores with criterion-referenced tests in instructional settings. *J Educ Measurement,* 15: 277-290.	Extended early description and defence of criterion referenced cut scores, including useful comments on the meaning of 'arbitrary'. Comments on the discrepancies observed in different standard setting methods. Some fine insults to other researchers, expressed with a generous strength, are included.
Hauer KE, Teherani A, et al. (2008) Approaches to medical student remediation after a comprehensive clinical skills examination. *Medical education,* 42(1): 104-112.	Useful description of qualitative findings on how educators view re-sitters with a view to remediation. Would inform further studies on IMG PLAB re-sitters.
Hays R, Gupta TS, Veitch J. (2008) The Practical value of the standard error of measurement in borderline pass/fail decisions. *Medical Education*, 42:810-815	Explore the value of the standard error of measurement in making decisions about students with exam scores at or below pass/fail borderline. New U/G course. Analysed de-identified pooled data for end of year assessments in the first 6 cohorts of medical students. Only scores around or below borderline mark, re-sit exams and pass/fail decisions. Cronbach's alpha calculated for each exam and used as measurement of reliability. Scores divided into 4 – pass score ±1 (68.26%), 1-2 SEM (95.44%), below the pass score 2-3 (97.44%) and >3 SEM (>97.44%) below the pass score. Standard Error of Measurement If scores fell within 1 SEM candidates likely to have deficiencies which could be remediated quickly, those who scored <3 SEM below the pass score were unable to remediate in time to pass a re-sit exam. Students in the middle 1-3 SEM below the pass score were unlikely to pass after remediation and re-sits. Scores in the pass score ±1 SEM band usually require a re-sit after remediation – this approach helps candidates progress but without significant weakness. Scores of 1-2 SEM below the pass score lead to firmer decisions about remediation and re-sits and scores of >2-3 SEM below pass scores indicate need for such broad remediation and re-assessment that candidates re-sit the year. Borderline candidates to undertake further assessments has positive implications for patient safety.
Hays RB. (2012) Remediation and re-assessment in undergraduate medical school exams.	Commentary on Pell at el (2012). Refers to difficulties in remediation, and suggests "The alternative may be more practical and more controversial: be tougher with students who clearly fail".
Hess B, Subhiya RG, Giordano C (2007) Convergence between cluster analysis and the Angoff method for setting minimum passing scores on credentialing exams. *Evaluation and the Health* Professions 30: 362-375	Compares cluster analysis with Angoff as applied to a primary care credentialing assessment. The results were similar. Using cluster analysis in this way (for instance, in association with Angoff) moves the process towards a compromise between prospective and retrospective methods.

Hess BJ, Weng W, et al. (2011). Setting a fair performance standard for physicians' quality of patient care. *Journal of General Internal Medicine,* 26(5): 467-473.	This paper demonstrates that carefully calibrated scores on an on-line 'practice improvement module' set by the American Board of Internal Medicine Maintenance of Certification programme, associated with an Angoff standard setting procedure, identified a small outlier group of physicians who had significantly lower clinical competence and professional behaviour ratings when residents, lower examination scores, and were more likely to work in solo practice. No sensitivity or specificity data is available from this study.
Hobma SO, Ram PM, et al. (2004). Setting a standard for performance assessment of doctor-patient communication in general practice. *Medical Education,* 38(12): 1244-1252.	Comparison of Angoff, Borderline regression and Borderline group methods for real video-recorded consultations by GP. Cut scores and RMSEs: Angoff 3.1 0.12 Borderline regression 2.5 0.05 Borderline groups 2.5 0.41 Regression was preferred by reason of consistency. Unsurprisingly, the GPs preferred regression to Angoff.
Humphrey-Munro S, MacFayden JC. (2002) Standard Setting: A comparison of case-author and modified borderline-group methods in small scale OSCE. *Academic Medicine*, 77(7): 729-732.	Canada, Ottawa 10 station OSCE. 61 4th yr medical students. 8 stations used in the study. Cut scores established from case-author and Modified Borderline Group methods. Case-author and Modified Borderline Group methods. Case-author cut scores = 5.77, fail rate 42.2% MBG cut scores = 5.31, fail rate 15.25% Conclusions: MBG score produced most realistic cut score and failure rates more reasonable than case-author method. MBG based judgements on all skills observed which produced better face validity, MBG costs more (examiners) but benefits out weigh these costs. Can also give immediate feedback to examinees.
Jalili M, Hejri SM, Norcini JJ. (2011) Comparison of 2 methods of standard setting: the performance of the three-level Angoff method. *Medical Education,* 45: 1199-1208.	Quant.105 med students in 14 station OSCE using 10 faculty members. Compared Angoff method (judges estimated whether Borderline students would pass station) compared with 3-level Angoff method (judges estimated whether a borderline examinee would perform the task correctly or not). Judges attended 2 half day workshops to help familiarise themselves with the methods. Angoff and three level Angoff mean scores 54.11 % sd 8.8 Angoff cut score 49.66 (95% CI 46.65 – 52.68) Angoff+ reality cut score 51.52% (95% CI 48.30 – 54.74) Angoff pass rate 65.7

	Angoff +reality pass rate 58.1 3L Angoff cut score 53.92% (95% CI 41.27-66.58) 3L Angoff+ reality cut score 63.09% (95% CI 51.80 -74.38), 3L Angoff +reality pass rate 44.8% and 12.4% p,0.001 Reliability using 95% confidence intervals of the cut off scores & inter-judge agreement. Differences between standards between OSCES were tested using a paired t-test; (p-values of,0.5 were considered to be statistical significance). Differences among pass rates arising from cut-off scores were tested using the McNemar test. Inter-judge reliability was calculated using 'intraclass correction coefficient. Correlation was assessed using the Pearson correction coefficient. A week after the comparison exam the procedure was repeated using normative data. Findings: Angoff showed higher agreement between judges & a narrower confidence interval in standards. Conclusion: Angoff more reliable and credible procedure for setting OSCE standards. Timely and costly to implement.
Jelovsek JE, Walters MD, Korn A et al (2010) Establishing cut off scores on assessments of surgical skills to determine surgical competence. Am J Onstet Gynaecol 203: e1-6.	Using Angoff, Hofstee and Contrsting Groups, applied to tests of surgical skill. More about the skill tests than the standard setting methods.
Jolly, B. (1999) Setting standards for tomorrow's doctors. *Medical Education*, 33(11): 792-793.	Commentary on Verhoeven et al 1999. Notes that on one occasion where he was involved, an expert group 'determined that a passing performance on the test as a whole should entail passing every station. In the pilot, all the examinees took all the stations. Not one passed'
Kaufman DM, Mann KV, Muijitjens, van der Vleuten. (2000) A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Academic Medicine*, 75(3): 267-271.	Canada – 12 station OSCE, 84 students to final year medical students x 2 (different years). 4 methods of standard setting were applied to the data and a cut off score for pass/fail decision to be made. Angoff, borderline, relative, holistic. Inconsistent results across the 4 types of standard setting. Angoff and Borderline similar scores (low failure rates under 2%) but Holistic (high failure rates 26%) and Relative (failure rates 8%) gave differing results. The Angoff method gave reliable results to be used in high stakes exams. Angoff method – pass score mean 52.00% before discussion, 51.17% after discussion. Variance among stations was 57.9% with 42.1% due to error variance. Root mean square error 1.45%. CI 95% Passing score 51.17% ± 2.90. The standard deviation of the actual exam scores were smaller 5.34%, mean score 63.21%. Second rating of the Angoff passing score 51.17%, resulted in a failure rate of 0.65% CI 95% Borderline method – mean value for the total test 52.46%, the average standard deviation was 9.74% Relative method – mean and standard deviation of score distribution 63.21% (1.96 SEMs below the mean).

	Failure rate 8.39%. The second relative method resulted in 95th percentile rank score of 72.27%, passing score 43.36% (.60 x 72.27%), failure rate of 0% Holistic method –(faculty wide standard used) passing score 60%, failure rate 26.45% Angoff and Borderline ,ethods provide defensible methods of setting standards. Borderline method cheaper to apply than the Angoff. However in high stake OSCES eg when licensing and recertification more investigation is needed.
Kiliminster S, Roberts T. (2004) Standard Setting for OSCE's: Trial of Borderline Approach. *Advances In Health Sciences Education,* 9: 201-209.	UK – 3rd and 5th year medical students. 22 station OSCE's held across 2 sites (n=210) students. Stations were marked out of 20 so total possible score was 440. The student had to pass 4 out of the 7 stations. Examiners briefed on borderline approach but given no training. Year 3 – Candidate's performance assessed at 19 of the 22 stations according to a 9-17 number checklist. In addition to the checklist assessment examiners were asked to mark candidate's overall performance for the 22 stations – pass, borderline pass, borderline fail, fail. Simulated patients made an overall rating (1 = low to 5 = high) for each candidate in response to questions at 14 stations. These were then multiplied by 4 to enable comparison with the other marks. Descriptive stats, correlations and Cronbach's alpha were used for analysis of data. Year 5 – An additional question in the checklist assessment asked examiners to give an overall mark honours, pass, borderline, fail. There were no simulated patient assessments. Borderline Results: Year 3 – Mean score 354 (out of 440) 81%, minimum 280, maximum 403. Reliability 0.66. For the 19 stations observed marks out of 20 were given – giving a total of 380. The mean was 307, minimum 236.33, maximum 349. Borderline students at ano-rectal examination = 18 (8.6%) at social history 94 (45%). Correlations between global ratings and checklist marks at each station ranged from 0.57 to 0.87. At the examination stations the correlations between the global ratings and the checklist marks ranged from 0.61 to 0.81 and to history stations the lowest correlation was 0.57, all other stations varied between 0.65 and 0.77. Pass mark 252.12. Overall mean = 307, standard error = 1.3057 Pass mark (total of the borderline means plus 1 SEM = 253.43 or 67%) Simulated patients only analysed history taking station – reliability 0.67 Overall marks for history taking was 0.65. Correlation between the simulated patients and the examiners was 0.55. Year 5 – Overall reliability 0.68 Mean score out of 140 was 98.3 (standard error = 1.3. Minimum was 66, maximum was 124. Total of borderline candidates mean 74.1. Pass mark 74.1 plus 1 SEM =75.4 (54%)

	Exam reliability was satisfactory but high stakes exams should see coefficients of 0.8. More training for examiners and assessment criteria to be improved. Reliability coefficient for simulated patients was higher than examiners and indicates simulated patients make credible examiners but also need training. Examiners used a smaller range of marks than simulated patients, showing reluctance on examiners part to mark students down therefore there is a need for training.
Kilminster S, Roberts T. (2004). Standard setting for OSCEs: trial of borderline approach. *Advances in Health Sciences Education: Theory and Practice*, 9(3): 201-209.	Review of Borderline Groups in practice (compared to set cut score of 65%). Reliability was 0.66-0.68. The SP scores showed a wider range and better reliability than that of the examiners, where comparable, ascribed to 'failure to fail'.
Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, Vlueten Cvd. (2003) Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education,* 37: 132-139.	Holland -16 station OSCE final year of PG GP training. 84 examiners. Reliability used generalisability theory. Credibility assessed by comparing pass rates of 86 trainees with 35 experienced GPs and by relating the passing scores to test difficulty. Modified Angoff. Borderline Regression (BR) used as an empirical procedure. Findings: Angoff pass score 73.4% reliability 2.1% Angoff +reality score 66.3% (n=16, P,0.001) reliability 2.1% BR 57.6% reliability 0.6% Angoff pass rates of students 19% and GPs 9% - correlation between test difficulty and passing score 0.69 Angoff +reality pass rates of students 66% and 46% correlation between test difficulty and passing score 0.88 BR pass rates of students 95% and 80% correlation between test difficulty and passing score 0.86 Conclusion: BR method more reliable and credible for OSCE. Reality check improves reliability of the modified Angoff method but does not improve reliability Modified Angoff without discussion and adjusted judgement, may lead to less precise and more variable estimates, resulting in a lower level of reliability than the Angoff method. Which is expensive timely and organisationally complex.
Mash B (2007) Assessing clinical skills – standard setting in the Objective Structured Clinical Exam (OSCE) *SA Fam Pract.* 49: 5-7	Comparison of Angoff and Borderline Regression in practice.
Nungester R J, Dillon GF, et al. (1991). Standard-setting plans for the NBME comprehensive Part I and Part II examinations.	Describes NBME approach in 1991, where test equating was used with an anchor test which had been standard set as the key. Standard setting appears to have been Angoff originally.

Academic medicine: journal of the Association of American Medical Colleges, 66(8): 429-433.	
Richter Lagha et al. (2012) A comparison of two standard setting approaches in high stakes clinical performance using generalizability theory. *Academic Medicine*, 87: 1-6.	Comparison of norm referencing and 'critical action' standard setting, where norm referencing performed better in terms of generalizability than CA setting in terms of dependability.
Ricketts C, Freeman AC, Lee R, Coombes. (2009) Standard setting for progress tests: combining external and internal standards. *Medical Education,* 43: 589-593.	Progress test sat by u/g students 4 times a year. 5 choice answer test from a question bank. Scoring: correct-item, minus 0.25 marks for incorrect and 0 for 'don't know'. Comparison of data from other schools made. Students in top 85% of cohort = satisfactory Progression decisions made from aggregating combined tests.
Rosebraugh C J, Speer AJ, et al. (1997). Setting standards and defining quality of performance in the validation of a standardized-patient examination format. *Academic Medicine,*_72(11): 1012-1014.	Examiner training and use of written scales improves reproducibility of results, and gives the subsequent OSCE greater reliability and correlation with other test forms.
Rothman AI, Cohen R, et al. (1991). Validity and reliability of a domain-referenced test of clinical competence for foreign medical graduates. *Academic Medicine,* 66(7): 423-425.	Explored candidate centred method (defining mastery group) in a context where there was some test-retest information. 24 passed, 47 failed, so numbers limited, but they conclude approach is valid and reliable. Rather old fashioned by today's standards.
Schindler N, Corcoran J, et al. (2007). Description and impact of using a standard-setting method for determining pass/fail scores in a surgery clerkship. *American Journal ofSurgery,* 193(2): 252-257.	Using Hofstee on a multi-assessment programme. Not strongly analytic.
Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Vleuten U. (2009) Who will pass the dental OSCE? Comparison of the Angoff and the Borderline regression standard setting methods. *Eur J Dent Educ*, 13: 162-171.	Aim: which standard setting method is best to prevent incompetent students to pass and competent ones to fail a dental OSCE. Applied to 119 3rd year dental students over 14 OSCE stations A comparison of the 3 methods – total compensatory (TC), a partial compensatory (PC within clusters of competence and a non-compensatory (NC). Reliability of pass/fail standard of error was calculated using root mean square error (RMSE). A criterion measure was taken using sample of students (89) who were rated by instructors in their clinics and divided into 'competent' and incompetent' students. The students clinical rating (true qualification) was compared with the pass/fail from the OSCE. Angoff 1, Angoff 2 (with reality check), Borderline Regression Angoff 1 TC pass rate: 86.6% RMSE: 1.3% PC: pass rate 30.3%, RMSE: 2.0-3.7%

	NC pass rate 0.8% Angoff 2 TC pass rate: 86.6% RMSE: 1.0% PC pass rate 34.5% RMSE:1.8%-2.2% NC pass rate 1.7% BR TC pass rate: 97.5% RMSE: 0.3% PC pass rate 61.3% RMSE: 0.6%-0.7% NC pass rate 7.6% BR method had highest number of incorrect decisions for the TC model than for the PC model Conclusions: BR method in a PC model provides defensible pass/fail standards. BR method showed more acceptable results and higher reliability than the two Angoff methods.
Schulz EM, Mitzel HC. (2011) A Mapmark method of standard setting as implemented for the National Assessment Governing Board. *Journal of Applied Measurement,* 12(2): 165-193.	A description of Mapmark standard setting in practice. This is a development of bookmark, which represents items by topics, and graphically, by difficulty. This enables judges to appreciate the level visually, allows topics to be differentiated, and can provide good feedback to candidates.
Shea JA, Bellini LM, et al. (2009) Setting standards for teaching evaluation data: an application of the contrasting groups method. *Teaching and Learning in Medicine,* 21(2): 82-86.	An analysis of Contrasting Groups methodology showing good and stable characteristics.
Southgate L, Campbell M, Cox J, Foulkes J, Jolly B, McRorie P Tombleson P. (2001) The General Medical Council's Performance Procedures: The development and implementation of tests of competence with examples from general practice. *Medical Education, Supplement***,** 35(1): 20-28.	This study compared the performance of 'reference' doctors (in good standing) with doctors referred to the GMCs procedures over concerns. Each group undertook a written knowledge test (EMQs), a simulated surgery test and an OSCE (drawn in part from PLAB). Standard setting was by Angoff, Contrasting groups, and the highest score of any failing candidate respectively. Extremely significant differences were obtained between the groups on all 3 tests, with high specificity and acceptable sensitivity. The correlation between the knowledge test and the OSCE was 0.69 ($p < 0.01$) and between knowledge and simulated surgery 0.72 ($p < 0.01$).
Southgate L, Hays RB, et al. (2001) Setting performance standards for medical practice: a theoretical framework. *Medical Education*, 35(5): 474-481.	Theoretical perspective – not relevant to report. Interesting analysis of different perspectives of stakeholders on professionalism.
Stern DT, Ben-David MF, et al. (2005) Ensuring global standards for medical graduates: a pilot study of international standard-setting. *Medical Teacher*, 27(3): 207-213.	Comparison of standards set internationally on IIME outcomes. Suggests that a true international standard might be settable.

Stone GE. (2001) Objective standard setting (or truth in advertising). *Journal of Applied Measurement,* 2(2): 187-201.	A thoroughly enjoyable polemic on what the author describes as Objective Standard Setting. Following the steps through, this involves selecting essential items, calculating item difficulty, having judges choose what percentage success should represent mastery, and calculating errors. The standard is a combination of these, for instance adding or subtracting the error depending on whether sensitivity or specificity is desired. The author suggests that this leads to greater consistency in fail rates. However, step 3 above will play a major component of this, and therefore 'objective' may be a little exaggerated.
Sturmberg JP, Hinchy JP. (2010) Borderline competence - From a complexity perspective: Conceptualization and implementation for certifying examinations. *Journal of Evaluation in Clinical Practice,* 16(4): 867-872.	Analysis of the meaning of 'competence', drawing on catastrophe theory.
Talente G, Haist SA, et al. (2003) A model for setting performance standards for standardized patient examinations. *Evaluation & the Health Professions,* 26(4): 427-446.	An example of modified Angoff in practice.
Taylor CA. (2011) Development of a modified Cohen method of standard setting. *Medical Teacher***,** 33(12): e678-682.	This study examines the consequenes of a modified version of Cohen standard setting, in which the guessing calculation is removed, the 90th percentile candidate is used, and the % multiplier is derived from Angoff standard set exams, rather than the arbitrary (or historical) 60% used in the original paper. The outcome showed lower variability than a fixed cut score, as would be expected from a norm-referenced approach, but did not explore the varying difficulty of the exam, or the validity of the outcomes.
Thanikachalam PM, Judson JP, et al. (2010) Inter variability in nedelsky and modified angoff standard setting in basic sciences of international medical university (IMU) undergraduate medical programme. *Histopathology,* 57: 192-192.	Poster. Angoff and Nedelsky gave acceptable results in a local trial.
Travis TA, Colliver JA, et al. (1996) Validity of a simple approach to scoring and standard setting for standardized-patient cases in an examination of clinical competence. *Academic Medicine,* 71(1): S84-S86.	Comparison of actual scores by SPs with 'case-author' ideal scores in a particular setting, showing little correlation.
Verhoeven BH, Verwijnen GM, et al. (2002) Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. *Medical Education,* 36(9): 860-867.	A comparison of standards set in a Progress Test by item writers compared to recent graduates, using Angoff approaches. The item writers appeared less consistent then the graduates, and the fail rate for writers was 57%, while that of graduates was 7%. While the graduates appear to have performed better, the authors question whether this would be viewed as politically acceptable. See Verhoeven et all 1999
Wayne DB, Cohen E, et al. (2008) The impact of judge selection on standard setting for a patient survey of physician communication skills. *Academic Medicine: journal of the Association of American Medical Colleges,* 83(10 Suppl): S17-	Angoff and Hofstee methods were used to set cut scores for a patient survey known as the Communications Assessment Tool (CAT). In general, patients and communications experts were much less lenient than trainees or programme directors. For patients, Hofstee methods were much more stringent than Angoff, failing 47% or 33% over two groups, compared to 7% for each group by Angoff.

20.	
Wayne DB, Barsuk JH, et al. (2007) Do baseline data influence standard setting for a clinical skills examination? *Academic Medicine,* 82(10 Suppl): S105-108.	Compared Angoff and Hofstee approaches in a clinical skills exam. Angoff gave pass rate of 89% to 96%, while Hofstee gave values of 68% to 86%, which the authors describe as 'too lenient' and 'too strict' respectively, but without true validity data. They recommend averaging the two cut scores.
Wayne DB, Fudala MJ, et al. (2005) Comparison of two standard-setting methods for advanced cardiac life support training. *Academic Medicine: journal of the Association of American Medical Colleges,* 80(10 Suppl): S63-66.	Comparison of Hofstee and Angoff methods used on advanced cardiac life support procedures. Both were reliable and stable. Hofstee was more stringent than Angoff but both were more stringent than historical approaches.
Wendt A, Kenny L. (2007) Setting the passing standard for the National Council Licensure Examination for Registered Nurses. *Nurse Educator*, 32: 104-108.	Describes standard setting for the NCLEX-RN exam, modified Angoff, followed by use of Beuk compromise method. In this paper, the difficulty was then further increased by 0.070 logits since "according to evidence from multiple and independent data sources and expert judgement, the passing standard needed to be increased". Note that this represents an attempt at 'retrospective validation' as mentioned in the main report, but on an informal basis.
Whelan GP. (1999) Educational commission for foreign medical graduates: Clinical skills assessment prototype. *Medical teacher,* 21: 156-160.	Historical account of the introduction of an OSCE by the Educational Commission for Foreign Medical Graduates.
Wilkinson TJ, Newble DI, Frampton CM. (2001) Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Medical Education,* 35: 1043-1049.	U/G med school in New Zealand. 18 station OSCE's run in 3 separate schools (looked at 3 years of data) but content and marking same. Study aims describe the method used to determine an OSCE pass mark and to evaluate its reliability and validity. 2 examiners per station. Performance of the stations marked out of 20 and using defined criteria and checklists. In addition a global rating of performance at each station: 4 point scale (fail, borderline, pass, above expected standard) – this score to be made independently of the mark out of 20 and from other examiners. These scores averaged out from the station mark for that student. If a student is rated as borderline the number out of 20 is noted. Giving a borderline mark for each station. They are then averaged for each station and school and then aggregated to give marks out of 360.The resultant % gives a score of a hypothetical student who was borderline at every station. The aggregate score is used to define pass/fail cut-off point. Reliability through interschool agreement, inter examiner agreement and intersection variation. Validity through comparing performance on the OSCE with performance on the MCQ and the aggregate mark from students' in-course assessment. A comparison was made with assessment marks of students who failed the OSCE with the lowest quintile of students who passed to see if examiners using norm-referencing. Comparison between means used Mann-Whitney *U* test. Correlations used Pearson's correlation coefficient. Yr 1: 181, Yr 2: 188, Yr 3: 203 students. Mean scores Yr 1: 95% CI 70.7% (69.7-71.7%) Yr 2:70.3% (69.4-71.2% Yr 3: 66.6% (65.8-67.4%). Variation considerable in number of students scoring borderline at a particular station. Variation indicates examiners basing decisions on a standard rather than norm-referencing.

	The average mark out of 20 for students judged to be borderline at a particular station varied between schools. Using a few stations to determine borderline cut-off score could produce unreliable results Method shows construct validity – students who fell below the pass mark also performed significantly worse than on other assessments. Validity: examiners not using norm referencing. Aggregate scores of 18 stations achieve greater internal consistency.
Wilkinson TJ. (2002) Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score - Reply. *Medical Education,* 36: 390-390.	A description of the practical use of Borderline group methods.
Wood TJ, Humphrey-Murto SM, Norman GR. (2006) Standard Setting in a small scale OSCE: A Comparison of the Modified Borderline-Group Method and Borderline Regression Method. *Advances in Health Sciences Education,* 11: 115-122.	Compare Modified Borderline Method with Borderline Regression Method. 10 station OSCE 59 clinical clerks. 8 stations involved patients and 2 stations were written questions. Only patient stations used in this study. Modified Borderline Method, Borderline Regression Method (6 point scale – inferior, poor, borderline satisfactory, good, excellent) 95% CI Comparison across standard setting methods Borderline Regression Method – cut score = 0.14 points lower (Mean square =5.14 vs. Mean Square =5.28) this was lower than the cut off score using Modified Borderline Group Method on 6 of the 8 stations. The pass rate for the regression method was 4% higher (M=71% vs M=67% respectively). The two approaches differed on 5 of the 8 stations. The 95% CI were smaller for the cut scores from the Borderline Regression method than the Modified Borderline approach by 0.09 (M=0.39 vs. M=0.48 respectively). t =2.93, p<0.05 Conclusion: Cut scores from Borderline Regression Method more accurate than from using Modified Borderline Group Method. Good to use for small scale OSCEs – doesn't require any complex procedure, cut off score easy to calculate, based on examinees actual performance so appears to have more face validity. Reduced statistical error with regression method.