Report of Analyses of PLAB Part 2 to the General Medical Council

David B Swanson, PhD – 1 March 2018

Executive Summary

Overview of Analyses

The analyses in this report were performed during and after a visit to the Manchester offices of the General Medical Council from 2 to 6 October 2017. While I was in the GMC offices, I had the opportunity to observe a portion of a PLAB administration, to discuss PLAB 2 with the Chief Examiner for that administration, to interact with some of the GMC staff involved in PLAB 2, and to discuss preliminary analysis results with Richard Hankins from the GMC and Drs Richard Fuller and Matthew Homer from the University of Leeds.

The analyses are based on a (fully anonymized) dataset that included performance information for the total test and individual scored stations for 2746 candidates who sat for the new PLAB 2 examination between 7 September 2016 and 14 September 2017 inclusive.[1] Scores for a total of 49131 stations were included, along with information about candidates, examiners, role players, and stations.

The following types of analyses were performed:

- Basic descriptive analyses including candidate counts, total test pass rates and mean scores, station pass rates and mean scores
- Generalizability analyses of the reproducibility of scores investigating the extent to which similar total scores would be received if candidates were retested with different samples of stations, examiners, and role players; indices of the reproducibility of scores with smaller and larger numbers of stations were projected statistically from these results
- Exploratory analyses of the consistency of station pass/fail standards across test dates and the reproducibility of domain scores in data gathering, clinical management, and Interpersonal skills to guide potential future improvements in scoring methods

Some of these analyses replicate work done in 2014 when the previous PLAB format was in place, allowing investigation of whether or not the anticipated benefits motivating the shift to the new PLAB format have been realized.

Conclusions and Recommendations – Test Length and the Reproducibility of Test Results

1. Changes made to the PLAB 2 examination format have resulted in a substantial improvement in the reproducibility of scores from 0.61 (old format) to 0.77 (new format).[2] These values can be interpreted as the expected correlation between the results of similarly structured (randomly parallel) exams using different stations, examiners, and role players. The standard error of measurement, another index of the reproducibility (precision) of scores, also improved by 17%. These results indicate that the changes made to the PLAB format have substantially improved the reproducibility of scores and pass/fail decisions.

[1] The dataset included test results for candidates from 82 test administration dates. The vast majority (89%) of the administrations included 18 stations per candidate, with 17 stations for all other candidates in the dataset. A few PLAB 2 administrations during this period had less than 17 stations; these were excluded from the dataset.

[2] While higher levels of reproducibility are desirable, the new PLAB yields scores at least as reproducible as other national examinations of clinical skills with which I am familiar.

2. There has also been an improvement in the reproducibility of "scores" representing the number of stations passed; this is important because candidates with passing total overall scores are also required to pass a minimum of 11 of 18 stations. However, these scores are still less reproducible than is desirable for making a high-stakes decision regarding entry into postgraduate medical education in the UK. At the same time, the reproducibility of these scores does seem sufficiently good for the limited purpose of serving as a secondary pass/fail hurdle. (See recommendation 4 for related discussion).

Conclusions and Recommendations – Standard Setting and Scoring

3. The use of the borderline regression method in setting standards is partially responsible for the improvement in the reproducibility of scores because the resulting pass/fail standards better reflect the day-to-day variation in the difficulty of the stations seen and the stringency of the examiners marking them. At the same time, there has been an unexplained increase in pass rates, and an adjustment to the standard setting procedure for individual test dates seems warranted. Results of generalizability analyses provide a better estimate of the standard error of measurement that may be useful in setting the pass/fail standard for individual administrations.
4. It appears possible to use borderline regression procedures to adjust the minimum number of stations required to pass PLAB 2 on a given test date to reflect the difficulty of the sample of stations on that date. This should improve the comparability of standards across test dates and, as a consequence, the reproducibility of pass/fail decisions. This may be difficult to explain to candidates, however, so careful exploration of the utility of such a procedure prior to implementation is warranted.
5. The use of global scoring for three domains (data gathering, clinical management, interpersonal skills) does not provide markers with much specific guidance for scoring, which can result in widely varying scores for the same stations on different days of test administration. Exploring the use of station development and marking methods[3] based on the "key features" seems desirable to improve station design and comparability in marking; this may also improve the reproducibility of scores and pass/fail decisions.
6. Investigate changes to examiner and role player training[4] to promote greater consistency in marking and portrayal across test dates.

Conclusions and Recommendations – Scheduling and Test Administration

7. For borderline regression methods to work well, it is desirable to have some strong and weak candidates sit for PLAB 2 on each test administration date; this should improve the precision of the estimated pass/fail standard for each station. Further, each examiner would, ideally, see a strong and weak candidate early in the day to aid in calibration in assignment of marks. The GMC possesses information (eg, IELTS scores, country of medical school, scores on PLAB 1, number of previous PLAB 2 attempts) useful in predicting how likely each candidate is to pass PLAB 2 before the test administration. These predictions could be used in assigning candidates to morning/afternoon sessions and to starting stations in morning circuits.[5]

---

[3] Drawing on the key features work of Bordage and Page, the Australian Medical Council and the Medical Council of Canada are doing interesting work in this area.

[4] In particular, it may be useful to consider the use of video in examiner orientation, training, and calibration on test days. With the new video capabilities, it appears possible to identify station-specific videos that illustrate typical and extremely good/poor performance on individual stations. Two or three videos of their assigned station could be viewed by each examiner and role player before test administration begins, assisting in examiner calibration and promoting consistency in role player portrayal.

[5] In principle, this information could also be used in assigning candidates to test administration dates, but it sounded like this would require a major modification in current scheduling procedures.

Miscellaneous Recommendations Unrelated to Analysis Results

8. Add an assessment of skills in "spoken medical English" to each station, with those skills rated by both examiners and role players. Report an overall score for spoken medical English and require candidates to achieve a passing score on medical English, as well as the total score, on the same test administration.[6]
9. Investigate the initial positions in UK postgraduate training obtained by candidates passing PLAB 2, and consider the implications of this information for the PLAB 2 blueprint and station formats used.
10. In reports to other organizations requesting information about the performance of candidates on PLAB (eg, the Australian Medical Council), provide a full "transcript" of PLAB results, including the date of each PLAB attempt and whether a passing or failing result was obtained.[7]

Thoughts on the Way Forward

The most important result reported here is that the changes introduced in the new PLAB 2 format have led to substantial improvements in the reproducibility of scores and pass/fail decisions – the General Medical Council is to be congratulated.

The most important recommendations in the shorter term are related to standard setting (Recommendations 3 and 4) to rectify the unexplained increase in pass rates so that unqualified candidates are less likely to pass the new PLAB 2. There are a number of (complementary) mechanisms by which this could be accomplished:

- Increase the number of standard errors of measurement by which the borderline regression pass/fail standard is raised[8]
- Use the larger standard error of measurement referenced in Recommendation 3 in making the above adjustment to the borderline regression pass/fail standard
- Increase the number of stations on which a passing score is required, perhaps exploring the use of borderline regression methods to do so as outlined in Recommendation 4

Also in the shorter term, exploring changes in scheduling and test administration (Recommendation 7) seems worthwhile, in part because it should be relatively straightforward to implement.

In the longer term, Recommendations 5, 6, and 9 seem likely to have the most benefit for both the reproducibility of pass/fail decisions and the validity of those decisions. They may also be useful to consider in guiding the development of the clinical skills component of the Medical Licensing Assessment.

[6] PLAB staff indicated that some candidates with high scores on IELTS (and those exempted from IELTS requirement) can be difficult to understand during PLAB 2 encounters, probably reflecting, at least in part, differences between the generic English-language skills measured with IELTS and those required to function effectively with patients and colleagues in a clinical environment. This may continue to be true even with a requirement for higher IELTS scores.

[7] During my 2014 visit, I got the impression that the number of PLAB attempts was sometimes considered by the GMC in making decisions, even if passing scores were achieved on both PLAB components. Regardless, it seems desirable for GMC to pass this information on to other institutions (eg, the Australian Medical Council) requesting information about PLAB results, perhaps in the form of a "transcript" summarizing ***all*** of a candidate's PLAB 1 and 2 attempts. I did not see a copy of the information sent to other institutions during this trip, and I apologize if this has already been done.

[8] My understanding is that such an increase is planned for early in 2018.

Overview of Dataset and Analyses Performed

Analyses were performed in the Manchester offices of the General Medical Council during the week of 2-6 October 2017. The (fully anonymized) dataset used in analyses included information about candidates who took the new PLAB 2 examination between 7 September 2016 and 14 September 2017 inclusive. Because some analyses used raw scores, administrations with less than 17 stations were dropped from the dataset. The resulting dataset included 2746 candidates and 49131 stations, with 89% of candidates completing 18 stations and 11% of candidates completing 17 stations. A variety of demographic and other information about candidates, examiners, role players, and stations was available.

Four sets of scores were calculated for use in analysis.

Station Deviation Scores. For purposes of analysis, raw scores for individual stations on each test date were transformed onto a percentage-of-possible-points scale by dividing each candidate's raw score (see domain scores below) by the maximum possible score on the station and multiplying by 100. The same transformation was applied to station pass marks determined with the borderline regression method in use for the new PLAB 2 format. The difference between the two was then calculated. This "station deviation score" expresses, in percentage terms, how much a candidate's performance on the station was above/below the station pass mark, placing scores on each station on a roughly comparable scale (which is desirable analytically). Positive values indicate the amount by which a candidate's performance exceeded the station pass mark; negative values indicate the amount by which a candidate's performance was below the station pass mark.

Total Test Deviation Scores. The "total test deviation score" was calculated by computing the mean of the station deviation scores. Positive values indicate that, across stations, a candidate's performance is above the overall borderline regression pass/fail standard for that test date, and negative numbers indicate that performance was below the pass/fail standard for that date. This places scores from different test dates onto a (roughly) common scale, regardless of whether 17 or 18 stations were used on the form.

Station Pass/Fail Scores and the Percentage of Stations Passed. In addition to requiring an overall passing score, PLAB 2 rules in place for most of the study period[9] required that candidates pass a minimum of 11 stations on an 18-station test form in addition to achieving a passing total test score. To investigate the reproducibility of "scores" indicating the number of stations passed, a station pass/fail score was calculated for each station and candidate by assigning a value of 100 if a candidate's performance on the station was equal to or greater than the borderline regression pass mark for the station and a value of 0 if a candidate's performance on the station was less than the station pass mark. The mean of these station pass/fail scores is the percentage of stations passed; the pass/fail standard on this scale is approximately 61% (100% X 11/18).

Domain scores in Data Gathering, Clinical Management, and Interpersonal Skills. These are the three (global) scores assigned by examiners at each station. Operationally, each domain is marked on a 0 to 4 scale, and the scores are summed yielding a station (raw) score varying between 0 and 12. For purposes of analysis, each domain score was divided by 4 and multiplied by 100% to place the scores on a scale in which scores are expressed in terms of the percentage of possible points for the associated domain.

[9] This requirement was introduced early in 2017; candidates taking the new PLAB 2 earlier than that were not subject to a secondary pass/fail hurdle requiring a passing score on a specified minimum number of stations.

Three general categories of analyses were performed:

- **Descriptive statistical analyses** of candidate counts, total test pass rates and mean scores, station pass rates and mean scores were run by year of test administration and PLAB 2 format and for 1st-time and repeat candidates.
- **Generalizability analyses** investigating the reproducibility of scores used in making pass/fail decisions and the extent to which similar scores would be received if candidates were retested with different (randomly parallel) samples of stations, examiners, and role players. These were done separately for station scores and for the number (percentage) of stations passed, reflecting PLAB's use of distinct pass/fail "hurdles" for each of these. The impact of increasing the number of stations included on test forms were also projected statistically for both station scores and the percentage of stations passed.
- **Exploratory analyses investigating** 1) the consistency of station pass/fail standards across test administration dates and 2) the reproducibility of domain scores in Data gathering skills, Clinical management skills, and interpersonal skills

Results – Descriptive Statistical Analyses

Table 1 provides overall pass rates by PLAB 2 examination format and year of test administration (cohort) for candidates on their first attempt, second attempt, third attempt, and more than three attempts. Comparing the rows for the Total group for the Old Format and the New Format (in red), it is clear that there has been an increase of roughly 10% in the pass rate, both overall and for groups defined by the number of attempts. This is also true in a comparison of the 2016 cohort taking the old format and those taking the new format, though the magnitude of the increase is smaller overall, varying somewhat in size by number of attempts. The reason for the increase in pass rate is unclear, with the shift in standard setting procedures to the borderline regression method providing the most likely explanation.[10]

		Number of Attempts at PLAB2 (regardless of exam format)									
		1st Attempt		2nd Attempt		3rd Attempt		> 3 Attempts		Total	
PLAB 2 Format	Cohort	Count	% Pass	Count	% Pass	Count	% Pass	Count	% Pass	Count	% Pass
Old Format	2011	1905	70.3	549	69.6	136	64.7	47	53.2	2637	69.6
	2012	1241	71.2	343	63.6	93	55.9	58	48.3	1735	68.1
	2013	1108	65.4	425	64.0	167	55.7	79	45.6	1779	63.3
	2014	1221	69.4	315	55.9	135	61.5	58	34.5	1729	65.1
	2015	1533	71.4	467	60.2	151	62.9	105	52.4	2256	67.6
	2016	1303	72.1	324	75.3	104	66.3	67	53.7	1798	71.7
Old Format	**2011-16**	**8311**	**70.1**	**2423**	**64.9**	**786**	**61.1**	**414**	**48.3**	**11934**	**67.7**
New Format	**2016-17***	**2248**	**81.6**	**373**	**75.3**	**91**	**69.2**	**53**	**56.6**	**2765**	**79.9**

**Reflects all administrations of the new PLAB 2 examination format through 14 September 2017*

Table 1: Candidate Counts and Pass Rates by PLAB 2 Examination Format, Number of Attempts, and Cohort

Over the roughly 12 months included in the study dataset, there were a total of 82 test dates with a different test form used on each date of test administration. The upper panel of Figure 1 provides a plot of pass rates by test administration date; the lower panel provides similar information for first-time and repeat candidates. The upper panel shows some day-to-day variation in pass rates from a low of 48% to a high of 100%. Some of this variation reflects the relatively small number of candidates testing each day: the average number of

[10] At the time this report was prepared, no demographic shifts (eg, associated with Brexit) in the 2016-17 candidate population have been identified as a more likely explanation, though this work is ongoing.

candidates tested per day is 33, ranging from a low of 17 to a high of 36, with 30 or more tested on the vast majority of test dates. In part, the variation may also reflect shifts in candidate characteristics, though the upper panel does not suggest much "seasonality." First-taker pass rates in the lower panel are somewhat less variable, ranging from 57% to 100%.[11] Compared with the analogous graphs in the 2014 report, day-to-day variability in pass rates is much smaller, despite the smaller number of candidates tested daily with the new PLAB 2 format. This almost certainly reflects the introduction of borderline regression methods for standard setting, which results in greater comparability in pass/fail standards across test dates by taking day-to-day variation in the difficulty of test forms into account. The latter was an important part of the rationale for shifting to use of borderline regression in setting pass/fail standards.

Figure 2 depicts the variation in mean total test deviation scores across the 82 test administration dates. These percentage scores are scaled so that a score of zero is assigned to the pass/fail standard for the associated test date. Some day-to-day variation is again evident, with most mean scores falling between 7% and 16%, as one might expect given the overall pass rate of 79.9% in Table 1. The day-to-day variation in mean scores depicted in Figure 2 is somewhat smaller than that seen in the analogous figure in the 2014 report, again reflecting greater day-to-day comparability in pass/fail standards achieved through use of the borderline regression group for the new PLAB 2 format. Standard deviations are typically in the range of 8 to 10 points, values consistent with the results of generalizability analyses presented later in the report.

[11] The sizable fluctuation in pass rates for the Repeater group simply reflects the small numbers of repeat candidates: on a number of dates, fewer than five were tested.

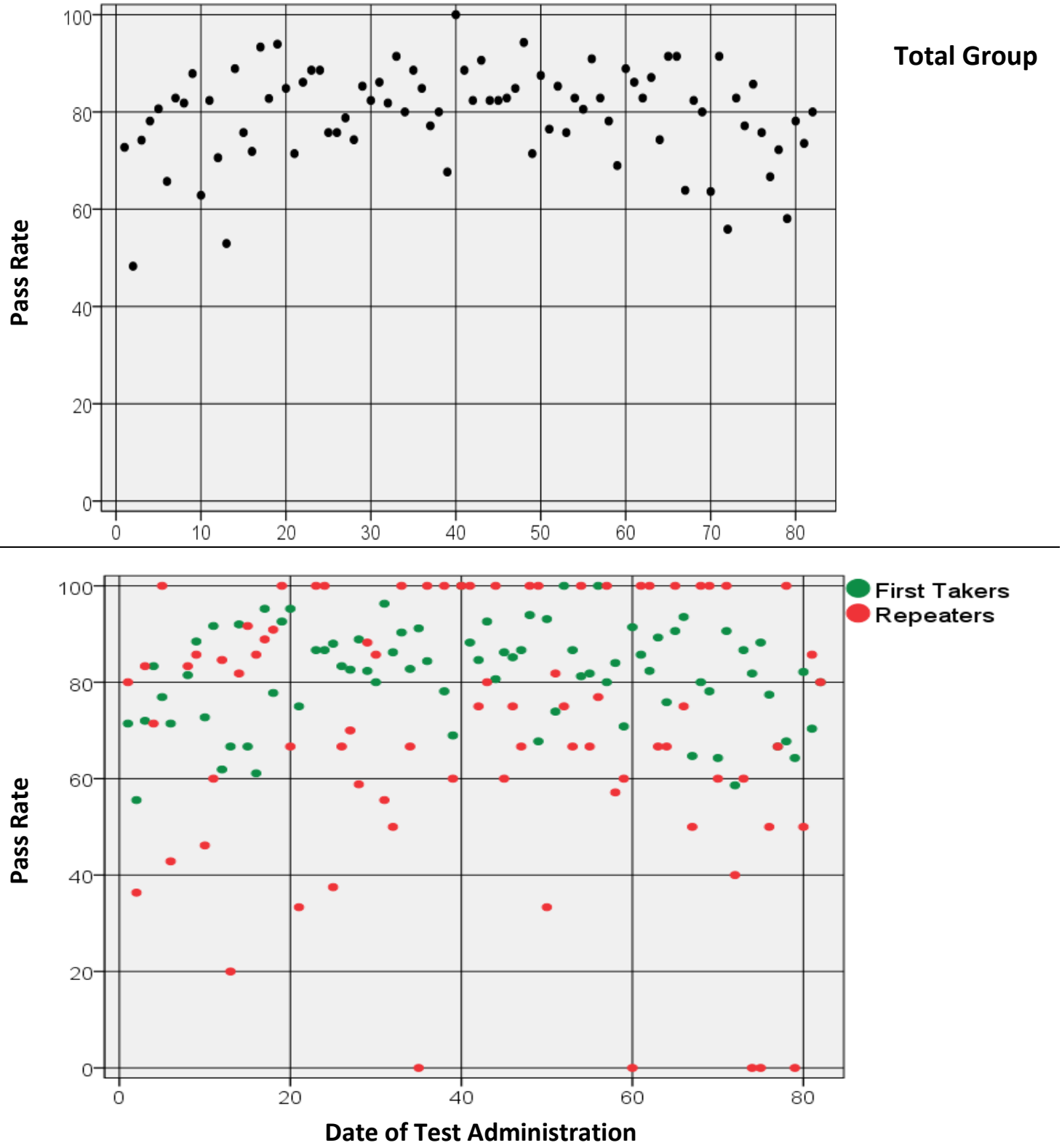

Figure 1: Pass Rates for the New PLAB 2 Format by Test Date

The upper panel shows pass rates for the Total Group of the each of the 82 test administration dates; the lower panel breaks the Total Group down into First-taker and Repeater groups on each date

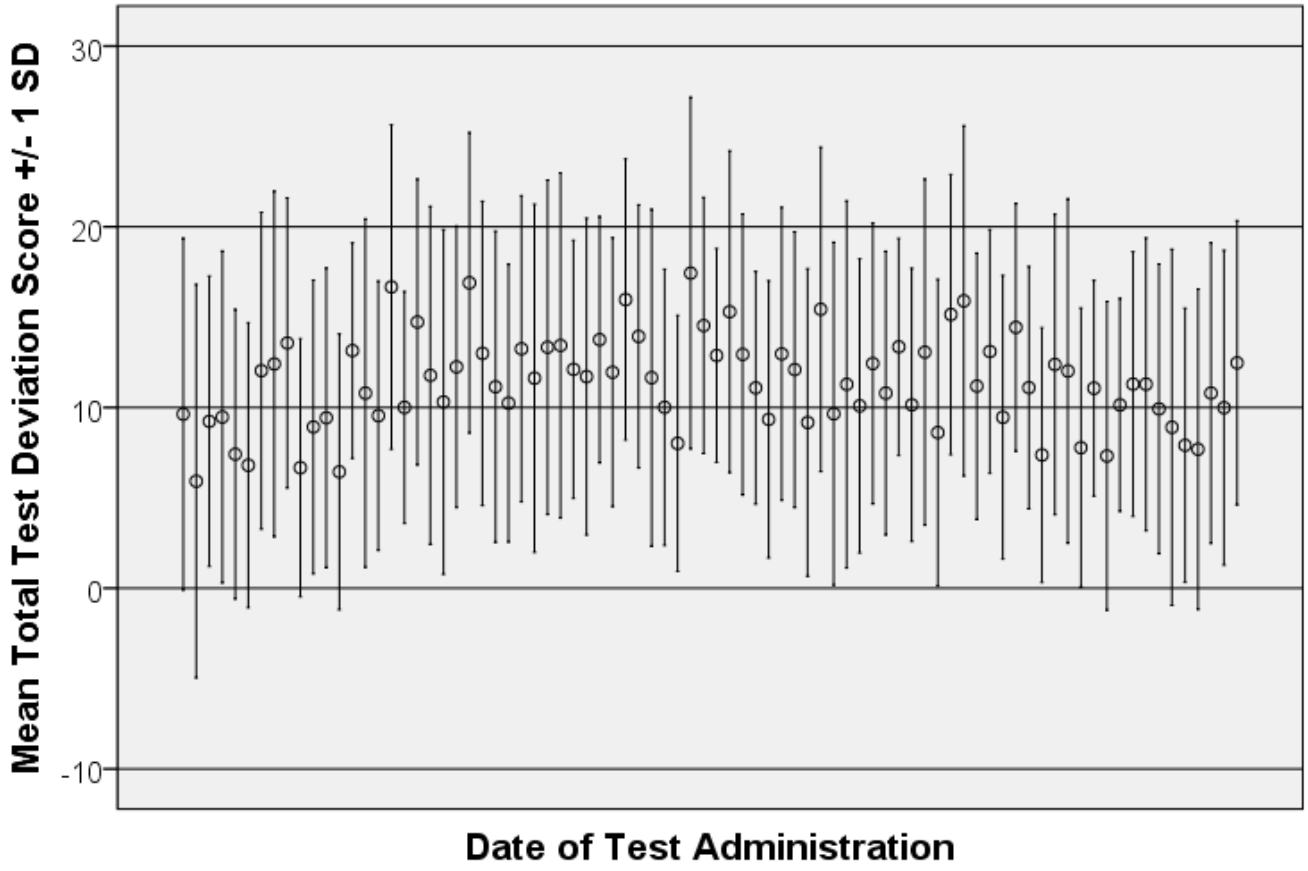

Figure 2: Mean Total Test Deviation Scores (+/- 1 SD) by Date of Test Administration

Results – Generalizability Analyses

The procedure summarized below was followed to estimate the reproducibility of PLAB 2 scores, including both generalizability coefficients and standard errors of measurement.

1. For each candidate and each station on each day of testing, the difference between a candidate's station score (on a percentage-of-possible-points scale) and the borderline regression pass/fail standard (on the same scale) was calculated. This station deviation score indicates how much a candidate's performance was above or below the pass/fail standard.[12]
2. A separate one-way random-effects analysis of variance (ANOVA) was performed (using the SPSS VARCOMP procedure) for each of the 82 days of testing included in the dataset, providing estimates of variance components for Candidates and Scores nested within Candidates [Scores(Candidates)]. The former is an estimate of "true score" variance (the variation in scores that would be seen if no measurement error were present), and the latter is an estimate of error variance.
3. Estimated variance components were averaged across the 82 days of testing to obtain a single, more precise (pooled) estimate of each variance component.
4. Pooled estimates of variance components were used to calculate generalizability coefficients and standard errors of measurement (SEMs) for the actual test length of 18 stations.
5. Pooled estimates were also used to project the effect of increasing the test length on both generalizability coefficients and SEMs.

A similar procedure was followed to estimate the reproducibility of station pass/fail scores representing the percentage of stations passed, reflecting the fact that candidates must also pass a minimum of 11 out of 18 stations in order to pass the exam.

Table 2 summarizes the results for both total test deviation scores and the percentage of stations passed. The top section of the table provides pooled estimates of the variance components. The variance component for Candidates provides information about the true variation in scores if no measurement error were present. The square root of this value (7.36) can be interpreted as the standard deviation that would be observed if the test were very, very long. Most (95%) scores would fall within +/- 2 standard deviations (a range of almost 30 points on the percentage-of-possible-points scale) of the mean for all candidates. Similarly, the value of 13.35 (roughly 2.4 stations) indicates the true SD for the percentage of stations passed. The variance component for Scores(Candidates) provides information about the amount of measurement error that is present; the square root of this value indicates the SEM for a test consisting of only a single station; this is a building block used in projecting the SEMs for other test lengths.

The lower section of the table provides projected generalizability coefficients and SEMs for a variety of test lengths for both total test deviation scores and the percentage of stations passed. At the actual test length of 18 stations, the estimated generalizability coefficient for total test deviation scores is 0.77 (shown in red) for

[12] If the mean of the station deviation scores for a candidate is greater than 0, the candidate exceeded the total test pass/fail standard for the test form used on that day of test administration; if the mean of the station deviation scores is less than 0, the candidate failed the exam. To the extent that differences in station difficulty and examiner stringency are accounted for by differences in station pass/fail standards set using borderline regression, this calculation places scores for test forms from different dates and different sets of examiners and role player on a common scale. The pass/fail standard for the total test is 0 on this scale.

total test deviation scores.[13] This indicates that if candidates repeated PLAB 2 with a different sample of 18 stations and examiners, the expected correlation between scores is 0.77. As discussed below, this is substantially improved over the value observed for the old PLAB 2 exam at a test length of 14 stations.

The SEM of 4.03% for deviation scores at a test length of 18 stations provides an index of the reproducibility of scores that is interpretable on the same scale as the scores. The SEM can be used to form confidence intervals around a candidate's score. For example, a candidate receiving a total test deviation score of -3 (failing by 3%), would have a 95% confidence interval ranging from -11.06 (-3 minus 2 X 4.03) to +5.06 (-3 plus 2 X 4.03). This is a fairly broad interval (from failing badly to slightly above passing), indicating that further improvements in reproducibility are desirable, but, as discussed below, it is substantially better than the SEM for the old PLAB 2 format at a test length of 14 stations.

Currently, the pass/fail standard set using the borderline regression method is adjusted upward by 1 SEM. The latter is estimated for each test administration date using a formula involving coefficient alpha. Because coefficient alpha provides an overestimate of the reproducibility of scores, the SEM used in adjusting the pass/fail standard upward is too small. Translating the value of 4.03% from the table onto the 216-point scale (18 stations X 12 points/station) used operationally for making pass/fail decisions can be accomplished by simply multiplying 4.03% by 2.16 (216/100), which yields a value of 8.07 – ***this is an appropriate value to use in adjusting the pass/fail standard upward***. [14] As expected, this is somewhat higher than the estimates based on coefficient alpha that are computed after each day of test administration.

Indices of reproducibility for the percentage of stations passed are substantially worse. At the actual test length of 18 stations, the projected generalizability coefficient is only 0.64, a value well below the 0.80 recommended for making high-stakes decisions like those associated with PLAB 2. Similarly, the SEM of 10.08% (roughly equivalent to two stations) is quite large, indicating that the percentage of stations that a candidate would pass is likely to vary substantially from one PLAB 2 administration to the next.

[13] This value is slightly lower than typical alpha coefficients for PLAB 2 test administrations on individual test dates. This is because alpha coefficients computed for test dates individually do not reflect measurement error due to differences in station difficulty and examiner stringency.

[14] Operationally, it would be straightforward to calculate a form-specific SEM using generalizability theory by doing a random effects oneway analysis of variance (station deviation scores nested in candidates) for any given date of test administration, though the pooled estimate provided here is probably more accurate.

	Total Test Deviation Scores (Percentage-of-Possible-Points Scale)		Percentage of Stations Passed	
	Candidates	Scores(Candidates)	Candidates	Scores(Candidates)
Variance Component (VC)	54.11	292.59	178.23	1829.81
Square Root of VC (% scale)	7.36	17.11	13.35	42.78

Test Length (# of scored stations)	Generalizability Coefficient	SEM (% scale)	Generalizability Coefficient	SEM (% scale)
14	0.72	4.57	0.58	11.43
16	0.75	4.28	0.61	10.69
18	**0.77**	**4.03**	**0.64**	**10.08**
20	0.79	3.82	0.66	9.57
22	0.80	3.65	0.68	9.12
24	0.82	3.49	0.70	8.73
26	0.83	3.35	0.72	8.39
28	0.84	3.23	0.73	8.08
30	0.85	3.12	0.75	7.81
32	0.86	3.02	0.76	7.56
34	0.86	2.93	0.77	7.34
36	0.87	2.85	0.78	7.13

Table 2: Results of Generalizability Analyses for the New PLAB 2 Format

The upper portion of the table provides variance component estimates from generalizability analyses. Those variance components are the basis for the generalizability coefficients and standard errors of measurement (SEM) presented as a function of test length in the lower panel. The row in red represents the new PLAB 2 test length of 18 stations.

To provide a basis for comparing the new and old PLAB 2 formats, Table 3, reproduced from the 2014 report, provides information analogous to that in Table 2.

	Total Test Deviation Scores (Percentage-of-Possible-Points Scale)		Percentage of Stations Passed	
	Candidates	Scores(Candidates)	Candidates	Scores(Candidates)
Variance Component (VC)	36.32	329.14	127.91	1964.18
Square Root of VC (% scale)	6.03	18.14	11.31	44.32

Test Length (# of scored stations)	Generalizability Coefficient	SEM (% scale)	Generalizability Coefficient	SEM (% scale)
14	**0.61**	**4.85**	**0.48**	**11.84**
16	0.64	4.54	0.51	11.08
18	0.67	4.28	0.54	10.45
20	0.69	4.06	0.57	9.91
22	0.71	3.87	0.59	9.45
24	0.73	3.70	0.61	9.05
26	0.74	3.56	0.63	8.69
28	0.76	3.43	0.65	8.38
30	0.77	3.31	0.66	8.09
32	0.78	3.21	0.68	7.83
34	0.79	3.11	0.69	7.60
36	0.80	3.02	0.70	7.39

Table 3: Results of Generalizability Analyses for the OLD PLAB 2 Format

The upper portion of the table provides estimated variance component from the 2014 generalizability analyses of the old PLAB 2 format. Those variance components were used to calculate the generalizability coefficients and standard errors of measurement (SEM) presented as a function of test length in the lower panel. The row in red represents the old PLAB 2 test length of 14 stations.

To investigate whether the SEM on the new PLAB 2 varies across the score scale, candidate-specific SEMs were estimated from station deviation scores by calculating the standard deviation of those scores for each candidate and dividing by the square root of the test length, producing a standard error of the mean score for each candidate – this is conceptually the same as an SEM. Results are plotted in the left panel of Figure 3. Each point in the plot represents a candidate, with total test deviation scores on the X-axis and SEMs on the Y-axis. Blue points indicate passing candidates, and red points indicate failing candidates. Points to the left of 0 are uniformly red because the pass/fail standard on this scale is equal to 0. There are also some red points to the right of 0; these candidates did not satisfy the requirement of passing at least 11 stations. The curved black line approximates the average SEM across the score scale. Although there are many outliers reflecting candidates with highly variable performance across stations, the average SEM is approximately 4% across the score scale, which is consistent with the (red) entry in Table 2 for an 18-station exam.

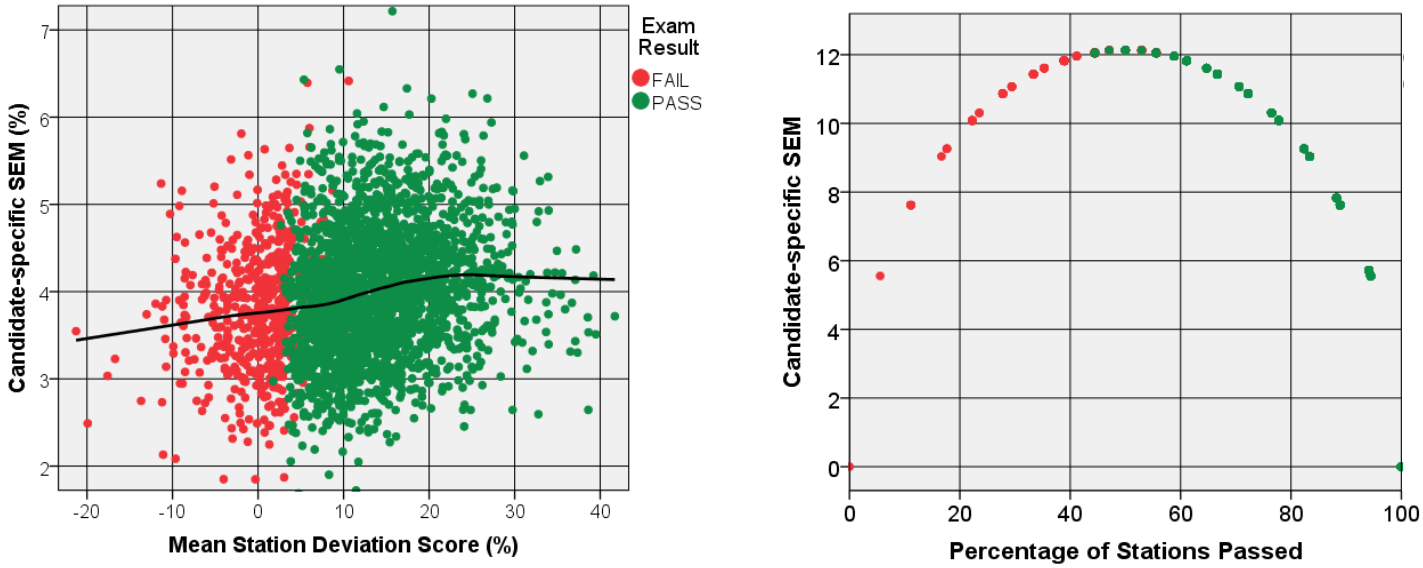

Figure 3: Candidate-specific SEMs for Total Test (Mean) Station Deviation Scores (left panel) and Percentage of Stations Passed (right panel), both at a Test Length of 18 Stations

The pass/fail standard is at 0 for the left panel and at 61% for the right panel. Some candidates in the right panel are indicated to have an overall pass on the examination because the secondary pass/fail hurdle for percentage of stations passed was not used for early administrations using the new PLAB 2 format.

A similar graph is shown in the right panel for the SEM as a function of the percentage of stations passed. The shape of the graph is quite different – the SEM is directly related to the percentage (proportion) of stations passed because the pass/fail standard for each test form is set independent of the difficulty of passing individual stations and overall form difficulty. As a consequence, the SEM on this "scale" is near its peak at the passing standard of 61% (11 of 18 stations). The value is close to 12% or about two stations (0.12 x 18), and a 95% confidence interval for scores near this value is +/- 4 stations, indicating that changes in pass/fail results on this scale are not very reproducible.

It would certainly be desirable for pass/fail results on this scale to be more reproducible from a candidate's perspective, but, from the perspective of "protection of the public," its use as a secondary pass/fail hurdle is probably reasonable. If the GMC is interested, It should be possible to use borderline regression procedures to adjust the minimum number of stations required to pass PLAB 2 on a given test date to reflect the difficulty of passing the sample of stations on that date. This should improve the comparability of standards across test dates and, as a consequence, the reproducibility of pass/fail decisions. This may be difficult to explain to candidates, however, so exploration of the utility of such a procedure prior to implementation is warranted.

Results – Exploratory Analyses

Consistency of Station Pass/Fail Standards across Test Administration Dates

The borderline regression standard setting procedure determines a pass/fail standard for individual stations by predicting (regressing) the sum of the domain scores in data gathering, clinical management and interpersonal skills from each examiner's global judgments of station performance. The pass/fail standard for each station is set at the point on the scale that corresponds to a global judgment of "borderline."

If examiners behave consistently from one test administration date to the next, one would expect that the pass/fail standards for a station would be similar across test dates.[15] To investigate the consistency of station pass/fail standards across test dates, stations used on at least 10 test dates were identified, the station pass/fail standard (on a percentage of possible points scale) for each test date was determined, and boxplots were generated to summarize the distribution of pass/fail standards for individual stations.

These are shown in Figure 4. Each vertical box plot shows the results for an individual station. The median of the standards for a station is indicated by the horizontal line near the center of each box (rectangle); the box delimits the interquartile range (the 25th to the 75th percentiles); "whiskers" on either side of the box represent the ranges for the bottom 25% and the top 25% of the data values with outliers excluded; and outliers are plotted as circles and asterisks (extreme outliers).

While the interquartile ranges (the boxes) for most stations fall between 40% and 55%, there are some interquartile ranges outside these values, and outliers frequently extend from 35% to 65%, with a few even more extreme values observed. Because of the small sample of marks (typically 30-35) available for estimating individual borderline regression equations, variation in the estimates of the pass/fail standards for individual stations are essentially inevitable because one or two outlier judgments can have a sizable impact on the estimates of regression slopes.[16] To the extent that the borderline regression procedure takes into account differences in station difficulty and examiner stringency in determining the station pass/fail points, the overall impact on day-to-day comparability of standards may not be great. At the same time, the variation in pass/fail standards depicted in Figure 4 also reflects differences in the basis on which examiners are assigning marks. Analyses of the reproducibility of individual domain scores in the next subsection are intended to provide additional insight into the causes of the variation shown in Figure 4.

[15] While not absolutely required, since the borderline regression procedure is designed to take into account variation in examiner stringency, the amount of variation provides an indication of whether examiners mark stations similarly.

[16] The GMC may wish to consider some measures to improve the precision of these estimates. There are statistical methods available (eg, parameterizing the regression equations to constrain slopes to be equal for a given station or across stations) that could be investigated to lessen the impact of outliers. In addition, it may be possible to schedule candidates so that each examiner sees a wide range of performance early in the morning, which should improve the consistency of their judgments over the course of the day.

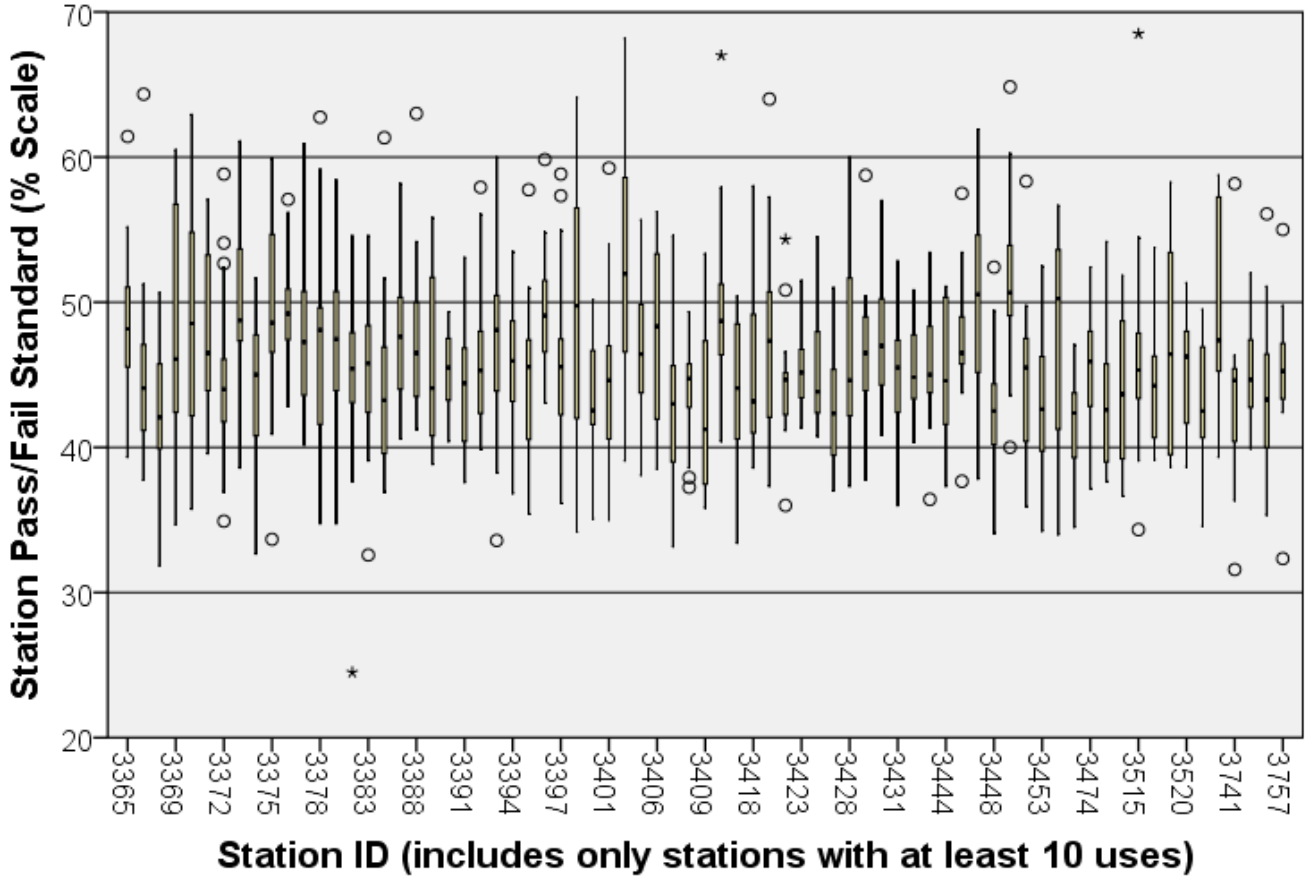

Figure 4: Boxplots of Pass/Fail Standards for Individual Stations

Only stations used on at least 10 test dates are included in the figure. Some stations had multiple versions that appeared to reflect adjustments to the associated materials; "versioning" was ignored in generating the figure.

Reproducibility of Domain Scores in Data Gathering, Clinical Management, and Interpersonal Skills

Because station pass/fail standards are not set for individual domain scores, the procedure described in the section on the results of generalizability analyses was modified to obtain information about the reproducibility of PLAB 2 domain scores in data gathering, clinical management, and interpersonal skills.

1. Domain scores for each candidate and station on each day of testing were assembled and transformed from the "raw" score scale of 0 to 4 into a percentage of possible points scaling by dividing by 4 and multiplying by 100%.
2. A separate one-way random-effects analysis of variance (ANOVA) was performed (using the SPSS VARCOMP procedure) for each domain score on each of the 82 days of testing included in the dataset, with each ANOVA providing estimated variance components for Candidates, Stations, and Error (the Candidate X Station interaction). The first is an estimate of "true score" variance (variation in scores that would be seen if no measurement error were present), the second is an estimate of true variation in station difficulties[17] that would be seen if no measurement error were present, and the last is an estimate of unexplained (error) variance.
3. Estimated variance components from the individual ANOVAs were averaged across the 82 days of testing to obtain a single, more precise (pooled) estimate of each variance component.
4. For each domain score, pooled estimates of variance components were used to calculate generalizability coefficients and SEMs[18] at the actual test length of 18 stations.
5. Pooled estimates were also used to project the effect of test length on both generalizability coefficients and SEMs for each domain score.

Table 3 provides a summary of results. The variance components at the top of the table provide a good basis for comparing the three domain scores. The square root of the candidate variance component can be interpreted as a standard deviation, indicating the true "spread" in candidate scores if no measurement error were present. It ranges from a low of 6.67% for data gathering to a high of 9.03% for interpersonal skills. Because these are SDs conceptually, estimates of the range of each domain score can be approximated by multiplying these values by 4, producing values of roughly 27%, 28%, and 36% for data gathering, clinical management, and interpersonal skills, respectively, all on a percentage of possible points scale. Thus, based on the current marking scheme, the true variation in interpersonal skills is substantially larger than that for data gathering and clinical management.

Using the same approach, the square roots of the variance components for station difficulty indicate that true variation in difficulty is largest for clinical management (11.93%) and smallest for interpersonal skills (9.03%), with data gathering (11.12%) falling in between. These results suggest that examiners apply more consistent marking criteria in judging interpersonal skills than the other two domain scores, though all of the values are fairly large, indicating substantial station-to-station variation in domain scoring.

[17] Differences in examiner stringency (hawk/dove effects) contribute to variation in station difficulty on individual test dates. The design used for PLAB 2 test administration does not allow separate estimation of station difficulty and examiner stringency due to confounding of the two.

[18] Both a relative and an absolute SEM were calculated. The former, numerically equivalent to the values that would be obtained using a computational approach based on coefficient alpha, excludes variation in station difficulty from measurement error. The latter includes that error. Because the borderline regression method, to some degree, takes differences in station difficulty into account, the actual SEM for each domain score is probably best viewed as falling between these two values.

	Data Gathering			Clinical Management			Interpersonal Skills		
	Candidates	**Stations**	**Error**	**Candidates**	**Stations**	**Error**	**Candidates**	**Stations**	**Error**
Variance Component	44.50	123.63	364.12	49.61	142.26	432.94	81.60	85.61	293.26
Square Root of VarComp	6.67	11.12	19.08	7.04	11.93	20.81	9.03	9.25	17.12
Test Length (# of scored stations)	**G-coeff (Relative)**	**SEM (Relative)**	**SEM (Absolute)**	**G-coeff Relative**	**SEM (Relative)**	**SEM (Absolute)**	**G-coeff (Relative)**	**SEM (Relative)**	**SEM (Absolute)**
14	0.631	5.10	5.90	0.616	5.56	6.41	0.796	4.58	5.20
16	0.662	4.77	5.52	0.647	5.20	6.00	0.817	4.28	4.87
18	**0.687**	**4.50**	**5.21**	**0.673**	**4.90**	**5.65**	**0.834**	**4.04**	**4.59**
20	0.710	4.27	4.94	0.696	4.65	5.36	0.848	3.83	4.35
22	0.729	4.07	4.71	0.716	4.44	5.11	0.860	3.65	4.15
24	0.746	3.90	4.51	0.733	4.25	4.90	0.870	3.50	3.97
26	0.761	3.74	4.33	0.749	4.08	4.70	0.879	3.36	3.82
28	0.774	3.61	4.17	0.762	3.93	4.53	0.886	3.24	3.68
30	0.786	3.48	4.03	0.775	3.80	4.38	0.893	3.13	3.55
32	0.796	3.37	3.90	0.786	3.68	4.24	0.899	3.03	3.44
34	0.806	3.27	3.79	0.796	3.57	4.11	0.904	2.94	3.34
36	0.815	3.18	3.68	0.805	3.47	4.00	0.909	2.85	3.24

Table 3: Results of Generalizability Analyses of Domain Scores for the New PLAB 2 Format

The upper portion of the table provides variance component estimates from generalizability analyses of the three domain scores, with each domain score expressed on a percentage-of-points scale. Those variance components were used to calculate the generalizability coefficients and standard errors of measurement (SEM) in the lower portion of the table. Relative and absolute SEMs differ in the definition of the sources of error included in the estimates: differences in station difficulty affect the absolute SEM but not the relative SEM. Calculation of the generalizability coefficients in the table exclude measurement error due to variation in station difficulty and are best viewed as an upper bound on the reproducibility of scores. The row in red represents the new PLAB 2 test length of 18 stations.

Comparison of the square roots of the variance components for Error shows a somewhat similar pattern: it is smallest for interpersonal skills (17.12%) and larger for data gathering (19.08%) and clinical management (20.81%). These values can be interpreted as the (relative) SEM for a test consisting of a single station.

The lower portion of Table 3 provides generalizability coefficients and relative/absolute SEMs for the actual test length of 18 stations and for a variety of shorter and longer tests. These are all projected using the pooled estimates of the variance components provided in the upper portion of the table. At the actual test length, the generalizability coefficient for interpersonal skills (0.834) is quite good. It can be interpreted as the expected correlation between scores on similar (randomly parallel) but not identical circuits of 18 stations. Values for data gathering and clinical management are substantially lower, reflecting both the smaller true variation in candidate scores and the larger magnitude of the error variance.

Taken together with the boxplots in Table 4, these results indicate that examiners are assigning domain scores in data gathering and clinical management somewhat inconsistently, both in terms of stringency (variance component for stations) and more generally (variance component for error). This suggests that improvements in the reproducibility of pass/fail decisions might be achieved by structuring the "rating task" examiners face in providing domain scores to increase comparability for the same station from test date to test date and across stations used on the same test date.

It is not completely clear how this might best be accomplished. Three (complementary) approaches to the problem are outlined below. They are ordered from most to least projected amount of work to implement.

Incorporation of Key Features in Marking. The use of a global 0 to 4 scale for the data gathering, clinical management, interpersonal skills domain scores does not provide markers with much specific guidance for scoring. This could produce widely varying scores for the same stations on different days of test administration; exploratory generalizability analyses of domain scores suggest this is the case for data gathering and clinical management in particular. Drawing on the key features work of Bordage and Page, the Australian Medical Council and the Medical Council of Canada are exploring the use of station development and marking methods based on "key features." This approach would provide examiners with more guidance in assigning domain scores by using short checklists related tied to key features to anchor global judgments, increasing comparability in use of the scales and improving reproducibility of scores and pass/fail decisions.

Enhanced Training of Examiners and Role Players. A complementary approach to increasing comparability could involve additional training of examiners and role players. Good video capabilities are available for recording encounters between candidates and role players. Station-specific illustrations of good and poor candidate performance could be identified, edited (shortened), and used to calibrate both examiner marking and role player portrayal to promote greater consistency in marking and portrayal across test dates.

Adjustments to Candidate Scheduling. For borderline regression methods to work well in adjusting pass/fail standards to account for day-to-day variation in form difficulty, it is desirable to have some strong and weak candidates sit for PLAB 2 on each test administration date. This has a direct impact of the accuracy with which regression equations can be estimated and the precision of the estimated pass/fail standard for each station. Further, each examiner would, ideally, see a strong and weak candidate early in the day to aid in calibration in assignment of marks. The GMC possesses information (eg, IELTS scores, country of medical school, scores on PLAB 1, number of and scores on previous PLAB 2 attempts) that could be used to predict how likely each candidate is to pass PLAB 2 prior to test administration. These predictions could be used in assigning candidates to morning/afternoon sessions and to starting stations in morning circuits, reducing the extent to which the criteria an examiner uses in assigning marks will shift over the course of the test day.