External Review of Assessment Structure for Dental Specialty Fellowship Examinations

Dr Matt Homer, Leeds Institute of Medical Education, University of Leeds

m.s.homer@leeds.ac.uk

Contents

Executive summary

This review provides an expert external review of the CSO element of the proposed assessment structure for Dental Specialty Fellowship Examinations. It draws on a range of documentation provided by the DSFE and the wider assessment literature. It is advisory in nature and intended to be proportionate in recognising the necessary balancing of competing demands that are always present when designing assessment systems – for example, feasibility/resources versus reliability/validity.

Key findings

- The review finds that the DSFE have been informed by a range of world-leading evidence and best-practice in assessment. This is important because the assessment design process, particularly in high-stakes settings, should be evidence-based and be seen to be so.
- There is a potential risk of low reliability given the relatively short length of exam (at nine stations) – see recommendations 1 and 2 below for what might be done to reduce this risk.
- The time allowed for preparation and assessment in each CSO station seem reasonable (but see also recommendation 1 below).
- The policy of using a single examiner in each station is completely defensible and entirely aligns with assessment guidance and practice in many healthcare and other settings.
- As it stands, there is the theoretical possibility that a candidate could 'fail' the majority of stations but pass the exam – see recommendation 3 for how this might be avoided.

Review recommendations

The following recommendations are made (full details are given in the main text):

1. Piloting of groups of stations under the new design should be carried out to give some emergent data around the likely reliability of the new exam structure.
2. Consideration should be given to increasing the length of the exam to ten stations if at all possible – this would help ameliorate the potential risk of unacceptably low reliability for the new exam.
3. To limit compensation across stations, consideration could be given to introducing a minimum-station-to-pass requirement for candidates - assuming a fully compensatory model is not an underlying key principle in the new design. This might increase the acceptability of the proposed exam to stakeholders.
4. The introduction of a global overall grade of performance in each station could be considered. This might have a number of potential benefits – for example, providing some triangulation of data in stations and allowing experimentation with empirical methods for setting standards in the station.

5. Steps should be taken to ensure that station materials promote best-practice examiner calibration. This will help ensure that scoring is as consistent and fair as possible across parallel circuits in the larger-cohort specialties.

Overall review summary

This review generally supports the current proposal for the design of the CSO – and is confident that it will produce an effective assessment model with valid outcomes. The design balances well a range of challenging issues and will likely have a beneficial impact on patient safety provided that the recommendations are considered carefully and acted on appropriately.

Introduction

This document is intended to provide an expert external review of the proposed assessment structure for Dental Specialty Fellowship Examinations – in particular, the clinical structured oral (CSO) component. It draws on a range of documentation provided by the DSFE and wider assessment related evidence (i.e. international assessment literature and best practice guides relevant to postgraduate clinical assessments). The intention is to provide an external and independent 'sense check' to the DSFE as to whether the proposed changes to the examinations are in line with current assessment thinking and, as appropriate, to suggest other issues that might not yet have been fully considered as the proposed changes have been developed.

The first comment to make on reviewing the proposals is that it is clear that the DSFE and those responsible for the proposed assessment structure have already been informed by a range of world-leading evidence and best-practice. The proposals clearly draw on key documentation – for example the Ottawa guide to good assessment – the most recent version of which is 2018 (Norcini et al., 2018), Kane's conceptual framework for assessment validity (Downing, 2003; Kane, 2013) and the Standards for Educational and Psychological Testing (American Educational Research Association, 2014). This underpinning with strong evidence is important as these sources are widely used by a wide range of institutions and regulators when developing or certifying effective high-stakes assessment practices across the world – for example programmes of assessment within medical education and other healthcare professions. The over-arching principle is that development of assessment practice needs to be considered as an evidence-based process – and requires significant theoretical and conceptual foundations to maximise the quality of the decisions that are made and the efficacy of the resulting assessment programme and its outcomes.

That said, no assessment system or individual exam is perfect and there are always many constraints that test developers are working under or trying to balance – related to logistics (e.g. is the assessment possible to run in terms of demands on examiners and candidates?), feasibility (e.g. is the exam affordable?) and a range of other challenges. Such limitations need to be recognised and explicitly taken account of when developing assessments – but, crucially, the evidence supporting the argument for the validity of the assessment and the inferences made on the outcomes (e.g. candidate pass/fail decisions) must be sufficiently strong. This is the essence of Kane's (2013) framework and one that is widely cited when investigating whether an exam's outcomes are defensible. Pragmatic decisions always have to be made taking into account the many constraints that are always present.

This review is intended to be proportionate and advisory in nature. It continues with sub-sections on five specific issues - three (number of stations, time allowed in stations, the use of a single examiner in stations) are those that have arisen as the proposals have developed and a further two (compensation across stations and maximising calibration) were identified during the course of the review.

The review finishes with a summary of the strength and risks of the proposal and a few final comments.

The number of stations in the CSO

The current proposal is for a nine station CSO with domain marking across four pre-defined skills areas (Clinical evaluation and assessment, Interpretation of investigations and diagnosis, treatment planning and management, and prognosis/maintenance).

The number of stations required in a high-stakes clinical assessment is a contentious issue with no single easy answer applicable across all contexts (Khan et al., 2013). There are clear trade-offs between reliability, wider validity, feasibility and demands on resources. One key issue is that of content specificity (Eva et al., 1998) where a candidates might do poorly in some areas/domains and better in others – which implies that a shorter test might not yield scores that reflect their real abilities very well. Clearly an assessment that samples more widely via more stations will have higher reliability (for essentially technical reasons) but also improved validity (the domains being assessed are better sampled) than they are in a shorter test. But obviously longer assessments are more resource intensive in many ways.

It is clear that nine stations is at the lower end of the spectrum for such an assessment according to the literature, although admittedly a lot of the research is focussed on undergraduate exams where, perhaps a wider range of domains/constructs are being assessed, in turn implying the needs for more stations. With nine stations, it is possible that the reliability (e.g. an alpha or G-coefficient) of the CSO might fall below the value preferred in such a high stakes exam (e.g. below the often-cited, though somewhat arbitrary, value 0.7/0.8). However, without some quantitative score data (e.g. from a pilot of stations), it is impossible to judge exactly the likelihood of this happening. Importantly, there is also a qualitative judgment to be made – can a blueprinted sample of nine stations represent the domain(s) being assessed well? Presumably, the current feeling amongst the assessment teams is that this indeed the case. It may well be that in quite narrow domains – as is perhaps likely the case in postgraduate dental speciality curricula – that nine will work well. It is also likely that domain marking will enhance reliability (and validity) with the same constructs being assessed across stations – leading to good correlations across these, which in turn lead to higher reliability coefficients.

Recommendations

1. Ideally, some piloting of groups of stations under the new design should be carried out – as well as gathering a range of other, more qualitative insight into station performance based on feedback from pilot candidates/examiners/observers (Khan et al., 2013). The scoring data will provide some formative quantitative data to be used to estimate the likely reliability of the full exam (using Spearman-Brown prophecy formula (Thompson, 2024). See also the next sub-section - piloting could also be informative on station timings.
2. Given the relatively low number of stations suggested, consideration should be given to making the exam a bit longer – perhaps increasing to ten stations if possible. This would almost certainly increase typical reliability/validity and would also help in cases where stations perhaps did not perform as expected – and might need to be removed from the exam *post hoc* based on feedback and psychometric data.

Time allowed for preparation and assessment in each CSO station

The proposed reading time of up to 10 minutes (varying by specialty) is quite long in comparison with many other scenario-based assessments, although specific evidence on this in the literature is limited. However, beyond lengthening the total testing time for a fixed number of stations, this between-station time is unlikely to be an important factor in determining the validity of the assessment and associated outcomes. In essence, this is a qualitative judgment to be made by the specialty assessment team – to allow a reasonable but not disproportionate amount of time for candidates to read/prepare for the station

scenario(s). Again, piloting, and previous and ongoing experiences will help inform judgments as to the optimal time required.

According to the wider literature, the actual time allocated to the station in a performance assessment like the CSO can vary considerably – from (say) 5 to 25 minutes (Harden et al., 2015, p.69). Obviously, the longer each station is (including reading time), the fewer stations can be administered in a fixed amount of total testing time. All other things being equal, a design with shorter but more plentiful stations is advantageous in likely leading to higher reliability coefficients. The obvious counter-argument is that if scenarios are artificially limited to smaller, unrealistic 'chunks' of cognitive questioning then shorter stations are a serious threat to assessment validity.

The fifteen minutes time, as proposed for the CSO, whilst being towards the longer end of the typical duration, seems entirely reasonable depending exactly on the nature of the higher-level complex scenarios within each station at this postgraduate level. Again, the time appropriate to each station is essentially a qualitative judgment for the assessment development team, informed by piloting and experience. It is important that there should be adequate time for minimally competent candidates to answer the structured questions - unless an important element of the assessment relates to measuring the speed of performance of candidates – since test 'speededness' can be another threat to validity (Lu and Sireci, 2007).

Even for complex and multi-layered cases, it should be possible with 15 minutes to design a range of valid scenarios in stations that allow candidates to demonstrate a wide range of multi-layered skills including, for example, elements of diagnostic reasoning and higher-level clinical reasoning skills. One has to accept that scenarios/questions in a structured exam necessarily remain imperfect imitations of genuine clinical practice – but, with good design, high-order cognitive skills can be assessed appropriately in a 15-minutes.

The policy of using a single examiner in each CSO station

The question of how many examiners should mark in a station is one that re-occurs frequently in discussions around the design of new performance assessments but is an issue that has been essentially settled for a long time. The evidence from medical OSCEs is unequivocal - it is far better to have a single examiner in a station (and more stations if possible), compared to having two (or more) examiners in a single station (van der Vleuten and Swanson, 1990; Harden et al., 2015, p.109). The reason for this is that the evidence suggests that examiner judgements actually vary little in comparison to how much candidate performance varies across cases/stations (content specificity again) – which in turn implies it is better (all other things being equal) to sample more across cases/stations than 'wasting' examiners marking in the same station. Whilst this argument might be seen a little technical (based on psychometric analysis of score patterns), it actually also speaks strongly to the issues of feasibility and logistics – trained examiners can be expensive and are often rare or difficult to supply in adequate numbers in many assessment contexts.

There is a degree of natural appeal with the idea that two examiners allow for real-time moderation and shared judgements in stations. However, it is generally to the benefit of candidates that examiners are allowed to carry out professional judgments alone in stations – guided by appropriate training materials, calibration discussions and scoring rubrics. As has already been discussed, the risk of variable decision-making and inconsistency in examiner judgments is best ameliorated having single examiners in stations and, other things being equal, more stations.

In summary, the proposal for a single examiner in each CSO station is entirely consistent with the assessment literature and wider performance assessment practices across a wide range of settings. As well as there being good psychometric arguments for a single examiner in a station, there are also likely important resource arguments in favour of this decision.

Given that the CSO assessment proposals are generally following best practice when it comes to design, scoring and so on, experience also suggests that the issue of candidate appeals is essentially a non-starter. With appropriate examiner training and a degree of ongoing observation in stations, there should be strong administrative evidence that examiner judgments are defensible. In rare cases where any concerns raised are found to have some justification, then steps can be taken to re-assure candidates and other stakeholders – for example, by adjusting marks or, as the ultimate action, removing problematic stations from the exam *post hoc*. There is evidence in the literature that 'group' effects (e.g. pairs or groups of examiners) are not immune from bias – for example, Angoff judges have been shown to be influenced by randomly generated data as to the difficulty of written items (Clauser et al., 2009). This suggests that psychological 'herd' effects can happen in high-stakes assessments and that pairing of examiners is no guarantee of improved 'objective' scoring of candidates.

Minimising compensation across stations

As it stands, the proposed assessment is fully compensatory in nature – the final score in the exam is the sole outcome that determines the exam-level pass/fail decision and there is no requirement for candidates to have to also pass a certain number of stations (a so-called 'conjunctive' standard). This is a challenging area of assessment practice and subject to long running debate and varying practice across healthcare professions (Clauser et al., 1996; Homer and Russell, 2021).

In the current proposal, with four domains in a station, and a 1-5 rating scale with 'borderline-pass' set at 3, the passing standard at the station level is 12 (=4×3) marks and a nine-station exam will have a passing score of 12×9=108. Imagine then a hypothetical candidate who performs relatively poorly on (say) five stations but strongly on the other four – see the tabulated example below. This is arguably an extreme example but there are other more moderate patterns of scoring that would have the same issue - the candidate has failed the majority of stations but has easily exceeded the passing score requirement.

Station	1	2	3	4	5	6	7	8	9	Total
Score	11	11	11	11	11	20	20	20	20	135
Pass/fail	F	F	F	F	F	P	P	P	P	P

Even if this is perhaps an unlikely scenario (and might never happen in practice), the stated emphasis on patient safety in the proposed assessment structure suggests that it would make sense to consider a pro-active change that would prevent this problem from occurring as follows - assuming a fully compensatory model is not an underlying key principle in the new design.

Recommendations

3. Consider the introduction of such a conjunctive standard – perhaps the simplest would be to say that candidates need to pass a minimum of five stations (i.e. over 50% of stations). Whilst this is a fixed standard, rather than being criterion-

referenced, it might re-assure a range of stakeholders – particularly those most concerned with patient safety issues in the proposed assessment structure. It may be that an example of candidate performance like that presented above would never happen, but this change might help 'acceptability' concerns and, perhaps also have beneficial educational effects (encouraging candidates to spread their learning across all domains being assessed rather than focussing on excelling in a few).

4. Consider the introduction of a global overall grade of performance in each station. This could provide triangulation of data in stations where there are big enough cohorts (are there stations where there is poor alignment between the global and domain scores?) but would also allow for experimentation with empirical methods for setting standards in the station – for example, borderline regression (McKinley and Norcini, 2014) - and the minimum station requirement proposed in recommendation 3 (Homer, 2023). Obviously, the global rating scale would need careful development – with appropriate levels/anchors (e.g. fail, borderline, pass, clear pass) and grade descriptors – particularly at the key borderline grade.

Maximising calibration across circuits

In the specialties with larger cohorts, there will be parallel circuits where pairs of examiners in the same station 'see' different cohorts of candidates. In order to ensure fairness and equivalent treatment across groups of candidates, it is important that calibration practices are sufficiently robust to minimise differences in patterns of scoring across circuits and to promote better alignment in examiner behaviours across circuits.

Recommendations

5. Steps should be taken to ensure that materials promote best practice calibration in stations. This might include more detailed and relevant scoring and calibration guidance/checklists and related support materials in stations – see for example Homer and Ababei (2026) for some ways that this can be done.

Summary of strengths and risks of the proposal

The proposal for the new assessment structure is evidence-based and founded on modern educational and assessment practices – and is not driven solely by financial or other resource considerations. The standardisation, whilst possibly bringing specific challenges for some specialties, also has clear benefits in terms of easing issues of governance and oversight across these, and also allows for the sharing of examiners (and possibly others involved in assessment development) across them.

The move to domain marking is fully supported in the literature – for assessing higher-level skills as candidates move towards independent practice - in comparison with, for example, checklists for more routine scenarios (Harden et al., 2015, p.135). Again, the decision to standardise the use of these across all stations - all four domains in all stations and across all specialties - is likely to bring a range of interoperability benefits to the overall system.

In terms of risks, there are always challenges in moving from a long-standing within-speciality practices to a standardised model across these but as already stated there are likely to be considerable benefits. Given how widespread the adoption of OSCE and related models of structured assessment have been over recent decades, both within and without of healthcare professions, it seems unlikely that there would be a particularly 'bad fit' of the proposed assessment model to certain specialties that meant these challenges cannot be

overcome. It is important to remember that the CSO is merely one assessment tool that sits within a broader programme of assessment (Torre and Schuwirth, 2025) for these specialties, and that no assessment can or should be intended to do 'everything' in terms of assessing all learning outcomes of each specific programme/curriculum.

The largest risk identified in this review is probably that of lower reliability than would be preferred for a high stakes exam – given the relative low numbers of stations. As one of the recommendations suggests, increasing the number of stations to ten alongside, appropriate piloting and psychometric analysis, would provide some mitigation to this risk.

The other risk, based on documented examiner concerns that have been communicated to the DSFE and to this review, is one related to acceptability of the CSO proposal – for particular specialties. There is little that can be said on this issue, other than to hope that this review itself, and the specific issues covered, will help to enhance the acceptability of the proposal to these groups.

Overall, this review concludes that the proposed changes to the CSO element will likely have a beneficial impact on patient safety provided that the recommendations are considered carefully and acted on proportionately. The proposed assessment model balances a range of competing issues and demands in a way attempts to be fair to all, is likely to be cost effective, and does not over-burden candidates – but is likely to produce an effective assessment model with valid outcomes.

References

American Educational Research Association 2014. *Standards for Educational and Psychological Testing, 2014 Edition*. American Educational Research Association.

Clauser, B.E., Clyman, S.G., Margolis, M.J. and Ross, L.P. 1996. Are fully compensatory models appropriate for setting standards on performance assessments of clinical skills? *Academic Medicine*. **71**(1), pp.S90–S92.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J. and Dillon, G.F. 2009. Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*. **46**(4), pp.390–407.

Downing, S.M. 2003. Validity: on the meaningful interpretation of assessment data. *Medical Education*. **37**(9), pp.830–837.

Eva, K.W., Neville, A.J. and Norman, G.R. 1998. Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Academic Medicine: Journal of the Association of American Medical Colleges*. **73**(10 Suppl), pp.S1-5.

Harden, R., Lilley, P. and Patricio, M. 2015. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment., 1e* 1 edition. Edinburgh ; New York: Churchill Livingstone.

Homer, M. 2023. Setting defensible minimum-stations-passed standards in OSCE-type assessments. *Medical Teacher*. **45**(10), pp.1163–1169.

Homer, M. and Ababei, V. 2026. Evidencing improvement in examiner calibration in OSCEs. *Medical Teacher*. **0**(0), pp.1–11.

Homer, M. and Russell, J. 2021. Conjunctive standards in OSCEs: The why and the how of number of stations passed criteria. *Medical Teacher*. **43**(4), pp.448–455.

Kane, M.T. 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*. **50**(1), pp.1–73.

Khan, K.Z., Gaunt, K., Ramachandran, S. and Pushkar, P. 2013. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Medical Teacher*. **35**(9), pp.e1447–e1463.

Lu, Y. and Sireci, S.G. 2007. Validity Issues in Test Speededness. *Educational Measurement: Issues and Practice*. **26**(4), pp.29–37.

McKinley, D.W. and Norcini, J.J. 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*. **36**(2), pp.97–110.

Norcini, J., Anderson, M.B., Bollela, V., Burch, V., Costa, M.J., Duvivier, R., Hays, R., Palacios Mackay, M.F., Roberts, T. and Swanson, D. 2018. 2018 Consensus framework for good assessment. *Medical Teacher*. **40**(11), pp.1102–1109.

Thompson, N. 2024. What is the Spearman-Brown formula? *Assessment Systems*. [Online]. [Accessed 10 June 2026]. Available from: https://assess.com/spearman-brown-prediction-formula/.

Torre, D. and Schuwirth, L. 2025. Programmatic assessment for learning: A programmatically designed assessment for the purpose of learning: AMEE Guide No. 174. *Medical Teacher*. **47**(6), pp.918–933.

van der Vleuten, C.P.M. and Swanson, D.B. 1990. Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*. **2**(2), pp.58–76.

About the author

Matt Homer has been a leading practitioner and researcher in medical education assessment for 20 years. His interests and experience in assessment also extend into a wide range of other healthcare disciplines and professions. His work focusses on improving quality assurance practices in high-stakes exams, developing standard setting methods and quantifying examiner variation in judgments of clinical performance.

He has an international profile within medical education, advising the General Medical Council (GMC) and other organisations/regulators on assessment practice and psychometric issues.

Matt also has experience in wider educational research, has led large externally funded research projects, has published over 90 research articles and has supervised 18 PhD students to completion.

Dental Examinations Executive

Response to external review of examination structure for Dental Specialty Fellowship Examinations

Background

In response to feedback and questions from examiners around the validity of the examination structure for Part 2 of the Dental Specialty Fellowship Examinations (DSFE), a decision was taken to commission an external review. This was undertaken by Dr Matt Homer. More information on Dr Homer and the remit of the review can be found in the Appendix. This document sets out the formal response of the Dental Examinations Executive (DEE) to the review.

Overall review summary

This review supports the current proposal for the design of the Part 2 Clinical Structured Oral (CSO) examination and provides reassurance that it will produce an effective assessment model with valid outcomes. It offers recommendations to further improve reliability.

Key findings of the review

The review provided five key findings:

1. The review finds that the DSFE have been informed by a range of world-leading evidence and best-practice in assessment. This is important because the assessment design process, particularly in high-stakes settings, should be evidence-based and be seen to be so.
2. There is a potential risk of low reliability given the relatively short length of exam (at nine stations) – see recommendations 1 and 2 below for what might be done to reduce this risk.
3. The time allowed for preparation and assessment in each CSO station seem reasonable (but see also recommendation 1 below).
4. The policy of using a single examiner in each station is completely defensible and entirely aligns with assessment guidance and practice in many healthcare and other settings.
5. As it stands, there is the theoretical possibility that a candidate could 'fail' the majority of stations but pass the exam – see recommendation 3 for how this might be avoided.

Recommendations of the review, with response from Dental Examinations Executive

The review provided five recommendations to the DEE for consideration, all of which have been carefully considered, and a formal response to each is provided:

Recommendation 1 - Piloting of groups of stations under the new design should be carried out to give some emergent data around the likely reliability of the new exam structure.

Recommendation 2 - Consideration should be given to increasing the length of the exam to ten stations if at all possible – this would help reduce the potential risk of unacceptably low reliability for the new exam.

Response to Recommendations 1 and 2 – DEE welcomes the recommendation to generate early evidence on the performance of the new examination structure. This would provide an important opportunity to demonstrate to examiners, specialty examination boards and the wider stakeholder group that the examination structure is not only evidence-informed, but also quality-assured before, during and after implementation.

DEE accepts Recommendation 1 in principle and will adopt a proportionate, practical approach to pre-implementation quality assurance across specialties. Recognising the logistical challenges of establishing a new examination across ten specialties, this work will focus on activities that provide meaningful assurance without delaying implementation. This may include trialling of stations with volunteer candidates or recent trainees, alongside the detailed calibration work referred to in Recommendation 5. The Executive is developing a programme of appropriate quality assurance activities, and further information will be shared with key stakeholders in due course.

Reliability data from other Royal College examinations using similar marking approaches provides confidence in the sufficiency of a nine-station model. However, DEE accepts the principle that increasing the number of stations is likely to further improve reliability and reduce the risk of unacceptably low reliability. The examination structure will therefore be increased to ten stations. This strengthens the sampling of the syllabus, broadens the evidence base on which candidate judgements are made, and further supports the defensibility of the exam structure.

This approach is particularly relevant given the review findings that, where resources are being allocated for an examination, overall reliability is generally better supported by more stations with a single examiner than by fewer stations with multiple examiners. DEE considers this to be a positive and pragmatic design choice: it preserves the breadth of assessment, uses examiner expertise efficiently and maintains a structure that is deliverable across specialties. The reliability, validity and acceptability of the revised structure will be reviewed routinely after every diet, with findings reported through the appropriate examination governance routes and shared with stakeholders.

Recommendation 3 – To limit compensation across stations, consideration could be given to introducing a minimum-station-to-pass requirement for candidates – assuming a fully compensatory model is not an underlying key principle in the new design. This might increase the acceptability of the proposed exam to stakeholders.

Recommendation 4 -The introduction of a global overall grade of performance in each station could be considered. This might have a number of potential benefits – for example, providing some triangulation of data in stations and allowing experimentation with empirical methods for setting standards in the station.

> **Response to Recommendations 3 and 4** – The new examination adopts a domain-based approach to assessing candidate competence in four key areas of clinical judgement across a range of different topics from the exam syllabus. Consequently, examiners are being asked to contribute to an overall picture of competence in these domains, rather than to provide summative judgements of global competence in the context of individual scenarios.
>
> This means that identifying station 'pass marks', introducing a minimum number of stations that must be passed, or using pass or global judgements for station performance – either in the results determination process or through feedback provision – presents candidates and examiners with a potentially mixed message about the examination structure and assessment approach. DEE therefore does not propose to introduce a minimum station to pass requirement or global overall grade at this stage.
>
> It is theoretically possible that a candidate could pass the examination overall whilst failing to achieve the station 'pass mark' in the majority of stations, but this is extremely unlikely in practice. This is because it would require candidates to be scoring at the extremes of the marking scale in different stations – a pattern which is not found in practice, based on existing data from similar examinations.
>
> The performance of the domain-based model, including the extent of compensation across stations, will be monitored routinely after every examination diet and reviewed by DEE.

Recommendation 5 – Steps should be taken to ensure that station materials promote best-practice examiner calibration. This will help ensure that scoring is as consistent and fair as possible across parallel circuits in the larger-cohort specialties.

> **Response to Recommendation 5 –** DEE recognises that effective examiner calibration is fundamental to the success of the domain-based approach being implemented with the new exam. A strong calibration process will be central to demonstrating that the examination is fair, effective and robust. Calibration will be a structured quality assurance process that begins during station development and continues through examiner training, delivery, results review and post-diet evaluation.
>
> Station materials and examiner guidance will be developed to support consistent interpretation and application of the marking criteria. The programme for each exam diet will include a structured session in which examiners can

familiarise themselves with marking criteria, discuss and agree how best to apply these and rehearse station encounters prior to examining candidates. This will enable examiners to apply their judgements with confidence and consistency across all circuits.

Full report

The full report, alongside this response, will shortly be available on the DEE website.

Next Steps

The DEE is working with DSFE Chairs and Examination Boards on implementation of the examination, including the modifications and assurance processes set out above. As part of a programme of ongoing quality assurance and improvement, all examination diets will be carefully reviewed, and outcomes will be used to refine station design, examiner guidance and calibration processes. Information will be shared with examiners and key stakeholders through appropriate governance routes so that the examination continues to develop transparently and constructively. DEE is grateful for the engagement of examiners, Specialty Examination Boards and specialist societies, and looks forward to continued dialogue as the examinations are implemented and strengthened.

Dental Examinations Executive, August 2026

Appendix

The External Review

The reviewer

A number of organisations and individuals were approached to carry out the review and the work was undertaken by Dr Matt Homer.

Dr Homer is a leading practitioner and researcher in medical education assessment, who has an international profile in this area, having advised the General Medical Council (GMC) and other organisations and regulators on assessment practice and psychometric issues. His work focuses on improving quality assurance practices in high-stakes exams, developing standard setting methods and quantifying examiner variation in judgments of clinical performance.

The scope of work

The review brief asked the consultant to undertake a structured review of the proposed assessment framework, with particular emphasis on the clinical structured oral component.

Three specific areas were highlighted for review, taking into account the aim of the examination and the postgraduate environment in which it sits:

- The number of stations;
- Time allowed for preparation and assessment in each station; and
- The policy of using a single examiner in each station and pros/cons of this.

Where areas are highlighted as having elements of risk, recommendations for mitigation were invited.

The reviewer was made aware of the communications received from a wide range of stakeholders and provided with details of the concerns which had been raised in this respect.