The marking and standard setting of PLAB Part 2

Note: This report was submitted to the PLAB Working Party in April 2013, and does not take any account of information, data or events taking place after that. It is put in the public domain for information, and should be read as a historic document in the context of what was known at the time of its writing. It has not been edited since submission to the Working Party.

Chris McManus
August 2014

Final version, 29th April 2013

The marking and standard setting of PLAB Part 2

Chris McManus
April 2013

Summary

1. The *current marking scheme* for PLAB Part 2 replaced the *original marking scheme* in May 2008 and resulted in a fall in the pass rate of about nine percentage points.
2. The original marking scheme was simple and transparent, requiring candidates to pass 10/14 stations with no more than one station as a serious fail.
3. The current marking scheme is more complex and far from transparent, using two marking schemes (A to E: Excellent Good, Adequate, Fail and Serious Fail) and EOJs (Examiner's overall judgements; Pass, Borderline, Fail), and candidates having to attain at least nine station passes and an overall pass mark.
4. The setting of the station pass marks and the overall pass marks currently utilises a variant of Borderline Groups which adds complexity.
5. Current and original marking schemes involve weighting of objectives, without examiners being informed about weighting schemes, resulting in a loss of transparency.
6. The use of an overall pass mark plus one Standard Error of Measurement (SEM) seems to have been imported into Part 2 from the Canadian MCCQE Part 2 (on which Part 2 was modelled.
7. The reliability of PLAB Part 2 is, at a best estimate, about .71, with several measures having values as low as .55.
8. Multi-facet Rasch modelling using FACETS suggests that there are undoubted differences between stations/objectives in difficulty, and between examiners in stringency (hawkishness).
9. Differences in examiner hawkishness are currently monitored and occasional examiners who are outliers are advised that they are at an extreme.
10. Differences in difficulty of stations/objectives are currently handled by the Borderline Groups method,
11. Overall there is an argument that the current marking scheme has sufficient problems to mean it requires revision. Options for change include: a) returning to the original marking scheme, which had the advantage of simplicity and transparency; or b) using FACETS to take station/objective difficulty and examiner stringency into account statistically, in effect statistically equating across different carousels and examiners.
12. There is apossibility that examiners are somewhat disengaged, not knowing weights or pass marks, not knowing the marks of their fellow examiners, and not knowing candidates' eventual outcomes. All of that could be helped by a post-circuit feedback and discussion session, which would help examiners to internalise the implicit standards of their colleagues, and might improve reliability.
13. Factor analysis suggests two underlying components in station marks, one for Communication and History-taking stations (C+H), and the other for Examination and Practical stations (E+P).
14. C+H scores correlate with IELTS measures of productive language use, particularly Speaking, whereas E+P scores correlate with Part 1 knowledge scores and receptive language skills, particularly Reading.
15. There is an argument that there should be separate pass marks for C+H and E+P skills.
16. Currently there are only 8 C+H and 6 E+P stations, so that reliability on C+H and E+P sub-skills may be low. There is an argument for extending the length of the examination.
17. The current analyses are only 'internal' analyses of PLAB. Future analyses of the relationship to performance on MRCP(UK) and MRCGP examinations should provide extra information on the predictive validity of the PLAB assessments.

The marking and standard setting of PLAB Part 2

Chris McManus
April 2013

Introduction

At the 29th January 2013 meeting of the PLAB Review group, several related questions were raised about the marking and the standard setting of the PLAB Part 2 examination. The questions revolved around current processes and their justifications, and the extent to which empirical analyses of examination data might resolve them. At the meeting I was asked by the Chair to look at these and related issues. In so doing I have received extensive help from Katharine Lang, William Curnow[1], and Michael Harriman, who clarified both current and previous procedures, and Daniel Smith has been of particular assistance, putting extensive amounts of effort into providing data for the analyses which follow. I am very grateful to them all for their help. Finally, I must apologise for a very long report on what at first sight appears to be two rather small and unimportant questions, although unpacking them revealed much more than originally expected. The whole project was done in rather a short time and, with apologies to the reader, I can only quote Pascal, who once said in his *Provincial Letters*, "I've made this longer than usual as I didn't have the time to make it shorter" [2]. It should also be emphasised that the present report is for informal and internal discussion, and represents my perceptions of the PLAB Part 2 assessment after reviewing the large datasets provided to me. The conclusions were presented orally at a meeting of the Review Group on 16th April, and minor changes made to the report after that.

The issues

These can be briefly summarised here:

1. **What is the method by which a pass mark is set in the Part 2 examination?** This issue arose from item 11 on the 29th Jan 2013 Agenda. Although driven by the issue of the role of the one SEM (standard error of measurement) which is added to the overall pass mark, this question needs also to consider the simultaneous use of both A to E and Pass/Borderline/Fail marking schemes, the use of the borderline groups method, and the role of weighting of stations. All of these all have relevance to issues of transparency of the marking scheme for both examiners and candidates.
2. **What is the underlying structure of the marks awarded to candidates?** This issue arose because of item 13 on the agenda, *Assessing competence in the PLAB Part 2*, which compared the current method of marking PLAB Part 2 with the 'horizontal' approach of the MRCP(UK) PACES examination which awards separate marks for different skills/competencies, with a candidate having to pass all components in their own right. The key question is whether all of the stations in PLAB Part 2 are measuring the same underlying competency, and hence it makes sense for them to be added together, with a pass in any one compensating for a failure in any other, or whether there are perhaps two or more separate underlying competencies, for which it might make more sense to have two or more separate pass marks.

A note on the structure of the examination and the data.

I have been provided with data on examinations taken between 13th June 2001 and 25th April 2012. It is necessary to differentiate *doctors*, who are single individuals, from the several *attempts* at the examination at each of which a doctor is a *candidate,* some doctors being candidates on several attempts. For some statistical analyses the appropriate level of analysis is the attempt (candidate) and for others it is the doctor.

Of 41,170 attempts at the exam, the majority were first attempts (79.5%), but 16.8% were second attempts, 3.0% third attempts, 0.6% fourth attempts, with some having up to nine attempts[3]. Pass rates were 76.2% at

[1] I'm particularly grateful to William for picking up an important conceptual error of mine on an earlier draft.
[2] "Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte."
[3] It should be remembered that some candidates are still actively making attempts at the examination.

the first attempt, 78.2% at second attempt, 72.6% at third attempt, and 60.6% at fourth attempt. The highest attempt when a pass was achieved was the seventh.

The examination changed its structure on 1st May 2008, from what I call the *original marking scheme* to the *current marking scheme.* From a candidate's point of view the exam structure remained the same, there being 15 OSCE stations, at each of which they were assessed on a Communication Skill, Examination Skill, History-taking Skill or a Practical Skill by a different examiner, so that each candidate is assessed by fifteen different examiners. The examiners also used exactly the same response form with the original and current marking schemes.

The nomenclature for the various components of an examination is somewhat confused[4]. At the examination centre, testing takes place over a consecutive group of several days, which I will refer to as a *diet*. On each separate *day* all of the candidates see the same set of 14 OSCE stations and this set of stations I will refer to as a '*carousel*'. The carousel accommodates sixteen candidates, as there is a pilot (non-scoring) station and a rest station, with no candidate starting at the rest station (and therefore each candidate has fifteen stations, fourteen of which are scoring. The same carousel is run three times in a *day*, twice in the morning and once in the afternoon, in three separate *circuits*. The carousel changes for each of the different days of a diet.

Numbers taking the examination have fluctuated for many reasons, often to do with governmental policy and other factors outside the control of the GMC which affect the numbers of IMGs (international medical graduates) wishing to work in the UK. For convenience, I have divided dates in each year into tertials (four-month periods, January to April, May to August, September to December)[5]. Figure 1[6] shows the numbers of candidates and the pass rate in each tertial, separately for the original marking scheme and the current marking scheme. It is clear that the pass rate fell with the introduction of the current marking scheme, and has remained lower at about 69.2% compared with a previous value of 77.9%, an 8.7% percentage point fall, and an 11.4% fall in the proportion of candidates passing. Data were available for 33,678 candidates who attempted 471,492 stations under the original marking scheme and 7,492 candidates who attempted 104,888 stations under the current marking scheme.

What is the method by which a pass mark is set in the Part 2 examination?

The original and current marking schemes.

Although seemingly only of historical interest, understanding the much simpler original marking scheme helps in understanding the apparent complexities of the current marking scheme. An example of the mark sheet used by examiners, taken from the Examiner Handbook, is shown in figure 2. In particular notice that the letters A to E are above each column of the lozenges where judgements are made on the objectives, and at the top it states that "A= Excellent, B = Good, C = Adequate, D = Fail, E = Severe Fail"[7]. At the bottom there is also

[4] In particular the computer-system has a variable called TESTID, which very confusingly refers to a set of consecutive days on which exams were run, even though the OSCE stations in these exams are different. I call it a diet.

[5] This happens to be the scheme by which MRCP(UK) PACES is currently reviewed, and it has the advantages of corresponding broadly to the three terms in universities, and also neatly dividing 2008 into the first four months, which the original marking scheme was used, and the later months when the current marking scheme was used.

[6] Figures are shown at the end of the text.

[7] According to the 2003 Review, the A to E grading system was, "derived from the grading system for the previous PLAB test and has remained in place. We have concluded that there is no merit in retaining a system that involves converting grades to marks and back to grades again and recommend, therefore, that the objectives for each station should be marked on a numbered scale." (Para 29). Removing the grades, as was recommended in para 30, has not actually occurred.

what is known as the *Examiner's Overall Judgement* (EOJ), which is described as "Overall Judgement (P = Pass, B = Borderline, F = Fail)"[8].

Original marking scheme, 2001 to 30th April 2008. The *Skill* on each station was marked on from 3 to 7 Objectives, with each objective given a *Score* on the 5-point scale, from A to E, these being scored as A=4, B=3, C=2, D=1 and E=0[9]. Each objective had a pre-determined weight which was set by the Part 2 Panel and was not known to examiners[10]. The set of weighted scores for the objectives were combined to give a *Station Score*, which was a (non-integer) number, which was then expressed as a *Station Grade*, expressed from A to E. (numerically as 4 to 0)[11]. **To pass the examination a candidate had to receive at least 10 A, B or C grades, and no more than 1 E grade**.

- *Setting a pass mark*. The pass mark under the original marking scheme uses what for want of any better term in the literature I call 'implicit criterion referencing'[12]. Each examiner has their own internal criterion of how a satisfactory or an unsatisfactory candidate would perform. For a particular station, examiners make a judgement for each of the Objectives, of whether performance is Excellent, Good, Adequate, Fail or Severe Fail. Those judgements take directly into account an examiner's perceived difficulty of the station (and stations can vary in difficulty intrinsically, because they are assessing different tasks or skills, and because there may be different performances by different simulated patients with different candidates). Perhaps the single most important thing about the original marking scheme is that there is a direct link between an examiner's implicit model of what they perceive as Adequate, and the judgements that that examiner makes about a candidate's performance. 10/14 marks of Adequate or better result in a pass (as long as there is no more than one Bad Fail grade). **The marking scheme is therefore transparent to the examiners and also to the candidates,** apart from the weighting scheme**.**

Current marking scheme, 1st May 2008 to present. Each station/skill has from three to six different objectives, with each assessed by examiners using the standard five-point scale (scored A to E). Each objective has pre-determined weightings which are not known to the examiner. Marking of the exam then proceeds in several steps:

- *Passing a station*. The scores on each objective are weighted together to give a (non-integer) S*tation Total,* which ranges from 0 to 4. Each station has a *Station Pass Mark*, typically from 1.6 to 2.1 (95% range), and if the Station Total is greater than or equal to the station pass mark, then that station is passed. **One of the criteria for passing the exam as a whole is that the candidate passes at least 9 stations**.
- *The Total Station Score*. The 14 (non-integer) totals are summed to give a (non-integer) total station score, which is in the range 0 to 56[13]. In order to pass the examination the total station score must be

[8] The EOJ has apparently been on the Examiners Response Form since the inception of the exam. However data are only available since Jan 2007, which nevertheless means that data are available for EOJs and the original marking scheme for the sixteen months from Jan 2007 to April 2008.
[9] In the past A to E were scored as 5 to 1, but for consistency they are now being reported as being scored from 4 to 0.
[10] Although examiners were not told the weights, they did of course know that if they scored all of the objectives as A to C then the candidate would pass the station, and if they scored them all as D or E then the candidate would fail the station.
[11] Station grades were given on the basis of the Station Total, with 0 to 0.5=E, 0.55 to 1.5=D, 1.55 to 2.5=C, 2.55 to 3.5 = B, and 3.55 to 4 = A.
[12] This incidentally is still how the standard is set for the nPACES examination of the MRCP(UK).
[13] The 2003 Review in fact recommended, "that candidates be given a total score for the OSCE: As now, each objective would be worth a certain percentage of the total mark for the station and the numerical mark given for each objective would be multiplied by the percentage allocated. Rather than adding up the results of these calculations station by station, they would be added up across all the stations to produce a total score for the whole examination. Research has shown that giving candidates a total score gives examinations more validity…" (para 30). The 2003 Review did not intend

greater than or equal to the *Total Station Pass Mark* plus one *Standard Error of Measurement (SEM)*. The total pass mark for the fourteen stations consists of the simple sum of the fourteen station pass marks, and typically the total station pass mark has a value of about 30. Calculation of the SEM is far less straightforward.

- *The Standard Error of Measurement*. The SEM is a measure of the accuracy of an overall score, and its calculation depends on a knowledge of the *standard deviation* of the marks of other candidates, and a measure of the *reliability* of the exam. The SEM is calculated separately for all of the candidates taking the three circuits using a particular carousel on a single day, which is usually about 47 candidates[14]. Each of the 47 or so candidates has a total score, from which the standard deviation (SD) can be calculated (and is typically about 4.3). From the marks on individual stations it is also possible to calculate the measure of reliability which is *Cronbach's Alpha,* and that typically is about .67. The SEM can then calculated using the conventional formula of SEM = SD.$\sqrt{}$(1-r), where *r* is the reliability[15]. The SEM typically has a value of about 2.4. The *unadjusted pass mark* (or notional pass mark) is the total station pass mark (see above). From that the *SEM-adjusted pass mark* is calculated by adding 1 SEM to the unadjusted total station pass mark. As an example, if the total station pass mark is 30, and the SEM is 2.4, then a candidate must have a total station score of at least 32.4 to pass the examination. Because a candidate can only pass if they are 1 SEM above the unadjusted pass mark, and the SEM depends in part on the SD of the other candidates, the current marking scheme in effect says that whether or not a candidate passes depends on how well other candidates perform in the examination (and that is the usual criticism of norm-referenced marking)[16].
- *The final pass-fail decision*. In order to pass the exam a candidate must:
    - **Have passed at least 9 stations, <u>and</u>**
    - **Have attained a total station mark greater than or equal to the SEM-adjusted pass mark.**
- *The setting of the pass marks*. The description of the marking scheme assumes that a pass mark is known for each station. Those pass marks themselves must however be calculated. The method used is a variant of the borderline group method. Each examiner, as well as scoring a candidate on the various objectives/criteria, also makes an Examiner's Overall Judgement (EOJ), rating the candidate as Pass / Borderline / Fail. It is important to note that the EOJs do not contribute to the outcome for the particular candidate who is being examined, but contribute later to a revision of the pass mark for that station. In a conventional borderline group analysis the mean scores of all the candidates rated as borderline are averaged and that average used as the pass mark for a station (or the pass marks can be summed across stations to give a pass mark for the OSCE as a whole). The 2003 Review was worried that with only about 48 candidates per day, the number of borderlines on a station was probably too low to be reliable[17]. The solution to that problem was to calculate pass marks "off line". Pass marks for a station in an actual exam are based on previous outings of the station, and all of the borderline judgements that were previously made for that station. After the exam is finished, any new

---

that a count of total stations passed should also be used in determining pass or fail, and that scores on each station would only be used for feedback to candidates (para 31).

[14] These calculations are not carried out in-house, but the datasheets are sent out to be scanned, the data are processed, and are returned to the GMC along with information on the SEM and the SEM-adjusted pass mark.

[15] The SEM can be calculated directly, without the indirect route though SD and reliability, but most statisticians calculate it this way as SD and reliability are easily found using modern software packages.

[16] There is a sense in which the SEM does not depend on the SD of the other candidates, but it still does depend on how other candidates perform. An examination taken by a single candidate cannot have an SEM.

[17] "The weakness is that only 48 candidates take each OSCE. This may cause difficulties in using this method because the distributions may not form regular curves. Because of this we recommend that some extensive modelling be done before it is introduced. We remain certain that standard setting is essential but if this method is not successful, it is possible to use other methods." 2003 Review.

borderlines are added into the cumulative set of borderline marks on a particular station, and a new pass mark calculated for that station.

Discussion of the method of setting a pass mark.

As can be seen from the description above, the process of setting the pass mark is complex and might indeed be described as convoluted. If members of the working party find the marking scheme complex and less than transparent, then probably others do as well, including examiners. Although there is a natural evolutionary progression which can be discerned in the current, apparently Byzantine formulation, the merits and de-merits of the current marking scheme require discussion, particularly in the light of the issues raised at the beginning of this document.

The practical implementation of the Borderline Group method.

The Borderline Group method requires a lot of work for GMC staff, pass marks for each of the stations being recalculated after each carousel of 14 OSCEs has been run. Although the method appears elegant in some ways, current examiners seeming to be able to reset the pass mark on the basis of the behaviour they are seeing in the most recent candidates, the reality is rather different. Figure 3 shows the minimum and maximum pass marks for all of the 220 stations used since 2008, and allows two features to be illustrated:

How much does the pass mark actually vary within a station? The horizontal axis of Figure 3 shows the lowest pass mark ever set for a station and the vertical axis shows the highest pass mark ever set. In practice pass marks hardly change at all, despite the complex and continual updating based on those graded as borderline. The average difference between highest and lowest pass marks is 0.04 scale points (a scale point is the difference between, say, Adequate and Fail). The implication is that a simpler method for setting pass marks could probably be produced. The reason for the lack of change is that a typical station has been used a median of 355 times, of which on 10% or so of occasions are borderline mark will be given. The 48 candidates on a new day of testing may contribute 5 more borderline marks to that total, but those additional marks are unlikely to alter the passmark to any great extent.

How different are pass marks between stations? What seems to be much more striking about figure 3 is the variation *between* the stations, the pass mark varying from 1.5 (halfway between D:Fail and C:Adequate) and 2.5 (halfway between C:Adequate and B: Good). In fact the standard deviation is .11, with 50% of of the pass marks being in the quite tight range of 1.77 to 1.91. The average pass mark is at about 1.84. Under the original marking scheme the pass mark in effect was set at 1.55 (calculated as midway between 1 (Fail) and 2 (Adequate), with rounding downwards). Several points emerge:

1. Although one of the rationales for introducing borderline groups was that the difficulty of stations varied, **in practice most stations do not differ very much in their pass marks**, with a few stations being exceptions to that rule.
2. On average **stations have somewhat higher pass marks under the current marking scheme than they did under the original marking scheme.** Whereas previously a candidate scoring just over midway between C and D on a station (i.e. 1.55) would have passed that station, under the current system that candidate would probably fail, most pass marks being higher than 1.55. The average pass mark of 1.84 is now close to the average mark arising from Adequate (2). Since the stations would appear to have become harder, that in part probably explains why, as figure 1 shows, the pass rate *fell* with the introduction of the current marking scheme.

How do EOJs compare with the overall station marks under the original marking scheme? With the original marking scheme examiners marked each objective (and hence indirectly the entire station) on a five-point scale from A to E, with each grade having a specific label, including C as Adequate, and D as Fail. Confusingly those labels still remain the same, but instead the pass marks are mainly set through the overall

judgements of Pass, Borderline and Fail. *That at best is potentially confusing to examiners, who cannot predict the consequences of using those labels*. The sixteen month period from January 2007 to April 2008 allows a direct comparison of how candidates performed on individual stations using the original marking scheme for stations (A to E) and the current marking scheme (Pass / Borderline /Fail). Table 1 shows how examiners used the two systems on 24,460 stations.

Table 1*: Relationship between original marking scheme (A to E) and EOJ scores (Pass, Fail, Borderline)*

	EOJ: Fail	*EOJ: Borderline*	*EOJ: Pass*
E: Severe Fail	402	0	1
D: Fail	4492	602	49
C: Adequate	1733	3690	6706
B: Good	33	161	7026
A: Excellent	0	3	1562

There is not complete agreement, and it must be emphasised that some inconsistencies probably result from examiners using a different implicit weighting in their judgements compared with the actual weighting (and that probably accounts for the occasional candidate given an EOJ:Pass on a station where they had a E:Severe Fail). Overall though the pattern is of good agreement. Those rated as EOJ:Fail mostly receive D:Fail, with a few receiving C:Adequate. Likewise almost all of those receiving EOJ:Pass receive pass grades of C:Adequate or above. The EOJ:Borderline group is the most interesting as they are the ones who are being used to set the pass mark. The vast majority of EOJ:Borderlines receive C:Adequate, with a majority of the rest receiving a D:Fail.

What are the semantics of Adequate and Borderline? The two marking schemes depend heavily on the meanings of the terms Adequate and Borderline[18]. *Borderline* means that a candidate is precisely on the border, and the examiner cannot decide whether they should pass or fail; a slightly better performance and they would pass and a slightly worse performance and they would fail. As it were, they are sitting precisely on the fence between pass and fail, between satisfactory and unsatisfactory. *Adequate* though surely has a different meaning. As the OED suggests (see the footnote), it has connotations of sufficient, acceptable, satisfactory and good enough[19]. On that basis, Borderline is at a lower level than Adequate, and is presumably at the boundary between Adequate and Inadequate. Since the semantics of the Adequate and Borderline are different it is hardly surprising that pass marks based upon them are different, being lower for Borderline than for Adequate[20].

How in practice do examiners rate the performance of candidates passing or failing PLAB Part 2? Table 2 shows, for those who *passed* the Part 2, the number of EOJ passes and EOJ borderlines. A key (but not the sole) criterion for passing is that candidates should pass at least 9 stations, and if that applies to stations in general it should probably apply also to EOJs. The *shaded area* shows those who have passed 9 stations, along with those who have 8 passes and 2 borderlines, 7 passes and 4 borderlines, etc also counting as a pass. The numbers in bold indicate candidates who have passed overall without achieving the criterion of 9 passes. Noteworthy is that one candidate who passed with only 3 EOJ passes and 4 EOJ borderlines (and therefore had 7 EOJ fails).

[18] For *Borderline* the OED says, "The strip of land along the border between two countries or districts; a frontier-line; often fig., **the boundary between areas, classes**, etc.". For *Satisfactory*, the OED gives, 3a "Fully satisfying what is required; quite **sufficient**, suitable, or **acceptable** in quality or quantity" and 3b, "**Satisfactory**, but worthy of no stronger praise or recommendation; barely reaching an acceptable standard; **just good enough**."

[19] The MRCP(UK) nPACES exam now uses Satisfactory, Borderline and Unsatisfactory.

[20] Later it will be seen that the Canadian MCCEQ Part 2 exam, on much of which PLAB Part 2 was modelled actually used both Borderline Pass and Borderline Fail grades. Adequate is probably closer to being Borderline Pass (i.e. just good enough).

Table 2: Number of stations with pass and borderline EOJs for candidates who passed the exam overall on the current marking scheme.

Candidates who pass		Number of EOJ Borderlines										
		0	1	2	3	4	5	6	7	8	9	
Number of EOJ passes	0											
	1											
	2											
	3					**1**		**1**				2
	4					**4**	**1**			**1**	**1**	7
	5			**1**	**5**	**12**	**9**	**10**	**1**			38
	6	**1**	**2**	**7**	**34**	**46**	**30**	19	5			144
	7		**18**	**57**	**132**	124	58	26				415
	8	**7**	**88**	219	228	169	57	10				778
	9	40	186	362	302	143	26					1059
	10	86	270	394	212	62						1024
	11	113	322	317	107							859
	12	116	252	144								512
	13	124	141									265
	14	79										79
	Total	566	1279	1501	1020	561	181	66	6	1	1	5182

Table 3 shows a similar table but for those *failing the exam overall*. The shaded area again shows the area for which a candidate might be expected to pass based on EOJs. The candidates in bold have failed overall, despite seemingly having a reasonable of passes. One candidate had 11 EOJ passes, 0 borderlines, and hence only 3 fails, and yet failed. Similarly, 2 candidates had 6 EOJ passes, 7 borderlines, and hence only 1 fail, but still had failed overall.

Table 3: Number of stations with pass and borderline EOJs for candidates who failed the exam overall on the current marking scheme.

Candidates who fail		Number of EOJ Borderlines										
		0	1	2	3	4	5	6	7	8	9	
Number of EOJ passes	0			1	1	1	1					4
	1	1	3	8	10	4	3	4		1		34
	2	1	8	13	17	11	5	4	1			60
	3	2	22	33	25	37	26	20	1			166
	4	2	22	40	64	56	49	11	6	2		252
	5	7	41	90	108	99	49	21	3			418
	6	19	61	132	138	88	49	**8**	**2**			497
	7	20	79	134	133	**63**	**21**	**4**				454
	8	28	77	**110**	**62**	**20**	**6**					303
	9	**15**	**35**	**31**	**15**	**2**						98
	10	**6**	**9**	**4**	**3**							22
	11	**1**	**1**									2
	12											
	13											
	14											
	Total	566	1279	1501	1020	561	181	66	6	1	1	

Tables 2 and 3 suggest that pass-fail decisions under the current marking scheme are somewhat removed from transparency, candidates passing when many examiners have seem them as a fail, and failing when many examiners have said they should pass. That might be a reflection of the statistical unreliability of the EOJs (although it is probably not), but an important next step is to assess the reliability of station scores and EOJs.

What is the reliability of station scores and EOJs? Table 4 shows the reliability of various scores for the original and current marking schemes. On the weighted total station scores the current marking scheme is slightly superior, but the station grades are superior for the original marking scheme, no doubt because they are on a five-point rather than a two-point scale. The EOJs are of similar reliability in both schemes, and while they are slightly less reliable than weighted total station scores, they are less reliable for the station grades in the current marking scheme. In absolute terms none of the alpha coefficients are terribly high, and certainly do not meet the conventional criterion of 0.7 (but see later for a discussion of that criterion)[21].

Table 4: Reliability (Cronbach's coefficient alpha) for various measures

	Original Marking Scheme (N=33,677)	Current Marking Scheme (N=7,492)
Weighted total station scores	**.614**	**.646**
Station Grades	**.588** (A to E)	**.516** (Pass-Fail)
Examiner Overall Judgements	**.564** (N=1890)	**.577**

Taken overall, the problem found in the previous section, of candidates who are failing the exam with many EOJ passes, or passing with many EOJ fails, is not due to the lack of reliability of the EOJs. That neither station totals, station grades, nor EOJs are particularly reliable does however mean that correlations between them will not be particularly high, and occasional discrepancies of the type found in tables 2 and 3 are possible. Either way, some consideration does need giving to whether weighted total station scores or EOJs are the better measure.

Using FACETS to assess ability, station difficulty and hawk-dove effects in PLAB Part 2

A problem in any clinical examination is that some examiners might be more 'hawkish' or 'doveish' than other examiners (McManus, Thompson, & Mollon, 2006). PLAB Part 2 has always treated this seriously but has responded to it primarily by measuring overall scores in examiners and then giving them feedback if they are outside particular limits. However that analysis can take no account of the particular case-mix of candidates and stations received by an examiner. A better way of handling the problem of hawks and doves, and indeed of a host of other issues is to use multi-facet Rasch modelling, using the program FACETS[22].

Rasch modelling, named after Georg Rasch (1901-1980) is the underpinning of much modern psychometrics, including Item Response Theory (IRT). Rasch modelling, also known as one-parameter IRT, solves important problems in measurement theory, putting all measures onto a single, logistic scale, so that direct comparisons are possible, with the measures on a 'ratio scale'. Measurement theory is subtle, but everyday comparisons can be made. Temperature is conventionally measured on the Centigrade scale on which intervals are equivalent, so that the same amount of heat is needed to raise the temperature of a body from 10°C to 20°C

[21] The reliability of the Part 1 examination is reported as being above 0.9, and that seems a reasonable estimate given the length and type of the examination, and the wide range of candidate abilities.

[22] A very preliminary FACETS analysis of the PLAB Part 2 was presented to the Part 2 Board at its meeting on 20th Dec 2011 by Dr Suzanne Chamberlain. The analysis however used only the cut-down, student version of FACETS, and as a result analyses could only be carried out on a single diet with a single carousel. However, within a diet, examiner and station are utterly confounded, and hence no assessment of examiner hawk-dove status could be carried out.

as from 20°C to 30°C. However ratios of Centigrade temperatures are not meaningful, and there is no sense in which 20°C is twice as hot as 10°C. For that one needs absolute measurement scales, as occurs with degrees Kelvin, the absolute measure of temperature, where 20°K is indeed twice as hot as 10°K. Rasch modelling allows absolute measures within psychometrics, although the units are arbitrary (as are degrees Kelvin), and are expressed on a scale of logit units, which work on a scale which in effect is of odds ratios.

The FACETS yardstick. FACETS allows Rasch estimation simultaneously of the scale on which responses are made (A to E for immediate purposes), the candidate's ability, a station's difficulty, and an examiner's hawkishness, with estimates of these facets all being placed side-by-side in a single diagram known as a *yardstick*[23]. Figure 4 shows the yardstick for PLAB Part 2, based on 7,798 candidates taking PLAB Part 2 from 1st January 2007, examiner ID data being available from that date onwards. The outcome measure is the A to E scoring scheme for each station, with objectives weighted in the usual way and mean scores converted to the five-point scale in a standard way.

At the right-hand side are shown the five outcome grades, A to E, and these are placed on the standard 'Measure' (Measr) scale at the extreme left-hand side, which is in logit units. The boundaries of A with B, B with C, C with D, and D with E are at 2.85, 1.14, -.77 and -3.21, and hence are separated by 1.71 units, 1.91 units and 2.44 units. The gaps between the steps in the scale therefore become larger as one goes from high grades to low grades, meaning that the scale is not strictly equal interval at the level of the raw scales, the distance from the D/E boundary to the C/D boundary being greater than from the B/C boundary to the A/B boundary. Rasch modelling can however calculate the true underlying distances.

The second column from the left, marked Candidate, shows the ability of the individual candidates on the Measure scale, each "*" indicating 84 candidates, and each "." Indicating from 1 to 83 candidates. The measures for each candidate take into account differences in difficulty of stations and hawkishness of examiners and hence are *estimated true scores*. The candidate measures are on the same scale as the A to E judgements, and hence the average candidate is a 'high Adequate', with the best being on the Good/Excellent boundary overall , and the worst scoring Serious Fail overall. A good assessment spreads candidates out as much as possible, and hence should find as much variability as possible in the performance of candidates. The standard deviation (SD) of the candidate measures in figure 4 is 0.55. In contrast, good assessments have as little variance as possible due to differences between examiners and differences in difficulty of stations, both of which are regarded as *construct-irrelevant variance*.

The third column shows the estimated variability in scores of examiners, with each * indicating 3 examiners. Some examiners are more hawkish than others. The measures for examiners are a little counter-intuitive at first, but *the most hawkish examiner*, with a measure of 1.75, is at the top of the yardstick; they had only made 46 judgements (a single diet), but 76.1% of their scores were in the range A to C (i.e. passes). In contrast, the most lenient examiner, at the bottom of the yardstick, with a measure of -1.25 had made 572 judgements (12 diets) and 98.8% of their judgements were in the range A to C. The role of case-mix can be seen in the fact that although for the most hawkish examiner the average mark on the A to E (0 -4) scale was actually 2.1, if they had seen candidates of average ability on average difficulty stations then their average would have been even lower at 1.35 (whereas a similar calculation of "Fair-M" scores for the dove gave an actual average of 2.8 and a Fair-M average corrected for case-mix of 2.87, which is almost exactly the same). Interpreting the hawks and doves, with the hawks at the top of the yardstick, means that as one reads across, that only a few of the weakest candidates would fail if they were examined by the extreme dove, but many more of the best

[23] There is an arbitrary constant in such calculations and in a single analysis it is usually set so that the scale values have a mean of zero and a standard deviation of one. Examiner effects and station effects are also set so that they have means of zero. In repeated assessments it is possible to fix values so that they are compatible with previous values for the purposes of statistical equating or standard setting.

candidates would still pass if they were examined by the extreme hawk. The SD of the examiner measures is 0.43, less than that of candidates, but probably not as little as is desirable. Having said that, *the candidate measures take examiner leniency-stringency into account*.

The fourth column shows station difficulty, and as with examiner differences, difficult (hard) stations are at the top and easy stations at the bottom, a weak candidate being more likely to pass if they encounter easy stations than difficult stations. The SD of station difficulty is 0.45, so stations differ in difficulty by about as much as examiners differ in hawkishness.

FACETS also calculates various measures of reliability (Schumacker & Smith, 2007), the so-called "Separation reliability" (R_p) being 0.68 when examiner and station difficulties are taken into account. Without correcting for examiner and station effects, R_p is .63, which is broadly equivalent to a Cronbach's alpha of .595 calculated on the simple A to E grades for each station[24] (and see Table 4, above).

Rasch modelling can also be carried out using the EOJ judgements rather than the A to E judgements. R_p is then somewhat lower at .56, which probably in part reflects there only being 3 points on the scale rather than 5, and that is supported by converting the 5 point A to E scale into a 2 point pass-fail score (ABC vs DE), when R_p then goes down to .51. Reliability diminishes as information is lost by the use of scales with fewer scoring points.

The Rasch Person Separation Index and the number of Strata. Although R_p is similar to Cronbach's Alpha, which is helpful for comparative purposes, both have the problem that the scale ranges from 0 to 1, so that although the differences between .7 and .75 and .9 and .95 are numerically similar, they are very different in their interpretations, the latter actually being much more substantial than the former. Rasch modelling instead uses a measure called the *Person Separation Index* (G_p), which extends from 0 to infinity. From it can also be calculated the number of *Strata* implicit in the data. In effect the number of Strata considers the extent to which individuals in the population can be differentiated on the basis of their measures into different groups[25]. PLAB Part 2 with A to E scoring has only two strata, suggesting that while there is some reliability in the measures, the G_p of 1.47 is far from adequate.

Why is the reliability of PLAB Part 2 so low? Whether calculated as Cronbach's alpha or its Rasch equivalent, the reliability of PLAB Part 2 is low, particularly given the GMC/PMETB advice that, "in high stakes medical examinations, we look for reliability coefficients around 0.9, and would probably find any components with reliability below 0.8 to be inadequate" (General Medical Council, 2010; Postgraduate Medical Education and Training Board, 2009)[26]. Although there are problems with the advice offered by PMETB (McManus, 2012; Tighe, McManus, Dewhurst, Chis, & Mucklow, 2010), that doesn't get around the problem that the PMETB advice is apparently still current. It could be argued that the true reliability of Part 2 is higher than it seems

[24] The Rasch model is still taking into account the non-linearity of the A to E scale whereas Cronbach's alpha assumes (incorrectly) that A to E are at equal intervals to one another (as their 4 to 0 scores suggest).

[25] Individuals are assumed to be different if they are at least three standard errors of measurement apart. A reliability of .435 implies that there are at least two strata (although the individuals are probably at the extremes of the distribution and most individuals are not separated). Likewise a reliability of .73 is necessary for three separable strata, .85 for four strata, .91 for five, .94 for six, and .96 for seven.

[26] The advice contrasts quite dramatically with that in the *GMC Glossary for the regulation of medical education and training* (General Medical Council, 2013), which says "There is much debate about what constitutes a 'good enough' co-efficient. As a general rule, anything **under 0.5 would be normally viewed as suspect** (and further investigation would be required), **between 0.6 and 0.8 is good** and **between 0.8 and 0.9 indicates strong reliability**. **Over 0.9 would be suspiciously good** – in which case, it would be worth checking that the data are correct, and rerunning the analysis" (p.9; my emphases). To complicate the matter further, albeit in a context which is not high-stakes, two recent reviews have talked of many researchers treating 0.7 as a measure of the adequacy of a test, but then state that it, "is clearly an improper usage of the statistic" (Cortina, 1993) [p.101], and that "the use of any cutoff value (including .70) is shortsighted" (Schmitt, 1996) [p.351].

because it is a part of a wider assessment system (just as the reliability of MRCP(UK) Part 2 is lower than Part 1 because candidates have of necessity already passed Part 1 and hence have a lower variance (Tighe et al., 2010), but that is not necessarily the case for PLAB Part 2, the exam being a very different type of assessment from PLAB Part 1.

Comparison of PLAB Part 2 with other assessments.

The reliability of PLAB Part 2 is not particularly high. A useful exercise is to compare it with other equivalent assessments for which FACETS analyses have been carried out (see Table 5). These studies were found by searching in a fairly informal way for FACETS and assessment on Google Scholar, there not being time to carry out a complete literature search.

Table 5 summarises six sets of medical data using FACETS for assessing clinical/OSCE type assessments. The studies are ranked from lowest to highest reliability, and that of PLAB Part 2 is the lowest. Comparing and contrasting suggests several factors which do not explain the low reliability. Low numbers of items on rating scales can reduce reliability, but there are other studies with 5-point scales with higher reliability than PLAB Part 2. Fewer stations reduces reliability, but PLAB Part 2 has more stations than the other studies. Other studies vary in the extent of the variance due to stations and examiners, but there are studies with both higher and lower values than PLAB Part 2, and FACETS takes account of such factors anyway. When candidates are more similar in ability then reliability is inevitably lower (until eventually when all the candidates have the same true ability there can be no reliability in distinguishing them, and the only variance is measurement error). The candidate variance for PLAB Part 2 is lower than some other exams but notably is lower than the most reliable assessment, COMLEX[27]. Given the range of candidate ability found (there are candidates averaging nearly Excellent and others averaging a Serious Fail), and given the unselected international origins of the candidates, it seems unlikely that the effective ability range is substantially less than in, say, MRCP, which also draws on similar international populations.

The two major features differentiating PLAB from the other assessments is that for PLAB Part 2 there is only a single eventual judgement being made by each examiner on each candidate, and there is only a single examiner at a station. In contrast, MRCP(UK) has two examiners at each station, who in (old) PACES made only a single judgement, but in nPACES make up to seven judgements. Likewise the Dundee MMI has 3 ratings per station, and COMLEX has 6 attributes per station.

A FACETS analysis by Objectives within Stations. Overall the most likely explanation for the relatively low reliability of PLAB Part 2 is to be found in the analysis reducing the multiple attributes assessed at each station into a single A to E score, thereby losing information. If that is the case then an analysis at the level of Objectives within Stations, rather than total scores of Stations, should be substantially more reliable, all of the detailed information on candidates and their performance on particular objectives being properly taken into account.

[27] Candidate variability undoubtedly has an effect. I explored that by dividing the PLAB candidates into high and low scoring candidates. Reliability was lower in both groups, as expected, and Candidate variance was also lower. However RMSE remained constant, as did Station variance and Examiner variance, as also expected.

Table 5: Comparison of six sets of data in which FACETS has been used to analyse assessment data in a medical context. Studies are sorted in ascending order of separation reliability (R_p).

Assessment	Scales for ratings	Description	Reliability	Standard deviations of measures			Comments
				Candidate	Station	Examiner	
PLAB Part 2, Jan 2006 to Aug 2012 [Current study]	*A to E (Excellent / Good / Adequate / Fail /Severe Fail)*	*7,798 candidates; 14 stations with 1 medical examiner; 142 examiners*	***R_p = .68***	*.55*	*.45*	*.43*	*A to E results used as it has the highest reliability.*
Medical school 4th year OSCE, University of Notre Dame, Australia (Tor & Steketee, 2011) (n=80)	Not stated	80 candidates; 11 stations with 1 medical school examiner; (?)64 examiners.	**R_p = .70**	.52	.59	n/a	A number of technical details are missing
MRCP(UK) PACES – old format (McManus et al., 2006)	ClearPass/ Pass / Fail / Clear Fail	10,145 candidates; 7 stations with 2 medical examiners; 1259 examiners	**R_p = .82**	.75	.10	.26	Based on 9 diets. Results similar for 26 diets. Stations in 7 groups as real patients
University of Dundee: Multiple mini interview (MMI) for medical school admission (Till, Myford, & Dowell, 2013)	Extremely Poor/ Poor / Adequate / Good / Ideal	452 candidates; 10 stations with 1 examiner (staff, student or SP) and 3 ratings per station; 156 examiners.	**R_p = .89**	.81*	1.03*	.20*	SPs more hawkish; students and staff similar.
MRCP(UK) nPACES (new format) (Elder, McAlpine, Bateman, Dacre, Kopelman, & McManus, 2011); Unpublished analyses	Satisfactory / Borderline / Unsatisfactory	15,106 candidates; 7 stations with 2 medical examiners and 4 to 7 skills per station; 1838 examiners	**R_p = .91**	.58	.11	.31	Diets 2009/3 to 2012/2. Stations in 7 groups as real patients are used.
Assessment of Humanistic Skills in a national osteopathic exam (COMLEX) (Zhang & Roberts, 2012)	Unacceptable / [Unlabelled middle] / Superior, each divided into 3 (i.e. 9 points)	4,564 candidates; 11 stations with 1 SP rater and 6 attributes per station; SPs confounded with station (85 combinations)	**R_p = .93**	.49*	.24*+		Notes that 95% of candidates passed and therefore range of abilities probably restricted

* Not given directly in paper but estimated from range of values shown on the yardstick.
\+ Estimates of station and examiner confounded.

The FACETS analysis was repeated but this time making the unit of analysis the individual A to E scores on the objectives at each station, scoring these 4 to 0. For the diets since Jan 2007 there were 545,900 such judgements, with from 4 to 6 objectives on each station, and a typical candidate having altogether 70 marks on individual objectives. As before candidate, and examiner were also put into the model. The yardstick is shown in figure 5. Compared with figure 4 it is clear that objectives are far more variable than were stations (which is not surprising), some being so easy that only a serious fail candidate would not pass them and others being so hard that only excellent candidates would pass them. FACETS takes into account those differences in difficulty, along with differences in examiner stringency, to estimate a true ability mark for each candidate. The separation reliability, R_p , is now 0.90, with the result that the Person Separation Index (G_p) is 3.00, meaning that there are at least 4 separate strata.

As an example of how FACETS analyses individual objectives, table 6 shows the objectives on a single station, 2626, taken by a fairly typical number of candidates. The difficulty of each objective is shown, and these are also indicated in red alongside the yardstick for objectives in figure 5. The easiest objective is #5 (Approach to the patient), and the hardest is #3 (Further management), both of which make sense. FACETS takes into account the differing difficulty of these and the other objectives, and uses information from them to estimate the overall ability of the candidates.

Table 6: Station 2626 taken by 138 candidates. The station asked about the History and Management of Lower Abdominal Pain. The standard errors of the Difficulty Scores are .10 in each case.

Objective	Weight*	Description	Difficulty score[+]
1	25%	History of Complaint	-.11
2	25%	Probable diagnosis and appropriate management plan	.23
3	15%	Further management	.82
4	15%	Explaining/Advising	.10
5	10%	Approach to the patient	-.40
6	10%	Fluency of performance	-.11

* Included for information but not included in the analysis
[+] The difficulty score is on the standard Rasch logistic scale, with high numbers indicating harder items

Overall it is clear that there is much more information in the detailed analysis of objectives than is apparent from using a simple, weighted sum as in the current marking scheme, and hence the reliability of the examination using such a system has risen to 0.90, putting it amongst the better performers in table 5.

Split-half reliability and an honest estimate of reliability. Reliability, even when expressed as the conventional Cronbach's alpha, is a difficult and sometimes paradoxical concept (Cortina, 1993; Schmitt, 1996). However those items must be *statistically independent*. In the case of an MCQ examination, as Clauser and Linacre have pointed out (Clauser & Linacre, 1999), running a candidate's mark sheets through three different optical scanners would increase the apparent reliability, but would not be acceptable because the items would not now be independent, but merely the same ones repeated[28]. A similar consideration applies with clinical examinations. Asking ten different examiners independently to assess a candidate does provide ten independent pieces of

[28] If I ask the question, "Are you happy?" then I get an answer. Repeating the same question 20 times and getting a similar answer twenty times does not truly increase the reliability of the question/scale, as the questions are not asking about independent attributes.

information, but asking an examiner to rate the same candidate on ten similar questions/objectives probably does not provide ten independent pieces of information (although it may provide two or even three pieces of information depending on the underlying factorial structure of the responses)[29]. In PLAB Part 2 the examiners are making assessments on up to six objectives, but are those assessments independent? FACETS assumes that they are and as a result separation reliabilities can be inflated (and the same happens with Cronbach's alpha, often resulting in claims of unlikely reliabilities).

There is a sense in which the only honest assessment of reliability is *split-half reliability*, where the two sets of measures are specifically created so that they are truly independent statistically. That can be done for PLAB Part 2. There are 14 stations, and each station has different content and a different examiner. If only the odd-numbered stations are analysed and only the even-numbered, then the two scores are genuinely independent, and the correlation between them is the correlation necessary for calculating the split-half reliability. Doing that with FACETS, and accounting for differences in examiner stringency and difficulty of specific objectives, using the method described in the previous section, the correlation of candidate scores for the odd-numbered stations and the even-numbered stations is 0.550. Using the conventional Spearman-Brown formula, which takes account of the fact that each half of the test necessarily has only half as many items, then the reliability of the whole of PLAB Part 2 is **0.710**, which is the split-half reliability. That is a lot less than the 0.90 suggested earlier by FACETS, but is a little higher than the 0.68 suggested in table 5, mainly because the analysis is at the level of objectives rather than stations.

Not all examinations necessarily have a split-half reliability which is substantially below the separation reliability reported by FACETS. For nPACEs, the separation reliability reported in table 5 when both examiners make judgements is 0.91. Calculating independent marks for each candidate, based on a FACETs analysis of either just an arbitrarily assigned first examiner or second examiner produces scores which correlate 0.853, giving a split-half reliability of 0.921. Although the FACETS separation reliability for nPACES happens to be a good estimate of *honest reliability* , that need not necessarily be the case, as seems to be the situation with PLAB Part 2. The extent to which reliability is erroneously inflated in the other studies in Table 5 cannot be ascertained without access to the raw data.

Summary. The net conclusion of a lot of complicated statistics is the sad one that the reliability of PLAB Part 2 is probably lower than is desirable, with 0.71 being the best value, which takes account of differences in examiner stringency and the difficulty of different objectives. Some of the problems with reliability may reflect problems with what has been called 'backfilling'[30], and also perhaps with examiners being disengaged (see the next section).

---

[29] As an undergraduate I heard a story about Wittgenstein, which shows the paradox neatly (although sadly I am unable to find the story anywhere in books about Wittgenstein or on the internet). The story was that Wittgenstein was sitting in his room in Trinity when a colleague came in holding a copy of The Times and said that there had been an earthquake in Peru. "Good Heavens," said Wittgenstein, "that cannot possibly be true!". Wittgenstein then jumped up, ran out into Trinity Street, bought a second copy of that day's Times, and came back saying, "It must be true, it says it in this copy as well". Replications to be valid must be true, independent replications, and not mere repetition of earlier information. Of course further mere repetitions of this story about Wittgenstein will not make it more likely to be true.

[30] In backfilling, examiners have a group of objectives to mark as well as an overall grade. Typically they decide on the overall grade, and having done that allocate their individual objective marks around the overall grade (and being doctors they are not so stupid as to 'tick down the column', and so instead they make one or two

The social psychology of examining. Increasing the reliability is not straightforward, but I would be concerned that examiners might be too disengaged from the process in which they are taking part, being treated as "marking machines", rather than as skilled professionals involved in complex judgements of potential colleagues in collaboration with their peers[31]. Examiners make marks on the A to E scales and the EOJs but there is no direct link from those marks to whether or not the candidate passes or fails, the weighting schemes are secret, the pass marks set by Borderline Groups are secret, and the examiners neither find out whether or not their candidates pass or what their fellow examiners thought of the candidates. All of that could well contribute to the low reliability of PLAB Part 2. One possible solution would be to have a 'post-circuit' feedback session, where examiners see on a screen[32] the overall marks given by all of the examiners to all of the 16 candidates, along with a discussion of obvious discrepancies (Extreme Fail vs Excellent, or, on EOJ, Pass vs Fail). That would all help to produce an esprit de corps amongst examiners, and a sense of collective process with mutual responsibility for outcomes. At the very least there is an argument for a qualitative research study of examiners to assess their perceptions of the role of being an examiner, how it is seen, what they like and dislike about it, etc..

EOJs, weighted station totals and the philosophy of pass marks and Borderline Groups?

The analyses of PLAB Part 2 using FACETS provide a lot of insight into the examination, and factors likely to influence its dependability, particularly differences in examiner stringency, station difficulty and difficulty of the various objectives on stations. However a key issue for any examination is not merely the placing of candidates in order, but the decision about where a pass mark is to be set; *Standard Setting*. For the current marking scheme of PLAB Part 2 the pass mark is based on the method of Borderline Groups, and for the original marking scheme it was based on judgements across each station of whether performance was Adequate or less than Adequate.

Any reader who has got this far may well be wondering what is the deeper philosophy behind standard-setting methods such as Borderline Groups. The method was introduced at the behest of the 2003 Review, which commented, "We would like to recommend the introduction of a formal standard setting procedure for the OSCE and believe that the borderline group method would be the most appropriate … [It has] … been validated by 10 years experience at the Medical Council of Canada". The statement is worth inspecting.

That the Borderline Groups method is formal will surely be accepted by anyone who has read through the details of how the pass mark is set in the current marking scheme. However the extent to which it has been "validated" is more contentious. A recent review for the GMC (Cleland, Dowell,

---

objective grades above or below the overall grade). The overall result is objective marks which are nearly carbon copies of one another and of the overall grade. A notable feature of nPACES is that there is no overall grade, examiners being told that each grade matters, and that sufficient borderlines or fails on any one skill is sufficient to fail a candidate. That may in part explain why the separation reliability of nPACES is almost exactly equivalent to the split-half reliability.

[31] I was struck when visiting the examination how much examiners seemed to crave more information about what was happening, about how little they seemed to know of what was actually happening, and how much they wished to discuss candidates they had just seen (and even if the rules forbid simulated patients and examiners discussing candidates, that inevitably, and perhaps justifiably, was what was happening in the one minute gaps between candidates). The video monitoring of examination rooms, while it has many advantages, also has the untoward effect of suggesting that any problems with the examination are problems caused by examiners, with the Big Brother technology being there to prevent this.

[32] Whilst electronic marking technology would help with this, it is straightforward to carry out with an assistant typing paper records into a spreadsheet.

McLachlan, Nicholson, & Patterson, 2012) discussed different types of validity, and while the Canadian method may have 'faith validity' and 'face validity', there is little evidence which I can find that MCCQE Part 2 has predictive validity (unlike MCCQE Part 1 which does have predictive validity for clinical outcomes (Tamblyn, Dauphinee, Hanley, Norcini, Girard, Grand'Maison, & Brailovsky, 2002)). Even if MCCQE Part 2 had predictive validity it would still be unclear of the extent to which that reflected the use of the borderline groups method *per se*.

A particular problem for Borderline Groups (and Borderline Regression) is understanding the precise point at which the standard is set, as numbers are calculated from other numbers, and eventually a pass mark emerges. That lack of transparency gives the impression, at the least, of smoke and mirrors, rather than a clear vision of how and when a candidate can pass or fail. Making proper sense of quite what Borderline Groups is doing requires a minor diversion into the nature of standard setting and the concept of the pass mark.

A diversion on the nature of standard setting.

Setting standards is hard, particularly if they are to be "credible, defensible and acceptable passing scores for performance-type examinations in real-world settings" (Downing, Tekian, & Yudkowsky, 2006).

A commonplace amongst those setting standards is that "Standards are arbitrary but they should not be capricious". The arbitrariness is apparent in fact even with seemingly hard-nosed standards[33]. Consider the level of lead in drinking water or the mortality rates of surgeons. Although everyone would like no lead in drinking water and a mortality rate of zero, those counsels of perfection cannot work in the real world, for a host of reasons. For other standards, such as a normal level of blood haemoglobin or serum potassium there are more complications, as zero is clearly not the optimal level, and too high or too low values need to be flagged up. The result is that some level has to be set, and that level is inevitably arbitrary, reflecting various conflicting results. The level of lead in the water where I live, in London, is 2.1 $\mu g/L$, which is clearly not zero. The permitted level is 25 $\mu g/L$, and hence the water 'passes' the test. Similar considerations apply with mortality rates for surgeons, with the situation made more complex by the subtleties of case-mix, etc., and the overall conclusion being that some level above zero has to be set. Merely to pluck numbers from the air would be 'capricious', as also would it be for boards of examiners to alter pass marks on a whim to satisfy some local purpose.

The setting of pass marks can be done in several ways, many of which have problems. In *norm-referencing* a fixed percentage of candidates, say 90%, passes an examination. That is regarded as a poor way of setting standards as whether or not an individual passes an examination should not depend on who else happens to take the assessment. Arbitrary percentage marks (such as the 50% used in many schools) are also not acceptable, particularly with multiple-choice or clinical exams, where the difficulty of items and stations can vary from occasion to occasion. *Criterion-referencing* is regarded generally as the best method of setting a standard, and so, in a surgery exam, a candidate

[33] The setting of standards should not be confused with producing reliable marks. A marking scheme gives marks to individuals, and allows them to be ranked from best to worst. If those marks are reliable then the same candidates on different occasions with different items will give the same ordering. Standard-setting is not about putting candidates into order, but about the more difficult question of where the line should be drawn between pass and fail, acceptable and unacceptable. Reliability tells one little or nothing about the latter, except that an unreliable test will inevitably have an unreliable pass mark, where chance alone determines to a greater extent whether or not a candidate passes or fails.

may be required to show that they can tie a certain number of knots successfully in a laparoscopic simulator in a fixed time. The criterion is clear and in principle everyone can pass or everyone can fail. Criterion-referencing often does not have such clear outcomes, and the quality of a candidate taking a history is less easy to quantify well, although a failure to ask about certain key symptoms may be a criterion. The ability to empathise well with a distraught relative is harder still to make clear as a criterion.

Very many assessments, particularly at higher levels of performance, depend in the final analysis on expert judgement. A hospital doctor with a new trainee will have certain expectations about what the trainee should and should not be able to perform on their first day on the ward. Experienced hospital doctors have seen many trainees, they know their own domain of expertise well, and if they feel that a trainee is less competent than they should be then that judgement is like to be valid. The validity of the judgement ultimately is tested by the long-term behaviour and the consequences of good or poor clinical practice in the trainee. The key feature is that *the judgement of an expert observer is likely to be the best way of assessing the ability of others who are training in the field*, with the proviso that conditions are standardised sufficiently to remove the possibility of caprice in the judgements. Almost all clinical examinations ultimately depend on that fact.

Borderline groups (BG) BG is an 'examinee-centred' method, rather than the 'item-centred' methods such as those of Angoff and Edel which are used for setting standards on knowledge-exams, as such as PLAB Part 1, by assessing the extent to which a 'just-passing candidate' should know the specific material. BG is a way of solving a particular problem which arises in clinical examinations such as OSCEs. Scoring sheets and checklists are used to measure the behaviour of candidates across a range of skills, and a summed, total number can be calculated for a particular station, with better candidates having higher total scores. The problem though is where to draw the line dividing pass and fail. In BG therefore the examiners, as well as completing checklists, also provide a global rating typically of the form Fail, Borderline, Pass[34], and this is the EOJ in PLAB Part 2. BG sets the pass mark by taking the total scores of all the candidates rated borderline, and uses the average of those scores as the pass mark (so in effect, about a half of the Borderline candidates pass the station and about half fail). Although the current marking scheme for Part 2 looks complex, it is actually a modified form of BG, the main difference being that the stations are sufficiently standardised across so carousels that the Borderline judgements of examiners on previous outings of the station can contribute to setting a pass mark on the current outing.

A key philosophical question concerning PLAB Part 2 concerns where and when the pass mark is set and by whom. For the original marking scheme that was obvious - - a candidate passed if, of the 14 expert examiners, enough, defined as 10, thought that they were at least Adequate, and no more than one thought that they were an Extreme Fail. That is entirely transparent, and as long as the 14 examiners are indeed experts, have been properly trained, and their judgements appear consistent with other examiners, etc, then the marking scheme is credible, defensible and acceptable.

Things are less immediately obvious for the current marking scheme, but **ultimately in the current marking scheme all of the numbers go back to the judgements made by the examiners of Pass, Borderline or Fail**. That is the locus at which all of the numbers, with all of their decimal places and

[34] Some systems use Fail, Borderline Fail, Borderline Pass, and Pass, the two Borderlines indicating on which side of the divide the examiner feels the candidate should lie.

apparent precision, are grounded in the clinical reality of whether or not the candidate is likely to be a competent clinician of an acceptable standard.

If the key decisions are those made by each examiner as to whether or not a candidate is a Pass, Borderline or Fail, then it has to be asked **why the pass mark is not based directly on the examiners' judgements themselves, whose meaning is clear and transparent, as in the original marking scheme, rather than on a set of derived numbers whose interpretation is unclear even for those who work with the examination regularly**. If the meanings of the numbers were clear then presumably examiners would have little difficulty in answering, say: What is the interpretation of a mark of 29.3? What is the meaning of an adjusted pass mark of 28.7? and how do either differ from a subsequent exam where a candidate may still have a mark of 29.3 but the adjusted pass mark is now 30?. The numerical machinations may be clear but the meaning of the outcomes is not. In contrast, as tables 3 and 4 show, to say that of the 14 examiners, 8 said a candidate was a Pass, 4 said they were a Borderline, and 2 a Fail, gives an immediately transparent meaning as to why it is thought that the candidate should pass.

Whether Borderline Groups provides any real advantage over the original marking scheme is far from clear to me. A cynical view might be that by being able to quote "Borderline Groups" when asked how the pass mark is calculated, the exam gains the ability to hide behind the apparent objectivity and status of a named method with references in the scientific literature, but ultimately it probably has no greater proven advantage than the original marking scheme, and it is clearly lacking in transparency. Unless solid reasons can be put forward for its use, beyond the mere assertion of the 2003 Review Group that it might be better, then it might be better dropped.

A particular problem with the A-B-C-D-E marking scheme and the Fail-Borderline-Pass marking scheme. Something that is far from transparent at present is why examiners are marking on two separate marking schemes, using the old A to E scale for rating individual objectives, but then use Pass/Borderline/Fail for their overall judgement. As has already been seen in table 1, C (Adequate) on the 5-point does not immediately translate to Borderline on the three-point scale. The problem, as described earlier, is that Adequate and Borderline are not commensurate.

The history and background to adding one standard error of measurement

As explained above, the PLAB Part 2 pass mark currently adds one SEM to the total pass mark of the individual stations[35]. The history of that is described in the paper for Item 11 of the 29th January 2013 meeting. The 1999 Review was vague on the details of standard setting, and commented only that, "A standard setting exercise will follow, using recognised methods which incorporate the performance of the reference groups mentioned in paragraph 25 and the judgements of assessors." (para 26). The 2003 Review said though:

> "As a result of the 1999 review it was agreed that borderline candidates should be given the benefit of the doubt on the basis that no assessment is exact. Consequently candidates falling one standard error of measurement below the notional passing score pass the examination. **At the time the**

[35] The current marking scheme, as did the marking scheme for the Canadian MCCQE Part 2, uses what strictly is SEmeas (which estimates the probable error of an actual score from a true score), whereas there is a strong argument that it should have SEest, the standard error of estimation (which estimates the probable error of a true given from an actual score)(McManus, 2012). The issue will not be considered any further here, as it rapidly becomes complex and technical. Suffice it to say that the difference between SEmeas and SEest can be fairly large, in part because the reliability of PLAB Part 2 is fairly low, at about 0.6 or so, and in part because the pass mark is a reasonable distance below the mean.

> **intention was that for Part 2 the reverse would be true but this was never implemented.** We took a different approach and explored the premise that it would be more important to prevent 'false passers' from going on to Part 2 by failing candidates whose scores fell one standard error of measurement above the notional pass mark. This would have reduced the possibility of a doctor who had not reached the required standard from slipping through the net. We asked a psychometrician to model this system on two past examinations. The table below [in the 2003 review showed] that a large number of candidates' scores are bunched around the mean score and the proposed system had the effect of reducing the pass rate considerably. [New para] We have, therefore, concluded that, for the time being, the notional pass mark should be used. The table below shows that this would reduce the pass rate to a lesser extent. (para 38 and 39)".

The use of SEM in the Part 2 examination is not discussed as far as I can tell, but current usage might be seen to derive from "for Part 2 the reverse would be true", even though the 2003 Review also removed the use of -1 SEM in Part 1.

As a result of the Review a Part 2 Review Implementation Sub-Group was set up (see agenda item 11, 29/1/2013) and it implemented the current marking scheme. In particular, a) the A-E scoring scheme was retained; b) two separate criteria were implemented, "to make sure that candidates could not compensate for low marks in some stations by achieving high marks"; c) the pass mark was set + 1SEM on the basis that the Medical Council of Canada ("on whose OSCE the PLAB Part 2 examination was originally modelled") implemented that system (and it also required passes on a minimum number of stations (variable in the Canadian case and fixed at nine for PLAB Part 2). Item 11 also noted, as mentioned above, that as a result of the various changes the pass rate dropped "from around 80% to around 70%"[36].

The impact of adding one standard error of measurement to the pass mark.

Current practice for the Part 2 examination uses a complex marking scheme whereby candidates have to pass 9 stations individually and also have a total station score greater than the sum of the individual station pass marks plus one standard error of measurement (see earlier). A notable feature of the use of +1 SEM is that it was also implemented at the same time as the number of stations which had to be passed was reduced from 10 on the original marking scheme, to 9 on the current marking scheme. A relevant detail for understanding these various changes is that, as Figure 1 has shown, the pass rate for Part 2 fell by 8.7 percentage points, which is not inconsiderable. *Clearly it is necessary to understand which features of the current marking scheme resulted in the substantial fall in the pass rate*.

Three things changed when the current marking scheme replaced the original marking scheme:

a. 9 passes were required instead of 10
b. The total score had to be greater than an overall pass mark
c. The overall pass mark consisted of the sum of the individual station pass marks with one SEM appended.

Table 7 shows the effect of remarking the examination in post-2008 candidates using different marking schemes.

[36] The Implementation also removed the -1SEM previously appended to the Part 1 pass mark (i.e. the pass mark increased), and " the annual pass rate for Part 1 dropped from around 60% to around 40%" (Item 11, para 11).

Table 7: Percentage passing using different marking schemes. The original marking scheme is pre-2008 and all others are post-2008.		
	Criterion/a for a pass	Pass rate
(a)	*Original marking scheme (≥10+ passes and ≤1 Severe Fail)*	*77.9%*
(b)	**Current marking scheme (≥9 passes and Total ≥ (Total Pass mark + 1 SEM)**	**69.2%**
	Alternative marking schemes for exams since May 2008…	
(c)	≥9 passes and Total ≥ (Total Pass mark + 2 SEM)[37]	57.6%
(d)	≥9 passes and Total ≥ (Total Pass mark) *[i.e. exclusion of 1 SEM]*	71.6%
(e)	≥ 10 passes *[only]*	58.0%
(f)	≥ 9 passes *[only]*	71.7%
(g)	≥ 8 passes *[only]*	82.9%
(h)	Total ≥ (Total Pass mark + 1 SEM) *[only, without requirement to pass 9 stations]*	78.6%
(i)	Total ≥ (Total Pass mark) *[only, without addition of 1 SEM]*	90.6%

With the current marking scheme (row b) it is much harder for candidates to attain 10 station passes than with the original marking scheme (row a), the rate falling from 77.9% to 58.0% (and we have already noted that the pass mark for stations has risen with the introduction of borderline groups). Even 9 passes alone (row f) are only achieved by 71.7%, and to get close to the original pass rate one needs a criterion of just 8 station passes (row g), which would be attained by 82.9%. That may well not have been acceptable to examiners since it is close to the 50% level of 7 stations.

The introduction of a requirement based on total score was new to the current marking scheme, and was imported from the Canadian exam. If candidates only had to attain a total mark greater than the Total Pass Mark (i.e. the sum of all the individual station pass marks, which is what the 2003 Review proposed) then 90.6% would achieve it (row i), the pass rate being so high as good performance on some stations is presumably compensating poorer performance on other stations. Adding one SEM to the criterion (row h) reduces the pass rate to 78.6% which is close to the pass rate under the original marking scheme (row a).

Taken overall it is clear that the main reason for the pass rate falling with the introduction of the current marking scheme was that it was much harder to attain nine individual station passes. Including 1 SEM had relatively little effect, reducing the pass rate only by a small amount (although had 2, 3 or 4 SEMS been added then the overall pass mark would have had a progressively greater effect, relative to achieving nine passes).

Why is it harder to pass stations with Borderline Groups than with the original marking scheme? The original marking scheme passes a candidate on a station if they receive a grade midway between a D and a C, a score of 1.55. Going back to table 1, of the 26460 judgements, 20914 (79.04%) are C or above and therefore passes. In contrast, for the F-B-P grades, 4456 are borderline, and from the mathematics of the Borderline Group method, only half of those cases will end up in a pass being awarded. In addition there are 15344 awarded a pass, meaning that only in 17572 (66.4%) of cases results in an outcome. The difference between 79% of passes and 66% of passes is the reason why candidates under the original marking scheme can achieve 10 passes but many fewer can even achieve 9 passes under the current marking scheme.

Although PLAB Part 2 uses Borderline Groups, the Literature Review (McLachlan, Illing, Rothwell, Margetts, Archer, & Shrewsbury, 2012), in its Appendix E, provided evidence that Borderline Groups

[37] This is equivalent to the analysis carried out in Annex B to Agenda Item 11, 29th January 2013.

generally produces higher pass marks and hence lower pass rates than do Borderline Regression and Contrasting Groups[38]. That in part accounts for the lower pass rates after the introduction of Borderline Groups. Which method is the more accurate, as the Literature Review concludes, depends ultimately not on mathematical manipulations but on predictive validity.

What role does weighting play in the marking of the exam?

One of the complexities of PLAB Part 2 is the role of *weighting*. Each of the objectives, of which there are usually three to six, has its own weight, which can vary from 5% to 70%. In practice very few are in the more extreme categories. For the 218 stations used since May 2008, only 9 (0.8%) of 1141 weights were of 5%, only 1 (0.1%) was of 70%, and the commonest weights were for 10% (37.8% of weights), 15 (7.5% of weights), 20% (25.4% of weights), 25% (8.8%of weights) and 30% (11.3%), these together accounting for 91% of the weights.

A key question for weighting, which undoubtedly contributes to the complexity of the exam, and also reduces its transparency to examiners and candidates due to the secrecy of the weights, is whether it is having any real impact on the marking of the examination. In advance it should be said that while Boards of Examiners in general like weighting, feeling that it is a way they can fine-tune assessments and make subtle distinctions, statisticians and psychometricians are generally sceptical of the utility of weighting. For them, weighting is rarely effective, in part because the marks made by examiners tend to be correlated, particularly within examiner but also between examiner (so that it is rare to find an A and an E made for the same candidate). Weights also complicate the calculation of standard errors and the like. Finally, if some topics are felt to be particularly important, then the argument is that they should be asked about more often, so that several judgements are made of them, rather than a single, possibly inaccurate, judgement being multiplied by a factor of two or three. Finally, if a weight is as low as 5%, then it is not clear that it was worth wasting an examiner's time to obtain that information, and it might have been better dropped entirely. Because the objective scores and their weights are available for the various stations it is possible to look at the role of weighting by *rescoring stations based on weighting and no weighting*.

Rescoring the original format examination is easier to make sense of numerically. In 440,530 stations marked before May 2008, the weighted final grade (A to E) was the same as an unweighted final grade in 89.5% of cases. Of the 10.5% of cases where the grade differed, in all cases it differed by only one grade, the weighted grade being higher in 5.5% of cases and lower in 5.0% of cases. In terms of a pass or fail, for 95.8% of stations the outcome was the same with weighted and unweighted scores, candidates passing on a weighted grade but failing on an unweighted grade in 1.7% of cases, with 2.5% showing the reverse pattern.

With the original marking scheme, weighting of objective scores has only a minimal effect on the proportion of candidates passing or failing a station. Numerically it is easy to see why. Consider the situation where there are five objectives, each weighted equally, and an examiner gives B (3 points) to each of them. The mean is 3 and the candidate passes with a B. To fail the station the candidate has to receive 1.5 or less. If the examiner gives one station a grade E (0 points) then the unweighted mean is 2.6 and the candidate still passes, albeit with a C. To fail the weighting on that one station would have to be increased from 20% to 50%, when the mean becomes 1.5 and the grade is a D.

[38] Appendix E of the literature review also asks the rhetorical question of why Pass, Borderline and Fail are always distributed at equal intervals along the abscissa in these methods, and wonders what would happen were it not to be the case. Rasch modelling sets out with the intention of asking precisely that question, and putting the various values on an equal interval scale.

Weighting only therefore affects pass/fail outcomes when weightings of some components are high, when there is a wide range of marks used within a station, and the marks are clustered around the pass-fail borderline. With the original marking scheme the pass/fail borderline is actually a fraction above 1.5, and yet the nearest marks are 2 (C) or 1 (D), meaning that individual marks, even when weighted, have little impact on the eventual outcome.

Exploring the consequences of weighting in the current marking scheme is more complex because the pass marks are calculated using the Borderline Groups method, and that can change with time. For simplicity a notional pass mark has been calculated by averaging the station marks of all candidates who had an EOJ of Borderline, and using that as the station pass mark[39]. The grades allocated, of A to E, are the same, and the method of weighting is similar. As before the stations were scored either with weighting or without weighting (and the pass marks are of course different when there is not weighting as the station totals change somewhat). Of 85,988 stations assessed under the new marking scheme since May 2008, 94.6% were unchanged with or without weighting. Of the 5.4% which did change, 2.7% passed with weighting and failed without, and 2.7% showed the reverse pattern. As with the original marking scheme, the overall conclusion is that weighting makes little difference to the outcome for most candidates.

An important question, which cannot be answered with the present data, is the question of validity. Any marking scheme is arbitrary in some sense, but some can be objectively better than others if they predict subsequent, unrelated outcomes better. The present dataset cannot really answer such questions, although in principle it should be possible if the PLAB data are linked in to other datasets, such as that for MRCP(UK) or MRCGP. If weighted marks were to predict those outcomes substantially better than unweighted outcomes, then that would be an argument in favour of weighting. Likewise, if Borderline Groups is indeed a better marking system than the original marking system, then results since May 2008 should predict outcomes better than results prior to that date. At present though there is no data to answer such questions.

How does the marking scheme of PLAB Part 2 differ from the Canadian MCCQE Part II?

The Part 2 Review Implementation Sub-Group report modified the Part 2 exam in somewhat different ways to that which the Review Group had intended, explicitly making it much more similar to the Medical Council of Canada's MCCQE Part II, which has been described in detail elsewhere in its full 20 station form (Dauphinee, Blackmore, Smee, & Rothman, 1997), and in an abbreviated 10-item sequential form (Smee, Dauphinee, Blackmore, Rothman, Reznick, & Des Marchais, 2003). There are a number of key differences from PLAB Part 2:

1. MCCQE II has 20 scoring stations, each ten minutes in length, some involving both a clinical and a written component, whereas PLAB Part 2 has 14 scoring stations each of five minutes.
2. Examiners make an overall judgement on each station, although for MCCQE II these are on a six-point scale of Outstanding, Excellent, Borderline Pass, Borderline Fail, Poor and Inadequate, whereas for PLAB Part 2 they are on three points, Pass, Borderline and Fail.

[39] The method gives a good approximation for most stations, although there are a few unexplained cases where the calculation is somewhat different. These are probably due to shifts in weighting or alterations in candidate performance (and hence altering performance of those in the Borderline group) over time. Time was insufficient to really dig into the occasional anomalies but they probably do not alter the big picture.

3. In MCCQE II there is a checklist of behaviourally-anchored scales as well as a checklist of information obtained (see http://mcc.ca/en/exams/qe2/scoring.shtml#SampleOSCE and http://www.mcc.ca/pdf/Rating-Scale_E.pdf), whereas PLAB Part 2 has 3 to 6 objectives each scored on the same generic scale from A to E. It is not clear how weighting works in MCCQE II.
4. Borderline groups is used in both examinations but in MCCQE II it is applied simultaneously to the large number of candidates taking the examination at a particular diet, whereas the small number of candidates in each day of PLAB Part 2 mean that historically derived BG scores are needed.
5. In MCCQE candidates need also to pass a certain number of stations, the number being decided by the examining board with an example cited at 12/20 (60%), compared with a slightly higher proportion of 9/14 (64%) for PLAB Part 2. MCCQE II candidates must also "perform in a professional manner during the whole examination"[40].
6. In MCCQE II there are two borderline groups (Borderline Pass and Borderline Fail), with the pass mark being set between them, whereas in PLAB Part 2 there is only a single Borderline group with the pass mark being set at its mean.
7. MCCQE II is an examination which is taken by all foreign graduates wishing to work in Canada *as well as all graduates of Canadian medical schools*. MCCQE therefore acts as a national licensing examination in Canada, with about 95% or more of Canadian candidates passing the examination[41]. Pass rates are much lower in international medical graduates, making the distribution of marks in MCCQE II somewhat complex and skewed to high levels overall by the Canadian candidates.

Cronbach's alpha for the reliability of the 20-item MCCQE II is reported as being about .73 (Dauphinee et al., 1997), and the reliability of the 10-item MCCQE II is said to be .66, with a comment that it is "quite high for a 10-station OSCE" (Smee et al., 2003). As seen earlier, the alpha reliability for PLAB Part 2 is between about .55 and .65 for 14 stations, which suggests it is less reliable than MCCQE II. For MCCQE II it is stated that with a reliability of about .74, "confidence around the pass-fail cutscore has been 0.99", adding that "adding 5 or 6 more stations would increase reliability as calculated by Cronbach's alpha, but would not increase confidence around the pass-fail score" (Dauphinee et al., 1997). How the value of confidence is calculated is not clear[42]. Downing (2004) refers to confidence in pass-fail outcomes, and references the method of Subkoviak (1998).

For PLAB Part 2, using Subkoviak's method, the confidence around the pass-fail cutscore is about .72 in terms of the coefficient of agreement (compared with chance allocation of about .57), and Cohen's Kappa of about .36 (which is corrected for chance, and random allocation gives a value of 0 and perfect agreement a value of 1)[43]. Neither seems particularly high, and both are compatible with the low-ish alpha values.

[40] It seems that a single failure on this criterion is sufficient to result in an overall failure.
[41] The piloting of a sequential examination was to assess whether or not a shorter first stage might be adequate for most Canadian candidates, with only candidates who perform less well on that part of the examination going on to the longer, 20-station assessment.
[42] Using Subkowiak's table 2 is does not seem possible to get an agreement of .99 for a reliability between .7 and .8.
[43] Based on a reliabilities of between .5 and .6 and a cut-off expressed as an absolute z-score of .50.

Summary: Weighting, standard error of measurement, and borderline groups.

The fairly complex analyses presented above can be fairly simply summarised:

a. The introduction of the current marking scheme in May 2008 resulted in an immediate and a sustained drop in the PLAB Part 2 pass rate by about 9 percentage points. Several things changed in the new marking scheme.
b. *Borderline Groups methodology was used to set the pass marks*. This accounted for most of the change in pass rate and resulted from moving from a system where passing was tied in to a mark of 1.55, which was between an overall grade of C (Adequate) and a grade of D (Fail), to a system which was tied in to a grade of 'Borderline', with half of those on Borderline failing. Despite shifting from 10 C grades to 9 pass grades on Borderline groups, the pass rate still fell, stations becoming harder to pass.
c. *An overall pass mark was also introduced, set at 1 SEM higher than the summed pass mark of all stations*. The original intention was seemingly to use -1 SEM for Part 1 and +1 SEM for Part 2. The former was not introduced but the latter was, and was responsible for about a 1.5% percentage point fall in the pass rate. There is no clear justification for it, either educationally or psychometrically.
d. Other problems also have become apparent, a number of which were present from 2001 but have been exacerbated since 2008.
e. The simultaneous use of two separate marking schemes (Severe Fail through to Excellent) and Fail, Borderline and Pass leads to potential confusion on the part of examiners.
f. Where pre-2008 examiners were mostly clear about the marking scheme (10 A,B or C to pass and no more than 1 E), post-2008 all became murky. A set of C (Adequate) grades could result in failing a station because the pass mark was above 2. Examiners were also not informed about pass marks on stations.
g. Different objectives on a station have, since 2001, been weighted, with the weights not being provided to examiners. In practice the weighting has little consequence and only a tiny percentage of station pass-fail judgements are altered by differential weighting as opposed to the simpler equal weighting.
h. The exam has become less transparent both to examiners and candidates. Whereas the pass criteria were once clear (ten examiners thought a candidate was Adequate), the use of Borderline Groups has obscured much of that clarity.
i. Taken overall, there would seem little support for the use of 1 SEM, the use of Borderline Groups, or the use of Weighting, making the examination less transparent, with no obvious improvement of internal consistency. The additional complexity of SEMs, weighting, etc, all result in significant costs in medical and administrative time, with no obvious benefits in compensation.
j. There may well be an argument for returning either to the original marking scheme or some variant on it.
k. Modelling using FACETS provides a clear demonstration of how effects due to stations and objectives within stations can be deconstructed, in order to obtain estimates of candidate ability which are independent of them.

l. FACETS also provides, for each candidate, an estimate of the standard error of the estimate of their true ability[44].

What is the underlying structure of the marks awarded to candidates?

The PLAB Part 2 examination has fourteen stations, and in the original and current marking schemes all of those 14 stations are treated as equivalent, so that a pass in any of them is regarded as equivalent in order to gain the nine or ten passes needed for the current or original marking schemes. The stations however measure different skills, and the present examination ensures that each carousel consists of stations assessing[45]:

- Communication Skills (C), n=4
- Examination Skills (E), n=4
- History Skills (H), n=4
- Practical Skills (P), n=2

At present the examination is entirely *compensated* in the sense that a mark in any can substitute for a mark in any other. The underlying assumption in mathematical terms is that candidates differ on only a single dimension, with better candidates being equally likely to be better on any of the four skills. Observation of doctors in practice suggests, however, that some are better at some skills than others, and on that basis the PACES exam of the MRCP(UK) has since 2009 assessed seven different skills in its exam, the marking scheme being changed so that a pass at a certain minimum level is required in all seven skills[46]. That change was driven by a realisation that under the previous, compensated, marking scheme there were candidates with poor communication skills who were compensating by having excellent examination skills, and vice-versa, and yet physicians believed that both sets of skills were essential for good practice. Whether there are truly seven skills is as yet not clear, but the statistical process of factor analysis has certainly confirmed that there are at least two distinct groups of skills, which can be described as Communication and Examination Skills. It therefore seems desirable to assess the question of whether PLAB Part 2 results provide evidence for two or more skill domains, and if so, what are the implications for marking the examination.

The fundamental tool in multivariate statistics for deciding on the dimensionality of a set of items is *Factor Analysis*. Without going into technical details, it looks at the correlations between items, and puts together those items that correlate most highly, and looks for structure in the patterns of correlations[47]. The analysis found evidence of two clear factors, with the four Communication and

[44] This standard error varies according to the overall ability of the candidate, estimates at the extremes of the range being less accurate and having larger SEs than those near the mean, as expected (but not taken into account by traditional estimates of SEM).

[45] This classification does not seem to appear in either the 1999 or 2003 Reviews, but it looks sensible and covers the major domains which a clinical (rather than an MCQ) examination would wish to assess.

[46] The skills, which were identified are studying various syllabuses for Physicians and other documents such as Good Medical Practice, are Physical Examination, Identifying Physical Signs, Clinical Communication, Differential Diagnosis, Clinical Judgement, Managing Patients' Concerns, and Maintaining Patient Welfare.

[47] For the technically minded, I extracted Principle Components in SPSS 20, used a scree-slope criterion to determine that there were two underlying factors in the fourteen station scores, and then Varimax rotated the two factors to give simple structure. The 14 stations of course differ in having different carousels, but I reordered the stations so that the first four as far as the program was concerned were Communications Skills, the next four were Examination, followed by History-taking and then Practical Procedures, which allowed the structure of the four skills to emerge. The eigenvalues for the candidates taking the examination since May 2008 were 2.579, 1.071, .984, .971, .942, .901, .894, .874, .851, .830, .809, .787, .767 and .741, showing a

four History-taking stations in one factor, and the three Examination and Practical stations in the other. On that basis it was straightforward to calculate simple summary scores for Communication and History (C+H) and Examination+Practical (E+P) skills. For correlational studies we calculated the average of the station totals for the eight or six skills in each set, and for considering possible marking schemes we counted the number of stations passed in each set (from 0 to 8 for C+H and 0 to 6 for E+P). It was also possible to calculate similar scores for candidates before April 2008, although there was a minor complication in that pre-2008 carousels were not always balanced to have the same mix of C, E, H and P stations. All stations were however classified, and therefore average scores were calculated for stations on the C+H skills and E+P skills for correlational purposes.

Correlates with Communication and History Skills (C+H) and Examination and Practical Skills (E+P).

If the two scores identified by factor analysis are real then they should have different correlates with external variables, and the first step was to confirm that that was the case. In the first instance it made sense to follow a recommendation from the 1999 Review, which said, "Research comparing candidates' IELTS scores with their performance in communication skills in the OSCE should be undertaken …"[48].

It should be noted that the following analyses are at the level of the *doctor* rather than at the level of the *candidate*. Candidates generally have only a single set of IELTS scales, and therefore it is misleading to correlate them across multiple attempts. There is a choice between analysing results at the final attempt, the most recent attempt, or the first attempt. In practice it is best to analyse data from first attempts at examinations, since those data tend to be normally distributed with a wide range, whereas pass attempts are inevitably severely truncated at the pass mark (McManus & Ludka, 2012). Overall 38,789 doctors had taken Part 2 at a first attempt.

IELTS scores in relation to C+H and E+P scores

IELTS[49] provides a total score, as well as scores on four sub-scales, Listening, Reading, Speaking and Writing. IELTS requirements by PLAB have varied, becoming higher over the years. Maximum scores are 9, and in PLAB takers in this data set range from 6.5 to 9. Some candidates are exempt from PLAB if they provided evidence of having taken medical training in a university using English.

Table 8 shows correlations of the C+H and E+P skills with IELTS scores. The largest correlations are with C+H skills, particularly with production skills, although reception skills also shows correlations. E+P skills show much smaller correlations. Taken overall it is clear that C+H performance differs substantially in relation to IELTS, as might be expected. The differences also validate the separation of C+H and E+P scores.

---

strong first factor, and then a second factor which appears above the screen-line based on the other eigenvalues.

[48] 1999 Review, p.1, recommendation *m*. It actually continued, "… as soon as reliable data are available for analysis", although to my knowledge this is the first attempt to do so.

[49] International English Language Testing System.

Table 8: Correlations of IELTS scores with C+H and E+P performance, and overall mean score on Part 2. N=30,750 and hence all correlations reach conventional levels of significance. Correlations >.2 are in bold, and correlations >.1 are in italics.				
		Communication+ History skills	Examination + Practical Skills	Total Part 2 score
IELTS scores	Total score	***.242***	*.100*	*.190*
	Listening	*.122*	.049	.094
	Reading	*.136*	.047	*.102*
	Speaking	***.202***	.055	*.143*
	Writing	***.224***	*.135*	***.201***
	Production	***.266***	*.121*	***.216***
	Reception	***.154***	.057	*.117*

Relationship of C+H and E+P scores to Part 1 PLAB results.

Candidates who take PLAB Part 2 have previously taken PLAB Part 1. An important practical question concerns the extent to which Part 1 results are predictive of performance in Part 2, particularly of the two subscores. Overall there is a correlation of .271 between Part 1 at first attempt and Part 2 at first attempt. This is likely to be an under-estimation of the true correlation because of restriction of range, only the best candidates passing Part 1 and then going on to Part 2[50].

Part 1 has a total score (relative to the pass mark, since pass marks vary for each day), and four sub-scores, on Context, Diagnosis, Investigations and Management. Table 9 shows the correlations with Part 2 overall performance and C+H and E+P sub-scores.

Table 9: Correlations of Part 1 first attempt scores with C+H and E+P performance, and overall mean score on Part 2. N=32,093 and hence all correlations reach conventional levels of significance. Correlations >.2 are in bold, and correlations >.1 are in italics.				
		Communication+ History skills	Examination + Practical Skills	Total Part 2 score
Part 1 mark (first attempt)	Part 1 total	***.230***	***.271***	***.258***
	Context	***.237***	***.274***	***.257***
	Diagnosis	***.259***	***.307***	***.295***
	Investigations	*.190*	***.221***	***.208***
	Management	.069	.087	.090

There is definite tendency for the E+P skills to be better predicted by Part 1 scores than are the C+H skills, although the effect is not large. Scores on Diagnosis are the best predictors of Part 2 scores, followed by Context and Investigations, with Management scores having low correlations with Part 2.

For completeness it is also of interest to see how IELTS scores relate to Part 1 marks (although it should be emphasised that this is only in the subset who have passed Part 1 and gone on to Part 2). Table 10 shows the correlations.

[50] Without full data on Part 1, rather than just data on candidates who passed Part 1, it is not possible to correct the correlation for restriction of range.

Table 10: Correlations of IELTS scores with Part 1 scores. N=30,750 and hence all correlations reach conventional levels of significance. Correlations >.2 are in bold, and correlations >.1 are in italics.						
		Part 1 Context	Part 1 Diagnosis	Part 1 Investigations	Part 1 Management	Total Part 1 score
IELTS scores	Total score	*.165*	*.162*	*.147*	*.168*	*.193*
	Listening	*.108*	*.109*	.088	*.157*	*.147*
	Reading	***.201***	*.199*	*.185*	***.205***	***.240***
	Speaking	.008	.004	.006	.024	.007
	Writing	*.124*	*.123*	*.113*	.055	*.119*
	Production	.086	.083	.077	.050	.082
	Reception	*.187*	*.186*	*.165*	***.217***	***.233***

In contrast to Part 2, and particularly to Part 2 Communication and History Skills, it is IELTS Reception Skills, in particular Reading, which correlate with performance in the Part 1 examination. The subscores show broadly similar patterns, although it is interesting that it is the Management Sub-scores which particularly correlates with Reception Skills, suggesting that the text in these questions may be particularly demanding.

Summary of C+H and E+P sub-scores.

The factor analysis strongly suggests that Part 2 stations can be sub-divided into two distinct groups, those assessing Communication and History Skills and those assessing Examination and Practical Skills. That division is validated by the distinct patterns of correlation of C+H and E+P with IELTS scores, C+H marks particularly correlating with production skills of Speaking and Writing, whereas E&P marks shows only small correlations with IELTS scores. Correlations of Part 2 scores with Part 1 exam results also suggested a division, it now being E+P skills which show the higher correlations. Finally it is worth noting that Part 1 scores mainly relate to IELTS Reception Skills, particularly Reading, in contrast to Part 2 C+H scores which relate to IELTS Production Skills. Taken overall, these analyses strongly support the idea that C+H and E+P scores are assessing different underlying constructs, and they raise the question of whether candidates passing in one type of station are compensating for poorer performance in the other type of station.

Passes attained in C+H and E+P stations.

Candidates post-2008 are assessed on eight C+H stations and six E+P stations, and receive a pass or fail grade on each of those stations. Table 11 shows the numbers of C+H and E+P passes, firstly for all candidates (with numbers gaining 9 passes in total indicated), and secondly for candidates who passed the exam under the current marking scheme.

Table 11*:* Numbers of C+H and E+P stations passed under the current marking scheme. N=7492 for all candidates, N=5182 for passing candidates. Shaded cells indicate candidates passing 9 or more stations. Candidates in bold have passed at least 50% of the C+H and 50% of the E+P stations.

a) All post-2008 candidates.

		Number of C+H stations passed									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	*Total*
Number of E+P stations passed	**0**	1	4	5	7	9	11	6	1	0	*44*
	1	3	8	19	31	34	32	26	17	2	*172*
	2	3	22	33	77	112	139	92	62	19	*559*
	3	5	20	64	117	**197**	**276**	**269**	**205**	**90**	*1243*
	4	3	31	55	145	**303**	**419**	**516**	**399**	**164**	*2035*
	5	3	11	45	129	**235**	**400**	**538**	**493**	**353**	*2207*
	6	2	3	18	39	**105**	**190**	**320**	**335**	**220**	*1232*
	Total	*20*	*99*	*239*	*545*	*995*	*1467*	*1767*	*1512*	*848*	*7492*

b) All post-2008 candidates passing under current marking scheme

		Number of C+H stations passed									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	Total
Number of E+P stations passed	**0**										
	1									1	*1*
	2								53	18	*71*
	3							**228**	**196**	**89**	*513*
	4						**362**	**496**	**398**	**164**	*1420*
	5					**202**	**391**	**536**	**493**	**353**	*1975*
	6				32	**105**	**190**	**320**	**335**	**220**	*1202*
	Total				*32*	*307*	*943*	*1580*	*1475*	*845*	*5182*

Under the current marking scheme, one of the criteria for passing Part 2 is that at least 9 stations should be passed. The shaded area in table 11a shows that *there are candidates who pass overall by passing all 8 C+H stations but only a single E+P station, and other candidates who pass all 6 E+P stations but only 3 C+H stations*, in each case the minima which are possible. Table 11b shows that incorporating the full marking scheme, with its requirement for a total score + 1SEM does little to mitigate that situation, there still being candidates with extreme combinations of C+H and E+P passes.

The consequences of setting pass marks for both C+H and E+P stations.

At present the marking scheme is compensated, requiring only that 9 stations are passed. A *non-compensated marking scheme* might require, for instance, that candidates pass at least 50% of the C+H stations (4), and 50% of the E+P stations (3); those candidates are shown in bold in table 11. Clearly that could also be used in conjunction with a requirement that candidates pass at least 9 stations (see table 11b), and there are many variations on the theme (such as requiring 5 C+H stations and 4 E+P stations). Some variations would result in the pass rates going up or down in relation to the current marking scheme, and all require modelling, particularly in relation to the inclusion of an overall passmark with or without 1 SEM.

At present the PLAB Part 2 exam has 8 C+H stations and 6 E+P stations. The latter is somewhat on the small side for setting a separate pass mark if it were felt to be desirable then *consideration may*

need to be given either to having 7 C+H stations and 7 E+P stations, keeping the exam at its current length, or increasing the number of E+P stations to 8, with a small increase in the length of the exam. There may also be an argument for splitting PLAB Part 2 into two separate components, one concerned with Communication and History-Taking and the other with Examination and Practical Skills.

Conclusion: The underlying structure of the marks awarded to candidates.

There is strong evidence that the Part 2 stations can be divided into Communication and History-taking (C+H) and Examination and Practical (E+P) stations, both from the internal factor structure and the correlation of sub-score marks with IELTS scores and Part 1 scores. Some candidates are stronger on one set of stations and weaker on the other, with the candidates who are weaker on C+H stations also having lower IELTS scores. Whether such candidates perform differently as doctors cannot be ascertained from the current data set, but it is at least plausible, and may well become more apparent when the data are linked to the MRCP(UK) and MRCGP datasets in work which is currently under way. If it is felt that candidates should achieve minimum attainment levels on C+H and E+P stations then there are various ways of doing that, although all would have consequences for the pass rate and require modelling.

Overall discussion.

Alternatives to the current marking scheme.

The 2003 Review Group, when suggesting the Borderline Group, in fact said, "The standard should be set for the OSCE using the borderline group method, **if modelling the system proves successful. Otherwise another standard setting system should be introduced**." It is not clear that the system has demonstrably proved effective in terms of transparency, it does not obviously provide any clear advantages in terms of reliability over the original marking method, and it is clearly less transparent. It may be then that "**Otherwise**" has now arrived and that "**another standard setting system should be introduced**". There are many possible options, but the two most extreme should perhaps be presented:

1. **A return to the original marking scheme.** It is not clear that the current marking scheme has any obvious benefits over the original one, it is more complex and less transparent, neither to any obvious benefit. The original marking scheme makes clear as its primary criterion that of 14 examiners, all of whom are experienced medical professionals who have been trained in the method of examining, a minimum of 10 are satisfied that the standard of the doctor is adequate. That is easy to explain to the public and to candidates, and examiners are also aware that their actions have consequences. It might be better if it were complemented by 'post-circuit review', examiners assessing the extent to which their judgements agreed, and in the case of large disagreements reflecting on how the different perceptions related to the performance of the candidates and the intentions of the examination.

2. **Using FACETS to set the standard using statistical equating**. Examiners would make judgements on an A to E or similar scale for each objective at each station. These judgements would be combined with all previous data on the examination for, say, the

previous three years, in order to take into account differences in examiner hawkishness and difficulty of objectives, and FACETS would calculate a "fair mark" on a standard logit scale. The p[51]ass mark would be set at the overall boundary of Adequate and Fail. In effect the method removes the need, a) for weighting or otherwise reviewing the difficulty of stations, b) to use borderline groups or other methods for setting the pass mark, and c) to monitor hawkishness of examiners and to try and influence a trait which may be fairly intractable to persuasion. The method is somewhat less transparent than the original marking scheme, but statistically it is undoubtedly superior. There would be a need to explain the method to candidates, but that is no different, for instance, to using item response theory for statistical equating of MCQ exams.

Separate marks for Communication and Examination stations? There seems little doubt that C+H and E+P stations are behaving differently, and in all probability that could well result in different outcomes in clinical practice. There are several options on handling this, which vary in their complexity and cost implications:

1. **Calculating separate marks for C+H and E+P stations and having a minimum pass mark on each**. This is easy to implement in the present examination, although having 8 C+H stations and only 6 E+P stations makes it a little unbalanced. Solutions might be:
    a. Having 7 stations on each;
    b. Increasing the examination to 16 scoring stations (17 actual stations), and having 8 stations on each station type.
2. **Splitting Part 2 into two separate exams**. If it were felt important to assess both C+H and E+P stations to current standards (and those do not have a particularly high reliability on their present length), then it might be possible to have, say, two C+H circuits of 16 stations in the morning, and only candidates passing C+H would then go on to a single E+P circuit with 16 stations in the afternoon (or vice-versa). There might be problems in balancing numbers, but no doubt that could be modelled and taken account of.

Questions that need answering.

This is a lengthy report, and it has raised many questions. For convenience, many of them are assembled in the Appendix, below, entitled *Issues to be considered*. Some issues are, of course, inter-related, so that answers to one may preclude the need for answers to others.

[51] MRCP(UK) Part 1 and Part 2 have been setting standards for their exams by statistical equating since about 2008.

Appendix: Issues to be considered.

1. **Is there any advantage to weighting objectives within stations?** In the current marking scheme and the original marking scheme the detailed weights for each objective are kept secret from examiners.
    a. **Should weights be kept secret from examiners?** The implicit assumption is that examiners are not to be trusted with this information as they might abuse it. If examiners cannot be trusted in this way, then they presumably should not be used at all. Examiners are intelligent, experienced, professionals and it seems unlikely that they are not divining most of the weights implicitly, and thereby making sets of judgements which give the result they intended.
    b. **Should weights be used at all?** If an objective has only 10% of the marks on a station, why is the objective there at all? An exam marked with equal weights for all objectives produces barely distinguishable results from the current marking scheme.
2. **Should the use of Borderline groups be altered or even removed?** Borderline groups ultimately relies on a circular argument for converting marks on individual objectives to total station marks relative to a station pass mark, allowing stations to be passed or failed.
    a. **Is there any benefit from the continual updating of the pass mark at each outing of a station?** At present pass marks hardly vary across the various outings.
    b. **Should updating of pass marks take place using only more recent outings?** There might have been an argument that pass marks should be based more on recent outings of a station than more distant ones, so that changes in expectations about currently acceptable practice could be incorporated into the pass mark. That might have been the original intention but is certainly not the case at present.
    c. **Should the EOJ scales contain Borderline Fail and Borderline Pass?** The Canadian MCCQE Part 2 uses both borderline fail and borderline pass marks with it the pass mark set between the two groups, in contrast to the single Borderline group in the current marking scheme.
    d. **Why are there two marking scales (A to E: Excellent to Serious Fail) and Pass/Borderline/Fail?** Is there an argument for replacing the two separate scales with a single one, perhaps with five or six scales incorporating Borderline Pass and Borderline Fail. The present scheme has few obvious advantages and is ripe for creating confusion.
    e. **Should objectives be scored not on common, generic scales (A to E) with only single or two word descriptions, but be replaced with behaviourally anchored scales at all points of the scale as with MCCQE?** Unless it is specified what Adequate means it is not clear quite how examiners should know what it is for any particular objective. However multi-facet Rasch analysis benefits from having a standard scale for all assessments.
    f. **Should Borderline Groups be replaced by Borderline Regression?** Borderline regression has the advantage that it could be used on single days even when no single candidate is scored as Borderline, as the regression line can still fit through the other groups. Borderline Groups is also known to have higher failure rates than Borderline Regression.
    g. **Is there an argument for scrapping Borderline Groups altogether?** The original marking scheme had the advantages of clarity and transparency, whereas the

current marking scheme has no such advantages. There is no strong evidence that examination reliability is higher with the current marking scheme. It may be that predictive validity is higher for the current marking scheme but that still remains to be seen in future analyses.

3. **Is there an argument for returning to the original marking scheme, rather than using the more complex current marking scheme?** Although Borderline Groups is much used, it is far from clear that is has obvious benefits over the original marking scheme, particularly since each objective is marked on an implicitly criterion-referenced scheme which includes the terms Adequate and Fail. If the exam were to return to the simpler, more transparent marking scheme, then some modifications might be desirable:
    a. **Should the five-point scale (A to E) be replaced by a six-point scale similar to that of MCCQE Part 2?** MCCQE uses Outstanding, Excellent, Borderline Pass, Borderline Fail, Poor and Inadequate, whereas PLAB uses Excellent, Good, Adequate, Fail, Severe Fail. That could be modified, perhaps to Excellent, Good, Borderline Pass, Borderline Fail, Inadequate and Serious Fail. An advantage would be that the scale a) separates out Borderline Pass and Borderline Fail, making examiners produce explicit decisions about which side of the borderline a candidate is, and b) the scale could also be used for EOJs.
    b. **Should examiners make decisions of Excellent vs Good, etc?** PLAB Part 2 is an examination which asks a single questions – Is the candidate competent or not? It is not an examination designed to separate excellent performance from good. With the current marking scheme, in effect a mark of Excellent allows a candidate in part to compensate elsewhere for a mark of Adequate vs Fail, which does not seem desirable if overall competence is the main decision of the exam. FACETS does take that into account to a large extent, and having more scale items generally produces higher reliability.
4. **What are the consequences of there being separate Communication and History (C+H) and Examination and Practical (E+P) Skills?** There seems little doubt that these skills are separate, both because of the internal structure of the stations revealed by the factor analysis, and because the C+H and E+P skills have different correlates with PLAB Part 1 and with IELTS scores.
    a. **Should there be separate pass marks for C+H and E+P skills?** Separate pass marks can be set for C+H and E+P. If problems of communication are felt to be particularly problematic in general for UK doctors, then a separate pass mark for each skill would flag up that both competencies are seen as important.
    b. **If there are separate pass marks for C+H and E+P, should there be equal numbers of the two types of station?** With the present exam structure, there could be seven C+H and seven E+P stations? Reliability of the exam overall though is not particularly high, and there might be a stronger argument for having eight C+H and eight E+P stations. Or perhaps even ten of each if reliability is seen as a major issue.
    c. **Should all stations potentially have both C+H and E+P marks?** Most real clinical interactions involve both C+H and E+P components. There might be an argument for classifying all objectives on all stations as C+H or E+P, so that all of the C+H objectives for all stations would contribute to an overall decision on C+H pass-fail.
5. **Practical issues of the implementation of PLAB Part 2.** The exam is very labour-intensive at present, in several ways:

a. **Should the scanning of mark sheets be brought in-house?** At present paper mark sheets are couriered from Manchester after each days examining to be scanned elsewhere, and for basic statistics, including the reliability, and the SEM to be calculated, and these values are then sent back to London to be processed into a pass mark. There might be an argument for the entire process to be brought in-house, particularly now that data-processing is becoming more efficient.
b. **Should paper mark sheets be replaced with iPADS or equivalent?** Electronic capture of marks would allow those monitoring the examination in Manchester to have simultaneous awareness of the marks being awarded to candidates by examiners. This might be an argument for not bringing paper scanning in-house but instead to concentrate on electronic data capture.
c. **Should examiners be discussing candidates as a group?** At present examiners have no idea who has passed or failed the exam overall? That contrasts with, say, MRCP(UK) PACES where, at the end of each circuit, examiners see all of the examination results, provisional pass-fail discussions are put up on the screen, and examiners who are clearly anomalous are asked to discuss their decisions. That makes examiners feel part of a corporate body, and helps to make clear what are the common standards for the exam.
d. **Should examiners exam on different stations during a day?** An examiner currently sees the same station all day long, typically on 48 occasions. That is tedious for examiners, and statistically also confounds examiners with stations within days. Were examiners to rotate through three different stations during a day then it would probably be less tedious, and would allow a clearer separation of station effects from examiner effects.
e. **How are hawks and doves handled?** At present there is a rather cumbersome procedure whereby hawks and doves are identified as statistical anomalies at the end of each day's examining and examiners are notified of their tendency. Whether that has any obvious effect on their future behaviour is far from clear, but it seems unlikely. The FACETS modelling suggests that hawk-dove effects are broadly similar to those in MRCP(UK) PACES where no feedback is given. Is the current feedback to examiners of any use?
f. **Might it be better to take hawk-dove effects into account statistically?** It is probable that hawk-dove effects have little impact on eventual outcome, mainly because effects are not large, and candidates are typically seen by 14 examiners, half of whom on average will be hawks and half will be doves. Hawk-dove scores for individual examiners can readily be calculated using FACETS, and, if necessary, the average hawk-dove score for a day or for a carousel could be taken into account statistically. Whether it would have any substantial effect seems however to be doubtful, but it can be calculated.
g. **Should candidate ability, examiner differences and station differences all be modelled simultaneously using FACETS?** Multi-facet Rasch modelling can simultaneously estimate candidate ability scores, examiner differences and station differences, with candidate scores taking into account the particular station mix they have received, or the hawkishness of their particular examiners. Those ability scores are then on a standard Rasch scale and are directly comparable across diets, etc.. The pass mark would differ slightly between diets, but it does so at present. A subtle

problem is that candidates 'passing' the same number of stations can either pass or fail according to which stations they passed, passing difficult stations counting more than passing easy stations.

h. **Could Rasch modelling be used to balance carousels/diets?** The difficulty of a particular mix of stations in a carousel can be calculated in advance from FACETS. By careful selection of stations it should be possible to ensure that the overall difficulty of a diet is within pre-determined limits. Likewise, the overall hawkishness of a set of examiners can be pre-determined from examiners' previous performance, and the average hawkishness of a carousel set within pre-determined limits. That would then mean that a standard pass mark can be used for each diet, which is easier for candidates to understand.

i. **Should examiners meet after each circuit for a review of marks, a discussion of differences, and a resolution of large differences?** As mentioned earlier, there is a risk that examiners are disengaged with the examining process, resulting in a lower reliability of marks. A debriefing meeting after each circuit would have positive benefits for examiners, both socially and perhaps also in terms of the reliability of the examination.

Bibliography

Clauser, B. E., & Linacre, J. M. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions (www.rasch.org/rmt/), 13*, 696.

Cleland, J., Dowell, J., McLachlan, J., Nicholson, S., & Patterson, F. (2012). *Identifying best practice in the selection of medical students (literature review and interview survey)*. London: General Medical Council (http://www.gmc-uk.org/Identifying_best_practice_in_the_selection_of_medical_students.pdf_51119804.pdf) .

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Dauphinee, D., Blackmore, D. E., Smee, S., & Rothman, A. I. (1997). Using the judgements of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Advances in Health Sciences Education, 2*, 201-211.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*, 1006-1012.

Downing, S. M., 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*, 50-57.

Elder, A., McAlpine, L., Bateman, N., Dacre, J., Kopelman, P., & McManus, I. C. (2011). Changing PACES: developments ot the examination in 2009. *Clinical Medicine, 11*, 231-234.

General Medical Council. (2010). *Supplementary Guidance: Reliability issues in the assessment of small cohorts*. London: General Medical Council (http://www.gmc-uk.org/Cohorts_GMCversion_31379265.pdf).

General Medical Council. (2013). *GMC Glossary for the Regulation of Medical Education and Training*. London (undated on original; dated Feb 2013 on website): General Medical Council: http://www.gmc-uk.org/GMC_glossary_for_medical_education_and_training_1.1.pdf_48186236.pdf.

McLachlan, J., Illing, J., Rothwell, C., Margetts, J. K., Archer, J., & Shrewsbury, D. (2012). *Developing an evidence base for the Professional and Linguistics Assessments Board (PLAB) test*. London: General Medical Council.

McManus, I. C. (2012). The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. *Medical Teacher, 34*(7), 569-576.

McManus, I. C., & Ludka, K. (2012). Resitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations. *BMC Medicine, 10*((doi:10.1186/1741-7015-10-60)), 60.

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education, 6: 42 (http://www.biomedcentral.com/1472-6920/6/42/abstract)*.

Postgraduate Medical Education and Training Board. (2009). *Reliability issues in the assessment of small cohorts (Guidance 09/1)*. London: PMETB (www.pmetb.org.uk).

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Schumacker, R. E., & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409.

Smee, S. M., Dauphinee, W. D., Blackmore, D. E., Rothman, A. I., Reznick, D. N., & Des Marchais, J. E. (2003). A sequenced OSCE for licensure: Administrative issues, results, and myths. *Advances in Health Sciences Education, 8*, 223-236.

Subkoviak, M. J. (1998). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Tamblyn, R. A. M., Dauphinee, W. D., Hanley, J. A., Norcini, J., Girard, N., Grand'Maison, P. et al. (2002). Association between licensure examination scores and practice in primary care. *Journal of the American Medical Association, 288*(3019), 3026.

Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The Standard Error of Measurement is a more appropriate measure of quality in postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations. *BMC Medical Education (www.biomedccentral.com/1472-6920/10/40), 10*, 40.

Till, H., Myford, C., & Dowell, J. (2013). Improving student selection using multiple mini-interviews with multifaceted Rasch modelling. *Academic Medicine, 88*, 216-223.

Tor, E., & Steketee, C. (2011). Rasch analysis on OSCE data: An illusgtrative example. *Australasian Medical Journal, 4*(6), 339-345.

Zhang, X., & Roberts, W. L. (2012). Investigation of standardized patient ratings of humanisc competence on a medical licensure examination using Many-Facet Rasch Measurement and generalizability theory. *Advances in Health Sciences Education, Published online 6th Dec 2012*, no page numbers.

Figure 1: Number of candidates (solid bars) and pass rate (red dots) for PLAB Part 2 from 2001 to August 2012, by tertiles (four-month periods). The examination changed its marking scheme in the second tertile of 2008, shown by the different colours of the histogram. The dashed lines show the average pass rate for the original marking scheme and the current marking scheme.

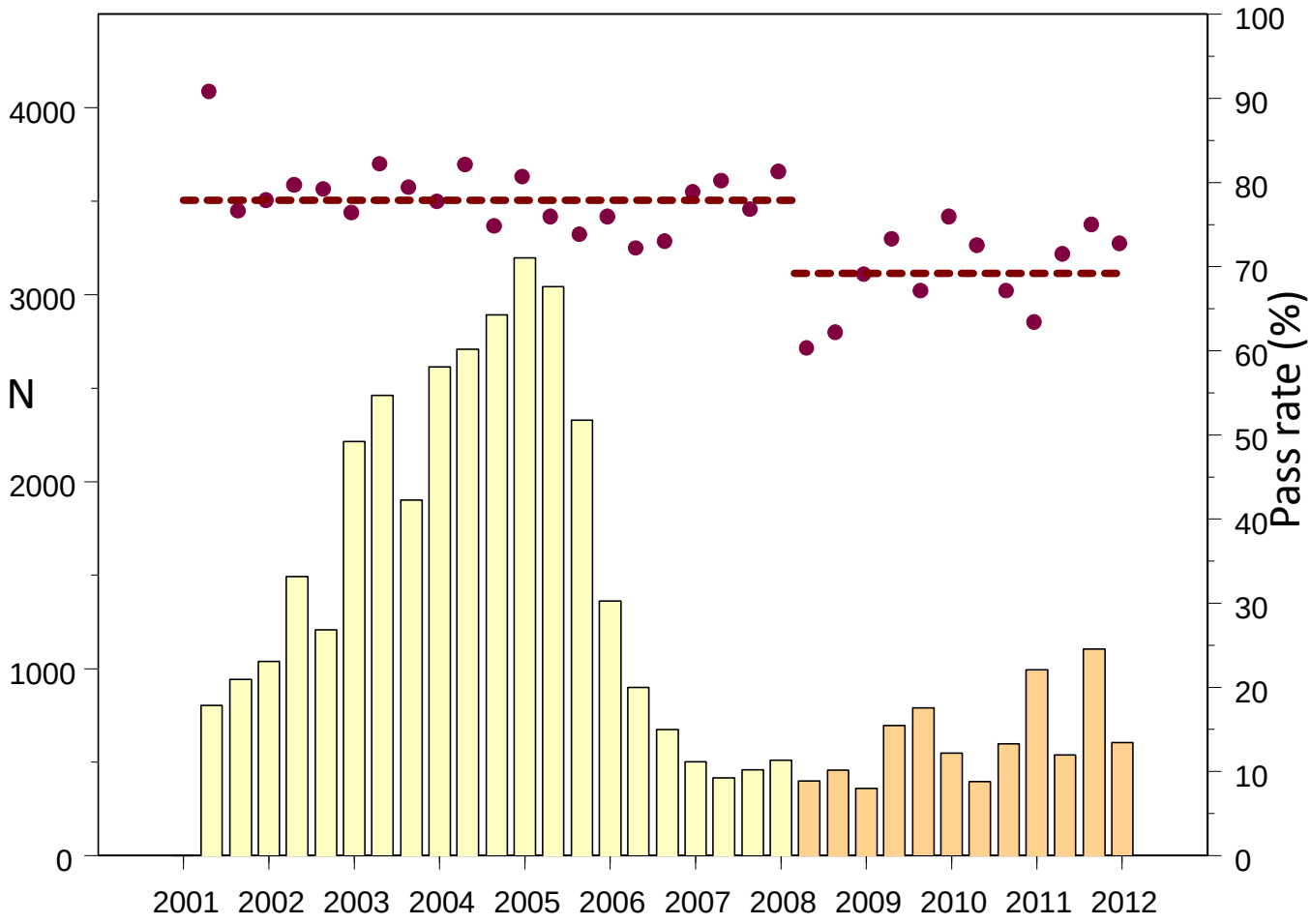
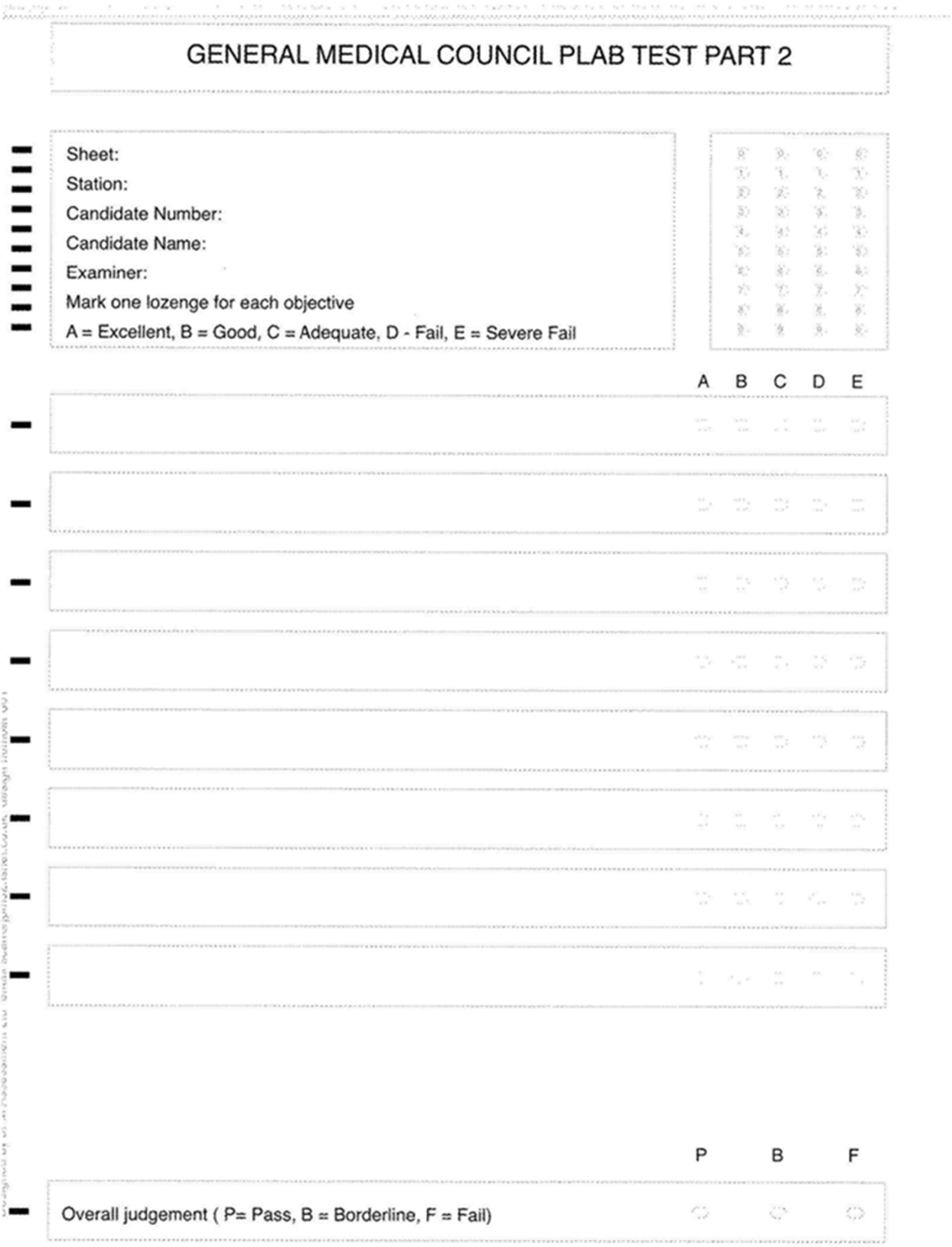


Figure 2: Mark sheet for a station at PLAB Part 2.



GENERAL MEDICAL COUNCIL PLAB TEST PART 2

Sheet:
Station:
Candidate Number:
Candidate Name:
Examiner:
Mark one lozenge for each objective
A = Excellent, B = Good, C = Adequate, D - Fail, E = Severe Fail

A B C D E

P B F

Overall judgement (P= Pass, B = Borderline, F = Fail)

Figure 3: Minimum passing score for stations (May 2008 to present) plotted against maximum passing score. Scores are on the standard scale of 0 to 4 for E to A. The pass mark on the original marking scheme was set at 1.55 (i.e. just above the boundary between 1 ('Fail') and 2 ('Adequate')), and is shown by the red lines.

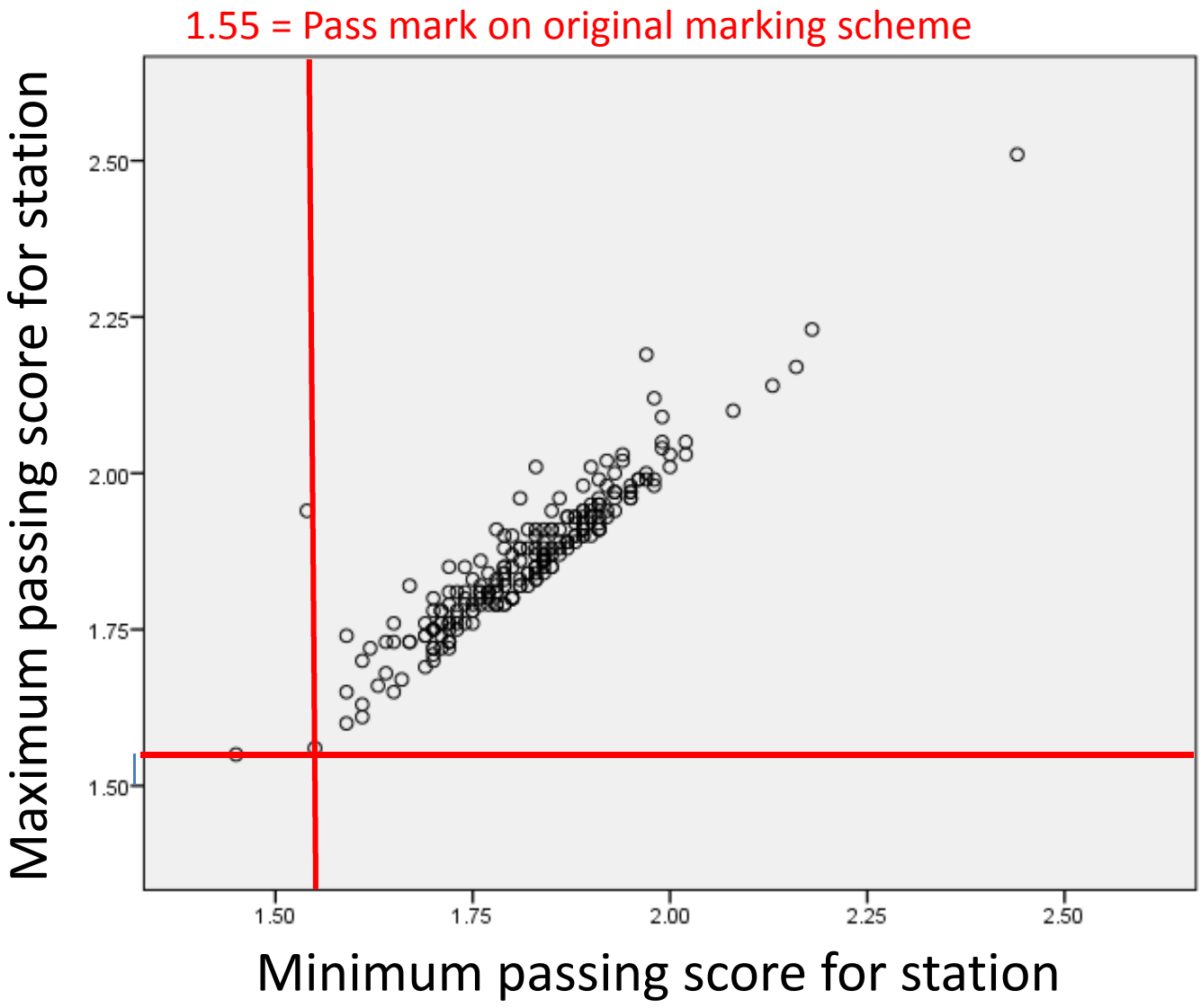


Figure 4: FACETS 'yardstick' for examinations post-Jan2007, using weighted station grade (A to E) calculated for each station (Note that for stations post-May 2008 this has been recalculated). See text for further details.

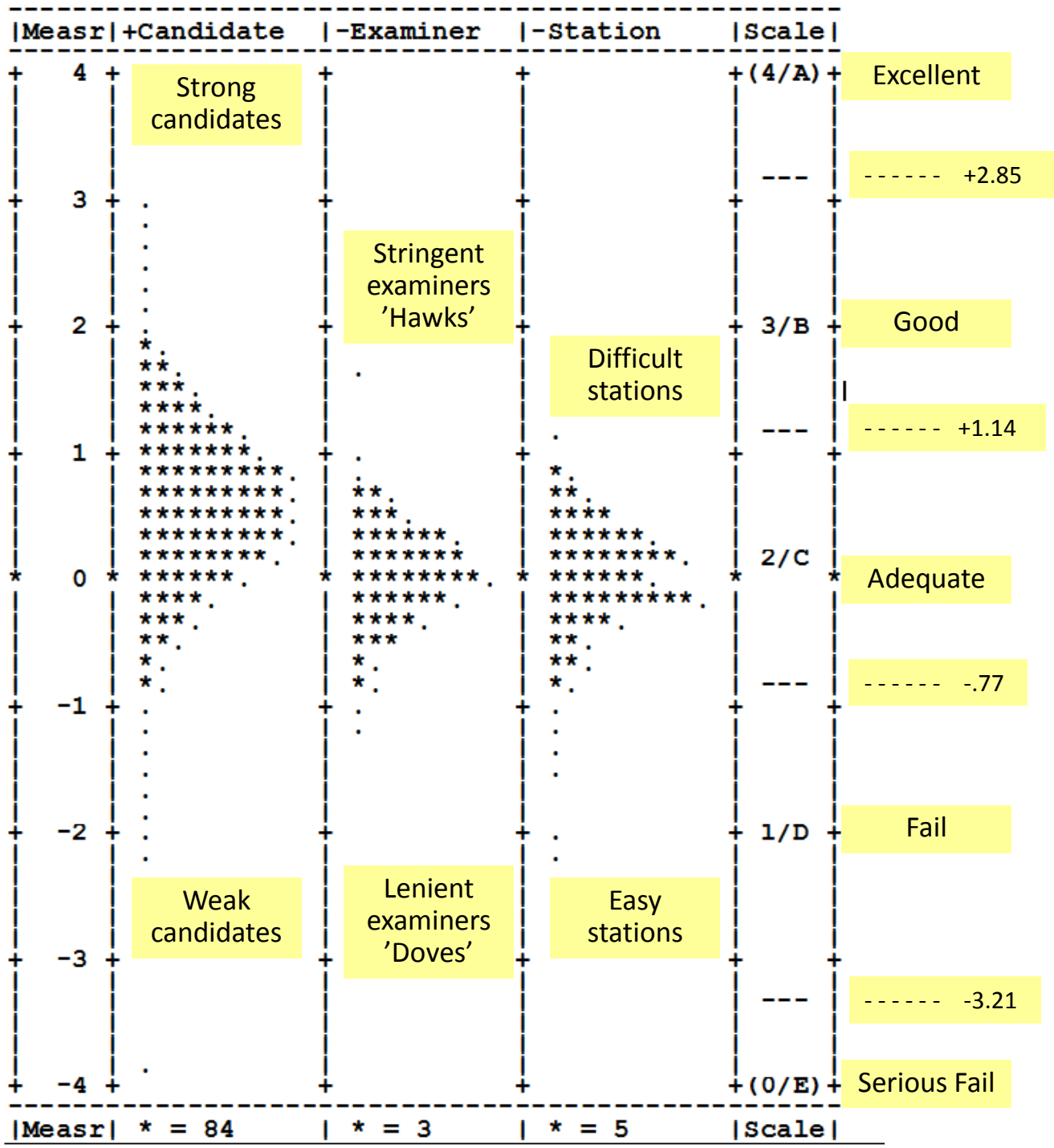


Figure 5: FACETS 'yardstick' for examinations post-Jan2007. Calculations use grades (A to E) on individual objectives within each station. See text for further details. Red arrows point to six objectives for station 2626, shown in table 6 and discussed in the main text.