Evaluation of a High-stakes Summative Examination in Oral Surgery with a view to Improving its Quality

Dr Yvonne Katharine Hurst

BSc (Hons), PhD

Kelso, Scotland

Submitted for the Degree of Masters in Health Professions Education: Assessment and Accreditation, Keele University

Submitted: 19th March, 2016

DECLARATION

I certify:

(a) That the above dissertation/project is my own account, based upon work actually carried out by me, and that all sources of material not resulting from my own experimentation, observation or specimen collecting, including observational data, have been clearly indicated.

(b) That no part of the work incorporated in the above dissertation/project is a quotation from published or unpublished sources, except where this has been clearly acknowledged as such, and that any specific direction or advice received is also properly acknowledged.

(c) That I have read, understood, and abided by the terms of Regulation VII1.5 below:

CONDUCT WITH REGARD TO DISSERTATIONS, PROJECTS, ESSAYS ETC., WHICH FORM PART OF A FINAL EXAMINATION FOR ASSESSMENT PURPOSES

(a) Titles must be approved or specified by the Department concerned in accordance with the provisions in the Calendar.

(b) The dissertation, projects or essays etc. shall be in the student's own words, except for quotations from published and unpublished sources which shall be clearly indicated as such and be accompanied by full details of the publications concerned. The source of any map, photograph, illustrations, etc. shall be similarly indicated. The student shall indicate clearly the sources, whether published or unpublished, of any material not resulting from his own experimentation, observation or specimen collecting, including observational data. Students will be required to sign a statement to that effect. Failure to comply strictly with these requirements may be construed as cheating.

Signed:

Date: 14th March, 2016

ACKNOWLEDGEMENTS

My thanks go to my supervisor John Boulet for his advice and support, my family, Edward, Emma and Bella Hurst, for their good humour and forbearance and Dr Cara Featherstone for performing the psychometric analyses for the study.

TABLE OF CONTENTS

SUMMARY

A successful outcome in the UK Tricollegiate Specialty Membership Examination in Oral Surgery (MOralSurg) is a prerequisite for entry onto the specialist list held by the regulator, the General Dental Council, and for onward progression to a National Health Service (NHS) consultant post.

This thesis is an evaluation of the quality of the MOralSurg against published criteria using the validity framework described by Kane (1994) and the utility framework described by Van der Vleuten (1996). A rationale is provided for the evaluation itself and for the frameworks used. The study was informed, in the main, by documentation and data relating to the examination structure, administration and outcomes, and observation of examination delivery; interviews with the examination administration teams and faculty were also conducted.

Many areas of good practice were identified, supporting the use of the examination for high-stakes decision making. Areas for further development were highlighted and suggestions for quality improvement provided.

ABBREVIATIONS

AoMRC	Academy of Medical Royal Colleges
BLR	Borderline Regression
COPDEND	Committee of Postgraduate Dental Deans
EMQ	Extended Matching Question
GDC	General Dental Council
GMC	General Medical Council
MOralSurg	Tricollegiate Membership in Oral Surgery
NHS	National Health Service
OSCE	Objective Structured Clinical Examination
RCSEd	Royal College of Surgeons of Edinburgh
RCSEng	Royal College of Surgeons of England
RCPSG	Royal College of Physicians and Surgeons of Glasgow
SAC	Specialty Advisory Committee
SBA	Single Best Answer
WBA	Workplace based assessment

INTRODUCTION

This section describes the role and place of the Specialty Membership Examination for Oral Surgery of the Dental Faculties of the three United Kingdom (UK) Surgical Royal Colleges (MOralSurg) in the UK training pathways that lead to recognised specialist status. It provides an overview of the history of the examination and the drivers for this evaluation study.

A successful outcome in the MOralSurg is a prerequisite for gaining entry to the specialist list, held by the regulator, the General Dental Council (GDC) via the traditional route. Inclusion on the specialist list is a requirement for securing a consultant post in the National Health Service (NHS). NHS consultants, having completed their postgraduate training and gained the necessary experience, lead clinical teams, provide expertise across the breadth of their specialty, undertake leadership and managerial roles and bear the ultimate responsibility for patient care (BMA CCSC, 2008). Being that they carry increased responsibility, consultant posts attract higher salaries than training or staff grade posts.

The traditional route to specialist status is open to UK trained applicants, who in addition to passing the exam, must also have been awarded a certificate of completion of training, having successfully completed a GDC approved deanery led training programme. Places on postgraduate specialty training programmes are limited and entry is competitive (COPDEND, 2013). Alternative routes to specialist registration are open to applicants trained outside the UK who must provide evidence of equivalence or meet the exemption criteria in accordance with regulators requirements. The only route available to UK graduates who have not followed the traditional route, is to provide

evidence of knowledge and experience gained through academic practice. These alternative routes are currently under review (Pierce et al, 2014).

Curricula for the specialty training programmes, which include the assessment blueprints, are set by the relevant Specialty Advisory Committee (SAC) of the Royal Colleges Dental Faculties and approved by the GDC. The Royal Colleges and Deaneries are responsible for the design and delivery of the different components of the assessment system for specialist training. A common approach to assessment blueprints designed by the SACs is to have the College specialty membership exams focus on the cognitive aspects of clinical competence that map to the lower "know" and "knows how" levels of Miller's (1990) competency pyramid. Interpersonal and clinical skills, mapping to the "shows how" level, are also included in a number of College examinations. Assessment of whole task performance, exemplified by Miller's "does" level, is commonly addressed using workplace based assessment (WBA) within deanery led training programmes. Skill areas such as technical dentistry, and domain independent areas such as team working, research, and professionalism, that are better suited to performance based assessment, are also covered by the Deaneries.

In the early 1990s, when specialist lists were introduced, each of the UK Surgical Royal Colleges developed and administered its own specialty membership exams. These unicollegiate exams all mapped to the same curricula, but there was no attempt to standardise formats or to provide evidence of equivalence. Candidates chose which exam to sit. However, in 2011 there was an agreement between the three Dental Faculties to develop and deliver joint (tricollegiate) exams in three specialties (Paediatric Dentistry, Special Care Dentistry and Oral Surgery) (Ibbetson, 2012; 2013). The exam department from each of the Colleges would take responsibility for delivering one of the

specialty membership exams on behalf of the others. The rationale behind this move was that as each examination diet attracted only small numbers of candidates; pooling would be cost effective and allow more robust analysis of exam performance needed to inform its ongoing development.

The decision having been made to proceed on a tricollegiate basis, exam departments and boards were required to deliver the first diets of these new exams expeditiously. The Tricollegiate Oral Surgery exam was the first to be delivered in March 2012, the regulations and guide to candidates having been published later than planned in December 2011. Delivery of the exam followed educationalist input, discussions on exam structure and eligibility arrangements, and a change in Board Chair. The examinations for Special Care Dentistry and Paediatric Dentistry were delivered later that year.

As part of the internal quality management processes, all examinations were reviewed after their first administrations and some changes were made to subsequent diets of the Special Care Dentistry and Oral Surgery exams to address issues identified. After three administrations, a more in-depth evaluation was requested by the Examination Executive through the three Deans of the UK Dental Faculties.

The Oral Surgery examination was the first to be evaluated as it was deemed most in need of review. Moreover, the board and exam department were very receptive to change**.** The exam structure had already been through a number of iterations during its short lifespan and the time available for its development had been limited. The Exam Board Chair, supported by external advisors, had identified potential areas for improvement relating to the design, delivery and cost effectiveness of the examination. The department tasked with delivering the exam had undergone significant staff

turnover and were receptive to external scrutiny of their processes**.** The specialty and the board presented some unique challenges; members had been trained via different programmes and pathways (e.g. training programmes in oral and maxillofacial surgery, oral surgery or surgical dentistry in the UK and their equivalents overseas, and transitional routes based on clinical experience) and many were newly appointed.

This paper describes an evaluation of the MOralSurg with respect to its planning (structure and policy), delivery (processes) and performance (outcomes), taking into account psychometric and educational criteria for good quality assessment and the purpose and place of the examination within the wider assessment and training programme.

The literature review considers the reasons and need for evaluating summative examinations used in the health professions. It examines the commonly cited criteria against which the quality of health professions assessments are judged and explores the rationale for their use. Frameworks for evaluating the quality of high-stakes summative assessment are compared and the evidence that can be accumulated is categorized and described. The methods section describes the rationale for the chosen evaluative framework, the sources and approaches used to gather the information, and the process used to compare exam design and delivery procedures against criteria within the framework. The results chapter describes the examination structure, policies, processes and outcomes as evidenced through data gathering. The discussion chapter identifies strengths and weaknesses of these components against the evaluative criteria, taking into account the relative importance of the different criteria (e.g. psychometric versus logistical) in relation to the purpose of the examination. The conclusion chapter provides

recommendations for the improving the exam structure, processes, and policies in line with its agreed purpose.

LITERATURE REVIEW

The purpose of the review

The literature review considers current drivers for evaluating summative examinations used in the health professions from the perspectives of different stakeholder groups. It examines the commonly cited criteria against which the quality of health professions assessments are judged, exploring and comparing the rationale for their use in relation to their purpose and the context in which they are delivered. It synthesizes different types of evidence used to support or challenge claims concerning the suitability and defensibility of high-stakes formal examinations. In so doing, it provides an overview of what constitutes good assessment practice.

Conducting the review

For an overview of good practice and evaluation frameworks based on empirical studies, theoretical considerations and suitable evidence, a literature review was conducted using the following terms interchangeably within each theme with terms from the other themes. Theme 1: “validation”, “validity” “evaluation”, “criteria”, “framework” or “quality”; Theme 2: “assessment”, “examination”, “assessment system” or “test”; Theme 3: “high-stakes”, “licensure”, “certification”, or “summative”; Theme 4: “competence”, “competency based”, “clinical competence” or “health professions”.

The titles generated were used to narrow down sources as appropriate to the study, and the abstracts were reviewed. The references cited in the shortlisted articles and the papers that subsequently cited them were considered in a second review of the literature. Authors consistently associated with relevant material were also used as search terms.

Sources reviewed included journal articles and book chapters. A number of websites were also explored, providing access to reports by key organisations (e.g. GMC). Of these sources, some could be considered primary in that they provided the author's original work, view, approach or interpretation of a concept (e.g. WFME, 2003). Others provide an evaluation, review or synopsis of other's work and, as such, could be considered as secondary sources (Popham, 2000). A high proportion of the text book chapters and journal articles were tertiary sources in that they provide commentaries on other's work but use this to synthesize new approaches (e.g. Kane, 1994).

While the review focused largely on health professions education literature, material from the broader field of education was included where the principles or approach were generalizable or it provided an extended explanation of a framework. As the review considered the evolution of criteria and frameworks, no restrictions were placed in relation to publication date.

The findings

Why evaluate assessments governing entry to the health professions?

Evaluation can be described as a thorough and systematic activity aimed at judging the efficiency, effectiveness and suitability of a process against its intended purpose (Cohen et al, 2013; Kusters, 2011). In health professions education, the practice of evaluation is widespread and can have many different drivers and stakeholders (Cohen et al, 2013; Dolmans et al, 2003). It can be triggered by a particular problem with an aspect or aspects of an educational programme; it can form part of a comprehensive mandatory activity directed by the profession's regulatory or accrediting authority (Coles and Grant, 1985; Iedema et al, 2004); and it can support organisational commitment to continuous quality improvement and the need to identify programme strengths and weaknesses,

informing decisions for future practice (Dolmans et al, 2003). In many cases, an evaluation can serve more than one purpose and address the interests of more than one stakeholder (Edwards, 1991).

As standards guiding judgements of the quality of health professions education become more prevalent, it is clear that concerns go beyond teaching and training activity to include the means by which students and trainees are deemed fit to progress to next stage in their careers (Karle, 2006). The World Federation for Medical Education Global Standards for Quality Improvement (WFME, 2003a; 2003b) and the UK General Medical Council (2010) describe a comprehensive range of criteria against which the quality of medical education should be judged. In addition to the criteria for formative assessment, which is often referred to as “assessment for learning”, there are others specifically aimed at summative assessment, or “assessment of learning”.

The important role evaluation can play in ensuring the quality of summative assessments should not be underestimated, in particular for those examinations that govern entry to a profession, and thus independent practice. Throughout history, there has been a long tradition in the health professions of assessing trainees’ knowledge, skills and attitudes to ensure they have achieved acceptable standards of performance before allowing them to practice independently. In the United Kingdom (UK), early dictates from Medical Royal Colleges include a requirement for those entering their professions to have first been examined and proved (Dingwall, 2005). This would seem appropriate bearing in mind the unique position of public trust afforded to these professions and, in current practice, the responsibility associated with the transition to specialist status and consultant posts (Higgins et al, 2005; Westerman et al, 2010).

There are many different stakeholder groups who have a vested interest in the quality and suitability of summative assessments used for health professions certification (or licensure), including the public, patients, government, regulators, professional bodies, employers, practitioners, examining bodies and examinees (Norcini et al, 2011). Recently, there has been increasing worldwide public and political interest in patient safety and the all-round competence of health professionals and, as a result, calls for greater accountability (Murray et al, 2000; Crossley et al, 2002; Walshe, 2002). In many countries where health professions are regulated, approval processes can be used to govern examining bodies and organisations that fail to meet their published standards.

The GMC (2010) and GDC (2015) both publish standards for postgraduate curricula and assessment, a number of which apply specifically to examinations and describe criteria that must be met. The GMC states that assessments, whether formal or workplace based, must be "chosen on the basis of validity, reliability, feasibility, cost effectiveness, opportunities for feedback, and impact on learning" and organisations must provide an evidence based rationale for their approach (GMC, 2010). The GDC stresses that assessment scores and decisions based on them must be valid and reliable (GDC, 2015). In both cases, requirements for summative assessment extend beyond technical specifications to include aspects such as standard setting, transparency, fairness and examiner recruitment. Curricula and assessment systems are subject to approval and any changes to these must be reviewed and approved by the relevant governing body prior to implementation (GMC, 2010; GDC, 2015).

While patient safety is a major concern for all stakeholder groups, there may well be other drivers of quality. For example, examining and professional bodies may have concerns regarding reputational risk to their organisation or profession should

assessments prove to be unfit for their given purpose. Formal examinations can be costly and resource intensive. Examining bodies, whether funded by government or via candidate fees, are accountable and must be able to justify the costs incurred and demonstrate efficient use of resources. Examinees might look for reassurance that the examinations that they are required to take, and the decisions made on the basis of them, are fair. The option of legal redress through appeal, should this be in doubt, may be an additional concern for the examining bodies. Employers are also likely to want reassurance that assessments provide suitable gateways to clinical practice and cover the relevant knowledge, skills and attitudes needed to provide effective patient care in their context.

With all this in mind, it is easy to understand why organisations responsible for creating and delivering assessments are increasingly looking to ensure that their approaches, instruments and processes have a strong evidence base. Professional requirements demand a quality assessment programme and justification of the approach taken for examination design and delivery in relation to the purpose and consequences of the assessment. The examining organisation or body must therefore be able to demonstrate the efficacy and fairness of their approach and processes. Standards published by the regulatory authorities are often necessarily generic and principle based (GMC, 2010) to ensure their applicability across a wide range of disciplines. However, these may fall short in providing sufficient detail to guide examining bodies in their planning, delivery and evaluation of assessments which are fit for purpose. To provide evidence for the suitability of their assessment activities, and to defend their approach, these examining bodies must not only be cognisant of the regulatory standards but must also look to the published literature on good practice.

Criteria and frameworks for evaluating assessment

A number of authors have highlighted the importance of selecting appropriate criteria for evaluating assessments of clinical competence in line with their intended use (formative or summative) and have proposed frameworks for this purpose (Linn et al, 1991; Messick, 1994; Van der Vleuten, 1996; Baartman et al, 2006; Norcini et al, 2011).

Two related criteria - reliability and validity - have long been used to judge the psychometric rigour of assessments and, as such, feature in many key texts, frameworks and published studies that focus on test evaluation (Anastasi, 1968; van der Vleuten, 1996; Popham, 2000; Boulet and McKinley, 2013; Norcini et al, 2011). Reliability is often said to be a requirement for validity but in itself not enough to ensure it (Linn and Gronlund, 1995). It is a property that can be difficult to conceptualise and describe and has been defined in the literature as the accuracy, consistency or reproducibility of scores (Streiner et al, 2014; Downing, 2004). It refers to the stability of a test score should the test be repeated. Estimates of test score reliability can be calculated in a number of ways and provide an indication of the proportion of error inherent in the score (Linn and Gronlund, 1995; Streiner et al, 2014) or the degree to which candidates' observed scores reflect their true scores (Clauser et al, 2008). It is considered of greater priority in high-stakes summative testing, where poor reliability can compromise pass/fail decisions. In such cases validity could be considered immaterial. Relatively speaking, error and precision of scoring is not as important an issue in formative assessment, where the purpose is to enhance learning, (Norcini et al, 2011), not to make competency decisions which might affect career progression.

Validity is cited as being the most critical of all criteria without which any test, summative or formative, would be rendered worthless (Linn and Gronlund, 1995; AERA,

APA & NCME, 2014). As with reliability, it can be difficult to describe, its conceptualisation and use as a criterion for test evaluation having evolved over many years (Goodwin and Leech, 2003; Cizek, 2012). Earlier definitions focused only on the technical merits of a test and described validity in terms of whether the assessment measured what it was intended to measure (Goodwin and Leech, 2003). In the literature, tests were often reported to exhibit a particular type of validity – criterion, content or construct and, more recently, consequential (Linn and Gronlund, 1995; Messick, 1995).

This fragmented view of validity has been largely replaced in the literature. Validity is now perceived as a "unitary" criterion that can be substantiated by accumulating different types of evidence (Messick, 1989; 1995). Current definitions of validity, although contested by some (Popham, 1997), focus on the technical merits *and* the social consequences of testing, and relate to whether the interpretation of scores from a test justify their intended use (Linn and Gronlund, 1997; AERA, APA & NCME, 2014; Kane, 2006; Norcini et al, 2011). Both Messick (1989) and Kane (1994) have highlighted the abundance of poor quality validation studies in the literature, and the need for more detailed frameworks on which to organise the relevant evidence. Kane (1994; 1996), building on the work of Cronbach (1988), suggests any claims that score interpretations are valid must be supported by a body of evidence. He uses an argument-based approach, which addresses the weak points in assessment design and delivery. He has proposed a new framework for building a validity argument that incorporates the reliability criterion and describes the types of evidence that might be collected to support it (Kane et al, 1999).

Messick (1995) suggests that evidence to support validity be collected in relation to six different aspects of the assessment process, namely; content, substantive, generalizability, external, structural, and consequential. The Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014), while incorporating the ethical or consequential dimension of validity favoured by Messick and Kane, follow more closely the Messick framework for describing how to gather validity evidence rather than adopting Kane's argument-based approach.

More recently, the criteria for evaluating assessments commonly cited in the literature have expanded beyond those of reliability and validity to include greater consideration of the educational effects, educational impact, and pragmatic implications of assessment. While this may reflect a wish to address more comprehensively the current drivers for evaluation, the expanding popularity of assessment systems, and performance and workplace based approaches to assessment, may also be contributing to the emergence of additional evaluative criteria (Linn, Baker and Dunbar, 2001; Gipps, 1994; AoMRC, 2009).

Norcini et al (2011) include "equivalence" as a criterion in their evaluative framework where the AERA, APA and NCME Standards (2014) refer to score comparability. These criteria relate to whether there is justification for using different versions of a test to inform the same decisions and, as such, will have implications for validity. If the tests are not equivalent, then decisions based on the scores may, or may not, indicate who has, or has not, met minimal requirements. Not surprisingly, equivalence has been described as a priority in high-stakes testing.

Fairness as a criterion has been viewed from many different perspectives. It is treated separately by some authors (AERA, APA and NCME, 2014), and incorporated within

validity arguments by others (Linn and Gronlund, 1995). It addresses bias and whether any examinee, or group of examinees, is likely to be unduly disadvantaged by the test design or processes. Test content and response formats are potential sources of systematic error, (AERA, APA and NCME, 2014), as are the raters used in performance assessments, be they examiners, actors or standardized patients (Boulet et al, 2003, Clauser et al, 2008). The processes for selecting and training of raters used in performance assessments can critically affect both test fairness and equivalence, and should be given careful consideration lest validity be compromised (Boulet et al, 2003; Clauser et al, 2008).

Van der Vleuten (1996) included "educational impact" in his framework for judging the utility of an assessment. He emphasised the value of reviewing whether an assessment is driving learning appropriate to the curricular aims and whether feedback informs and encourages future learning. Norcini et al (2011) make similar points but consider these two effects as separate criteria: educational effect relating to how learners prepare for an assessment and catalytic effect relating to the impact of feedback. While educational effects are often said to be of higher priority for formative and programmatic assessment, their importance in summative assessment is gaining ground (Norcini et al, 2011). The UK Academy of Medical Royal Colleges (AoMRC) having consulted its members on the feedback they currently give candidates sitting high-stakes summative assessments, has issued new guidance that aims to drive improvements to catalytic effect (AoMRC, 2015).

Feasibility, as described by Norcini et al (2011), addresses the practicalities of accessing the human, physical and fiscal resources required for an assessment. Organisations responsible for funding assessments may have to account for their expenditure, while

those designing, delivering and evaluating assessments must consider how to fulfil their obligations considering available resources.

Acceptability is a criterion included in a number of frameworks (van der Vleuten, 1996, Norcini et al, 2011). While at first glance it seems to address only the reaction of different stakeholder groups to an assessment, should an aspect or aspects of an assessment prove unacceptable to a key stakeholder group, it could ultimately impact on validity or feasibility (e.g. via perceived worth by employers, its execution by examiners and assessors).

A number of recent studies have suggested a need for additional, or different, criteria to address the newer complex performance based assessments and competency based assessment programmes that incorporate them (Linn et al, 1991; Baartman et al, 2006). Baartman et al. propose a new framework of criteria, rejecting traditional criteria such as validity and reliability as "container concepts", and suggesting terms that might be more clearly defined and interpretable. They include in their proposed framework authenticity, cognitive complexity, fairness, meaningfulness, transparency, reproducibility of decisions, comparability, costs and efficiency and educational consequences.

Some of these criteria might offer a different lens with which to judge the value of an assessment and enable a fuller understanding of traditional criteria. For example, authenticity relates to how well different aspects of the assessment reflect the workplace and could be useful when considering both the validity of scores in relation to their intended use and the educational and catalytic effects of an assessment. Transparency relates to the type and quality of the information on the assessment made available to stakeholders, which could be useful when considering the acceptability,

validity, educational effect and catalytic effect of an assessment (Baartman et al, 2006). Other criteria appear to be more a rebranding of existing well established and understood criteria such as those described by van der Vleuten (1996). For example, the criteria of educational consequences, costs and efficiency, reproducibility of decisions and comparability seems to mirror respectively those of educational impact (educational effect and catalytic effect), feasibility, reliability and equivalence described in the traditional models yet, in most sources, are described in less detail.

Conclusions

Frameworks for evaluating traditional summative examinations

When identifying an appropriate evaluative framework for an assessment, it is widely acknowledged that the underpinning criteria should be appropriate to the purpose and, arguably, also to the type (traditional or workplace based) of assessment (Baartman et al, 2006; Norcini et al, 2011; Cook et al, 2015). A number of the aforementioned frameworks, or combinations, would be applicable to a traditional high stakes summative examination. It is also advisable that the results and recommendations of the evaluation are reported in a manner that can be easily understood by all stakeholders and, in particular, the intended recipient (Coles and Grant, 1985).

The model described by van der Vleuten (1996) fulfils these requirements in that it was proposed for the evaluation of formative or summative assessment, is applicable to traditional styles of testing, and employs terminology that is in common use and well understood in health professions education. It also has the added advantages that it addresses a range of criteria that reflect the priorities of different key stakeholders, and aligns with the guidance and regulatory standards familiar to many UK organisations.

The model described by Norcini et al (2011) treats equivalence as a separate criterion and key consideration in high-stakes assessment. It also highlights educational effect and catalytic effect as separate criteria

The broader framework suggested by van der Vleuten (1996) and expanded by Norcini et al (2011), if used alone, may provide insufficient detail on the technical criteria of reliability and validity to guide full and robust justification for test score use. Many attempts at validating assessments have been rightly criticised for their lack of rigour, where the evidence cited was selective and fragmented, and thus failed to add up to a cohesive argument supporting or opposing their use (Kane et al, 1999; Messick 1994). The six aspect model proposed by Messick could be used to further explore validity evidence, but the similarity of the terminology to the fragmented view of construct, content, criterion as different types of validity may hamper its interpretation and utility.

Kane's framework addresses validity and reliability in relation to an assessment's intended use, and provides a structure against which evidence can be collected. It encompasses, directly or indirectly, the majority of criteria in other frameworks (Table 1) and highlights the interdependence of evidence supporting validity claims, describing underpinning inferences as links within a chain. Since its inception, a number of authors have published reports on the practicalities of applying Kane's framework to a variety of different assessments and assessment programmes, and have suggested the types of evidence that might be used to support each aspect of the argument (Kane et al, 1999; Clauser et al, 2008; Schuwirth and van der Vleuten, 2012; Boulet and McKinley, 2013; Cook et al, 2015; Hatala et al, 2015).

Gathering evidence to support the quality of high stakes formal examinations

If judgements are to be made concerning a trainee's readiness for professional practice, and these judgements are based on the scores or outcomes from a test, the inferences made on the basis of the test design and delivery must be supportable (Kane, 1994). Kane's framework allows evidence to be collected to support each of four inferences required to justify a claim of validity. The categories of evidence are scoring, generalisation, extrapolation and interpretation (Kane et al, 1999; Boulet and McKinley, 2013; Cook et al, 2015). The framework can also be used to explore the weaknesses in exam design and delivery. According to Cronbach (1980), validation is about finding flaws in the interpretative argument and, as such, evaluation should also consider the weaknesses - obstacles and threats to validity - of the assessment in addition to the strengths (Crooks et al, 1996).

Potential issues with assessment design and delivery that could compromise validity, as distilled from all the aforementioned frameworks, are summarised in Table 2. Aspects that should be scrutinized to support or challenge each inference are described below. The level of standardisation of each stand-alone component, the alignment of scoring rubrics to the trait being measured, the processes used for recording and calculating scores, the examination environment, item and station quality, examiner selection processes, examiner turnover, actor training (for OSCEs), question reuse policy, question security, sufficiency of examiner training and inter-rater reliability could be reviewed in relation to the first inference that scores on a test are accurate measures of the trait of interest (scoring).

The second inference, that the sample of items or stations used in a test is sufficient to allow generalizations to the wider domain of all possible items (generalization), could be explored by considering; the length of the test in relation to the heterogeneity of

items, the number of marking events, the sampling plans, blueprints, sources of measurement error, whether candidate submissions can be truly attributed the candidate, the rules whereby good performance in one skill can compensate for poor performance in another, the weighting of items, score variance estimates from different administrations of the examination.

The third inference is that the items used and traits being tested mirror important tasks and skills required in professional practice (extrapolation). If the knowledge, skills and attitudes measured do not capture or transfer to the clinical environment, the assumption that a higher score will be indicative of a greater readiness to practice at the required level cannot be supported. For example, if the degree of standardization is so stringent that a complex task is reduced to a series of simple tasks tackled in isolation, the result may not reflect performance in real practice. Associations between scores on a test and level of expertise, future performance, or other validated measures of the same trait, congruity with curriculum aims and syllabus, and examiner feedback on test stations can be investigated in this regard.

The final inference in the chain is that test scores are interpreted accurately and used appropriately by decision makers, taking into account the imperfections of the processes used to generate them. The standard setting processes, use of the standard error of measurement in establishing a cut-score, the discriminating power of items/stations, and score reporting practices can be used to ascertain the plausibility of this inference. For example, a norm-referenced standard arbitrarily set at a fixed percentage score, would be inappropriate for a competency based assessment as it would not be anchored to any performance criteria. The standard would change relative to the cohort and could allow progression of candidates whose performance falls below

a minimum acceptable standard or hinder progression of candidates whose performance meets or exceeds the required standard (Norcini, 2003; Cizek et al, 2004)

In addition to the criteria of reliability and validity, which can be addressed in detail using Kane's validity framework, van der Vleuten's (1996) utility framework includes educational effect, acceptability and cost or feasibility. A number of investigators describe the use of this framework for evaluating the quality of their assessments and provide examples of the type of evidence collected (Southgate et al, 2001; Yudkowsky et al, 2006; Roberts et al, 2008). The acceptability criterion could be investigated through stakeholder feedback, appeals and complaints received and review of information provided. The advice and information provided to candidates prior to the examination, eligibility and question reuse policies and the extrapolation, scoring and generalization aspects of the validity argument can be considered in relation to educational effect. Catalytic effect would be influenced by score reporting, feedback and resit policies, and could be explored using longitudinal data from candidates resitting the examination. Evaluation of feasibility could take into account availability of resources required to deliver and audit an examination, balanced against the need for ensuring psychometric rigour in relation to its intended purpose.

A number of other investigators, while not subscribing to specific published evaluative frameworks, provide examples of evidence used for validating their licensure or certification examinations. Some take a comprehensive approach to validation, while others focus on providing specific and singular types of evidence that, in themselves, would not constitute a solid case, but could form part of a more thorough investigation (Hutchison et al, 2002). Examinations such as those that confer Membership of the Royal College of General Practitioners (MRCGP), Membership of the Royal College of

Physicians (MRCP), Educational Commission for Foreign Medical Graduates (ECFMG) certification, the United States Medical Licensing Examination (USMLE) and the Medical Council of Canada Qualifying Examination have been the subject of in-depth validation studies (Munro et al, 2000; Tamblyn et al, 2002; Tamblyn et al, 2007; Van Zanten et al, 2003; Van Zanten 2007; McManus et al, 2013). The associated studies have utilised, amongst others, factor analyses to explore content comparison of extreme groups, and correlations with alternative assessments or workplace performance indicators. Each of these approaches to providing evidence can be considered for inclusion within evaluative frameworks.

METHODS

The methods section describes the framework of criteria against which the quality of the MOralSurg (2013) was judged and the rationale for its selection. The information sources that were used to facilitate an evaluation of quality in relation to the chosen framework are detailed.

A framework for evaluating the MOralSurg

The quality of the MOralSurg was evaluated primarily against the validity criteria described by Kane (1994), taking into account wider interpretation of the framework where it overlaps with other models. Two additional criteria (educational impact and acceptability), described by van der Vleuten (1996), were also considered. Taken together, these frameworks encompass the quality criteria relevant to the examination purpose. The structure and terminology used in these frameworks should be familiar to those requesting the evaluation, and be interpretable by the wider stakeholder groups, as they mirror criteria used by the regulators for medical and dental education in the United Kingdom.

The MOralSurg examination is summative in purpose and clearly of high-stakes in that, as part of a wider assessment system, it can have an impact on the applicant's career progression. A successful outcome in the examination is required by UK trainees in the specialty to be eligible for the inclusion on the General Dental Council (GDC) specialist list and entitlement to use the title "specialist". Eligibility to the more senior position of consultant in Oral Surgery within the NHS requires specialist status.

The reliability and validity of MOralSurg assessment scores, and associated decisions, is crucial: reliability, to ensure acceptable levels of measurement accuracy for reproducibility in pass/fail decision making, and meaningful extrapolation of candidates observed scores to true scores; validity, to ensure that the examination was assessing the intended content and constructs, and that score interpretation is justified in the context of career progression. Kane's (1994) framework, taking into account the criteria described by Messick (1995), was considered the most appropriate for evaluation of the examination against these two criteria, allowing judgement to be made on sufficiency of evidence supporting the argument for its use on psychometric and ethical bases.

The inclusion of additional criteria from the van der Vleuten (1996) framework addresses the educational and practical considerations of testing. Educational and catalytic effects, although generally thought to be of lesser importance in summative assessments (Norcini et al, 2011), were of interest with regard to the Specialty Membership Examination in Oral Surgery. The AoMRC, which represents all the Medical Royal Colleges in the UK, has issued new guidance for providing feedback to candidates in summative examinations. The dental faculties of the UK Surgical Royal Colleges will therefore need to consider more carefully the catalytic effect of their assessments. The educational effect may be of interest to those involved in supporting trainees in preparing for the examination, including the deaneries that provide GDC approved training programmes and universities that offer Master's Degree programmes in Oral Surgery.

The acceptability of the examination is likely to be influenced by all of the above criteria and the detail of the information available to all stakeholders on the aims, process and outcomes of the examination. It too was considered an important criterion.

Feasibility will also be factored into any recommendations made to improve examination quality. The joint UK dental faculties are reliant on goodwill and availability of their exam board members and examiner teams to both examine candidates and to develop questions and cases. Examination fees, where possible, are set to cover the overheads for examination development and delivery. As such, the Colleges will be desirous of an approach that ensures psychometric rigour, but is mindful of the financial burden placed on candidates.

Evidence gathering

Information on the examination was gathered from various sources using a number of different approaches, both qualitative and quantitative, as described in Table 3. This included the Guide to Candidates (GtC) and Examination Regulations (ER) retrieved from the websites of the UK Surgical Royal Colleges, the Oral Surgery Curriculum (OSC) (2012) available from the GDC website, marking rubrics (MR), written papers (WP), complaints (C), OSCE circuit scenarios (OCS), feedback letters (FL), educationalists' reports (EdR) examiner feedback (EF), candidate feedback (CF), examiner job descriptions (JD)and data (Da) provided by the Examinations Department of the Royal College of Surgeons of England

Where documentary evidence was unavailable or inappropriate for evidencing a particular aspect of examination design, delivery or outcome, other approaches were used. A number of examination processes relating to the 2013 diets, including delivery of the viva and OSCE components, recording of exam data, examiner training, standard setting, and debriefing sessions were directly observed.

Reliability coefficients were calculated for each of the three components of the examination. Cronbach's alpha was used to estimate the internal consistency of the first component – a selected response written paper (March diet only) ,the combined structured oral and case report components (November diet only) and for the Objective Structured Clinical Examination (OSCE) (November diet only). The standard error of measurement (SEM) was calculated for each of the components using the relevant r-coefficient. Cronbach's alpha and the SEM do not take into account observer variance. The weighted kappa statistic was used to explore inter-observer agreement in the viva/case report component where two examiners were used to mark each station and report. Item exposure was determined by reviewing all written examination papers delivered to date. Where there were at least eight candidates for a component (March 2013 diet), item analysis was performed on the written papers to generate p–values as a measure of question difficulty, point biserial correlation coefficients as a measure of discriminatory power of items and distractor analysis as a measure of response format quality.

Written papers from both the 2012 March and October diets and those from the 2013 March and November diets were analysed to determine the extent to which items had been reused. Written papers from March and November 2013 diets were reviewed to provide an indication of the proportion of recall and application oriented items in accordance with the classification described by Norman et al. (1996) and Case and Swanson (2002). These papers were also scrutinised for the presence of technical flaws that might cue or assist a test wise candidate (e.g. repeated words in stem and correct option, the longest answer being correct) or introduce construct irrelevant difficulty (e.g. negatively phrased question stems).

RESULTS

The results section comprises a detailed description of the examination and its components, including its design, delivery and psychometric properties of the scores.

The MOralSurg Examination 2013

The Purpose

Two diets of the MOralSurg examination were delivered in 2013; the first in March and the second in November. These followed two initial diets run in 2012. The curriculum document, guide to candidates, and exam regulations pertaining to the MOralSurg all provide information as to its intended purpose and its format. These documents are updated and new versions issued as changes are made to the examination.

The Specialty Training Curriculum; Oral Surgery (Specialist Advisory Committee (SAC) in Oral Surgery, 2010) was the active version of the curriculum during 2013. It describes the examination as one of two approaches that would be used for the assessment of core competencies required of a specialist in Oral Surgery; the other being a system of workplace based assessment and appraisal (SAC in Oral Surgery, 2010, pp7-9). No rationale was provided for the use of the two approaches. The syllabus table included as an appendix of the curriculum lists the training objectives and the assessment approach (MOralSurg or WBA) that will be used for each. There is no indication how constituent knowledge, skills or attitudes should be covered by the appointed assessment methods. Both approaches are mapped to objectives for evidence based practice, clinical governance, health promotion, legal issues and biological sciences. Objectives relating to teaching and communication, history taking and examination,

personal health and probity, clinical record keeping, team working, lifelong learning, management and administration, use of information technology, risk management, informed consent and all core clinical competencies and the underpinning knowledge, skills and attitudes for each are mapped exclusively to the workplace based assessment methods and approach. National Health Service based practice, biostatistics, practice management and working with external bodies are mapped exclusively to the MOralSurg.

The October 2012 versions of the MOralSurg Regulations and Guide to Candidates relate to the 2013 examination diets. The regulations describe the purpose of the exam as testing "knowledge and understanding relevant to the practice of a specialist in Oral Surgery" as described in the curriculum learning outcomes. The aim of the examination is described as being to "allow candidates to demonstrate knowledge and understanding of the principles, practice,….. planning and delivery of oral surgery" (RCSEd, RCSEng & RCPSG (2012a, p4).

The candidate guide describes the aim as testing candidates' "range of knowledge… at the level of a specialist practitioner" (RCSEd, RCSEng & RCPSG (2012b, p2). The scope of the examination is described as including the applied sciences relating to Oral Surgery in addition to its practice and principles. The candidate guide includes an examination blueprint. Learning outcomes listed in the blueprint use the terminology to be "familiar with", "knowledge of" or "competent at", and each is assigned an importance level (essential, important, supplementary) and mapped to the assessment method or methods used to address it.

The Format

The Specialty Training Curriculum (2010) describes the *intended* format for the MOralSurg. It proposes use of written papers, vivas and clinical examinations to assess the core competencies outlined in the syllabus. The October 2012 versions of the MOralSurg Regulations and Guide to Candidates provide more detail on the second iteration of the MOralSurg examination used in the 2013 diets. The examination is described as consisting of three conjunctive components: a written paper containing questions in single best answer (SBA) and extending matching (EM) formats; an objective structured clinical examination (OSCE) of 150 minutes duration; and case reports review and viva (Table 4). However, the marking schemes reveal that the case report and vivas were treated as two separate stand-alone components in the March diet. Neither the regulations nor the guide to candidates cover the purpose of the individual components or rationale for their use (RCSEd, RCSEng & RCPSG, 2012a; 2012b). The intended scope of each component is described in the examination blueprint.

While each component must be passed independently before a candidate qualifies for award of MOralSurg, a pass in a component can be carried forward for up to 3 further diets within a timeframe of two years. Case presentations can be submitted for more than one diet (RCSEd, RCSEng & RCPSG, 2012a).

Examination Policies

Administration

For all three components, candidate responses were manually entered into a spreadsheet or recorded in another electronic format by a member of the examinations team. Total score was summed manually for the written and case report components.

A spreadsheet was used to calculate the final OSCE scores. In the November diet, the report from the Board Chair describes all data entry and analysis as being double checked by a second member of the examination team.

Score reporting

Interview with examination administrators revealed that candidate scores for each component are reported as percentages (number of points awarded/ maximum points available x 100) and candidates are informed of the pass/fail outcome for each component. Failing candidates can request more detailed feedback (e.g. marks given against station number and examiner comments) for the case report and OSCE components. No further feedback is available for passing candidates.

Examiner recruitment

Appointment as an examiner requires possession of a GDC recognised primary dental qualification, entry on the specialist list, being of good standing with the regulator and being active in dental training and clinical or academic practice. All examiners were required to attend examiner training before taking up the role and serve a standard 5 year term. No information was available on requirements for examiners to undergo refresher training.

Questions, items and OSCE scenarios are held centrally as documents in electronic (Microsoft Word) format. These are password protected when sent by e-mail to members of the board and examiner group. Candidate feedback was unavailable for review.

Written component

Format and scope

No information is given in the candidate guide or regulations as to time allowed for this component or the number of questions to be included in the paper. Examination scripts and results spreadsheets show that the March paper contained 128 questions (90 SBAs and 38 EMQs) and the November diet 119 questions (73 SBAs and 46 EMQs). The time allowed for the component was 3 hours.

The blueprint indicates this component addresses basic science and clinical knowledge, clinical judgement, operative and technical competence, evidence based practice, and aspects of clinical and professional practice. No item specifications were available to describe the relative proportions of topics, learning outcomes or levels of items (recall and application) to be included in each paper although items were allocated a topic area.

Marking scheme and standard setting

Candidates were awarded one mark for each correct answer. The standard setting session for the March diet of the exam used a modified Angoff approach and was directly observed. Each question was discussed by a panel of eleven examiners and a consensus reached as to the proportion of minimally competent candidates who would answer correctly. Discussion of the minimally competent candidate was not included in the process. An administrator recorded and totalled the estimate for each question to generate the pass mark. Data from previous diets was not included in the process. Several changes were made to the process in a subsequent standard setting exercise (July 2013) which included discussion of the minimally competent candidate, independent estimation of probability by panel members, discussion of the effects of guessing on probability estimates, recording of each estimate and double entry of data. No documentation was available on policies for rounding (candidate marks and passing standard) or managing borderline candidates.

Paper review

SBA questions had 5 options and EMQs had between 7 and 11 options. Master copies of the written papers included a description of the topic area for each question. Review of the content of the 2013 papers revealed that all but 4 of the questions in each paper addressed basic science or clinical knowledge or its application: the November paper included 2 questions on biostatistics and 2 on clinical governance, the March paper 2 on biostatistics and 2 on informed consent.

Review of the paper format revealed that questions and options within each question are routinely presented in the same order as they appear in the question banks. Furthermore, 92% of questions in the March paper used in both previous diets were placed in the first quarter of the paper. Of the 15 questions in the November diet that were used in three previous diets, all are placed within the first quarter of the paper.

Forty-nine (38%) questions (48 SBA; 1 EMQ) used in March 2013 had features in keeping with a true / false format (e.g. heterogeneous options, minimal relevant context in the stem and a lead in which implied that only one option was correct). Twenty-nine (23%) questions (19 SBA; 10 EMQ) had features commensurate with recall level items and 50 (39%) questions (23 SBA; 27EMQ) had features in keeping with application. Overall, this indicates a paper where 48% of items address knowledge recall and 52% knowledge application.

Forty-one (34%) questions (36 SBA; 5 EMQ) from the November paper had features in keeping with a true / false format, a further 24 (20%) questions (17 SBA; 7 EMQ) had features commensurate with recall level items (minimal relevant context in stem), and 54 (45%) questions (20 SBA; 34 EMQ) had features in keeping with an application level

items (multiple relevant pieces of information contained in the stem). Overall, this indicates a paper where 54% of items address knowledge recall and 46% knowledge application. Overall, across both the March and November papers, a greater proportion of SBA questions (52%) than EMQs (7%) followed a true / false format. A greater proportion of EMQs (61%) than SBAs (26%) were written such that they would require application of knowledge.

Item review

Questions from both papers were analysed for technical flaws that might cue or assist a test wise candidate or introduce irrelevant difficulty according to published guidelines (Case and Swanson, 2002). Thirty-five (39%) of the March paper SBA questions included one or more potential cueing issues (e.g. the longest answer was correct, convergence strategy allowed for a paring down of possible options, word/s in the stem were repeated in the options, precision of the correct option differed from the distractors). Five (6%) of the SBAs included elements of construct irrelevant difficulty (e.g. negatively phrased lead-in, used vague terminology). Of the 38 EMQs, there were no apparent cueing flaws but 4 (11%) questions included irrelevant difficulty (e.g. complex option combinations). Twenty-seven (37%) of the November paper SBA questions included one or more potential cueing issues and 1 (1%) included irrelevant difficulty. Of the 46 EMQs, there were no apparent flaws relating to irrelevant difficulty, but 4 (9%) questions had potential cueing flaws.

Delivery and administration

No protocols were available that covered the delivery of this component and it was not observed. However, based on the answer sheets, results spreadsheets, and

educationalist's report it can be ascertained that candidates recorded their answers manually.

Statistical analyses

The written paper for the November diet attracted 3 candidates and, as such, calculation of reliability estimates was not undertaken. While a reliability estimate and standard error of measurement (SEM) were not calculated for the March diet, these were estimated as part of this evaluation. The Cronbach's alpha based on 8 candidates was 0.84 (68% CI: 0.76 - 0.92) and the SEM was 3.23%.

The Board Chair's report on the March diet describes quality assurance as including calculation of p-values and the review of items where p=0.25 or below. As part of this evaluation, corrected item total correlations, reliability coefficients if item were removed, and distractor analysis were also performed. The mean p-value across all items was 0.78. Thirty-seven (28%) of 128 questions included in final score had a p=1, 39 (30%) questions had p=0.875. Twenty-four (26%) SBA questions used and 3 (8%) EMQs had negative corrected item-total correlations ranging from -0.03 to -0.49 and from -0.08 to -0.24 respectively. For the November 2013 diet, only p-values were calculated. The mean p-value across all items was 0.67. Forty-one (34%) of the 119 questions had a p value =1.

Question reuse

No policy governing reuse of questions was available. Four papers from March 2012 to November 2013 were analysed to quantify question exposure to date. In the March 2013 paper, 31 (24%) of the 129 questions had been used in both previous diets, 55 (43%) had been used in 1 previous diet and 43 (33%) were first introduced for the March

diet. In the November paper, 15(13%) of the questions (n=119) had been used in all previous diets, 25 (21%) had been used in two previous diets, 54 (45%) had been used in 1 previous diet and 25 (21%) were new.

Question writing guidance and review

No records were available relating to item writing training received to date by members of the question writing group; a training session delivered in July, 2013 was observed. This session covered: reference to any blueprints, descriptions of recall and application level questions, guidance for writing questions at knowledge application level, avoiding cueing the test wise candidate and introducing irrelevant difficulty. Attendees were given the opportunity to critique, write and review questions and provided with suggested further reading (e.g. Case and Swanson, 2002). It was unclear whether this training was required for all item writers.

All of the 8 candidates for the March 2013 diet passed the written component. The pass mark was 63.0%. The mean score achieved was 77.4% and the score for one candidate fell within 1 SEM of the pass mark, another candidate within 2 SEMs. Two of the three candidates for the November 2013 diet achieved the passing score of 65.7%.

Case report and Viva Component

Format and scope

This component has 2 sections. For the first section, candidates submit four 2000-word reports describing the management of patients they have treated. For three of the reports, the type of case is prescribed, the fourth case is of the candidate's choice. As described in the Guide to Candidates, patient consent and a signed statement from their supervisor confirming their "substantial involvement in the treatment" in each case

must be provided (RCSEd, RCSEng & RCPSG, 2012b, p3). Candidate reports were marked independently by two examiners across 4 domains (Relevance; Information gathering; Understanding; and Structure, Organization and Presentation). Reports are anonymised and copies e-mailed to the examiners. As each examiner pair marks one report type from each candidate, this means each candidate will receive marks from eight examiners in total.

The second section involves a 10 minute structured face-to-face oral examination on each case report using the same examiner pairs charged with marking the report. Examiners mark independently across four domains (Critical appraisal of case presented; Awareness of risks and complications; Familiarity with other treatment; and Evidence-based practice)

The blueprint shows the case report component as addressing basic science and clinical knowledge, clinical judgement, operative and technical competence; competence to complete oral examination; evidence based practice, "attitudes to clinical and professional practice" (RCSEd, RCSEng & RCPSG, 2012b, appendix A).

Marking schemes and standard setting

An interview with examination administrators and examiners, combined with scrutiny of marking rubrics, highlighted changes made to the marking scheme and standard after the March and prior to the November diets. For the March diet, each domain in the case report and viva sections could be given a mark of 0,2,3,4 or 5 corresponding to the descriptors "well below standard/information absent", "just below standard"", "meets standard" "exceeds standard" and "outstanding". Marks were averaged across examiners, reports and domains. The passing standard was 3. The case report and viva

sections had to be passed independently. Generic performance indicators were provided for each domain and level.

For the November diet, the domains were retained from the March diet but a 1-4 scale was adopted. The descriptors and performance indicators were rewritten to achieve alignment across the scale, and anchored around the description of the minimally competent candidate. The passing standard was changed to 2.5. Marks were averaged across examiners, reports, domains and sections (case report review and viva). These changes were made in consultation with the examination board and complied with current regulations and guidance to candidates.

Delivery and administration

Information on the delivery of the case report section of the component was derived through interviews with the examination administrators. Case reports were submitted no less than 4 weeks before the exam date, anonymised and sent by e-mail to the examiners in advance of the examination delivery date along with the marking rubric. No calibration session or discussion was undertaken by the examiner group. Anti-plagiarism software was not utilised. Marks were made available to the examiners prior to the viva section.

Information on the delivery and administration of the viva section of the November diet was obtained through observation. No protocols were available. Examiners were briefed by the Board Chair just prior to the viva. Arrangements for rotations, working effectively in pairs, recording of marks, the importance of marking independently, and good practice in conducting a viva were covered. Candidates were briefed by the examination administration team and then by the Exam Board Chair.

Examiners were trained on the new marking scheme. They discussed minimal competence and the performance expected for each domain at each level of the rubric. Examiners were provided with generic questions for each domain but were allowed to ask their own questions based on their knowledge of a specific case. Each examiner pair was observed with at least one candidate. Examiners shared the questioning task, each having approximately five minutes with each candidate. Five minutes were allowed after each viva for the candidate to move to the next station and for examiners to record their marks. Examiners later reported that they found 5 minutes insufficient to cover the questions adequately. Examiners did occasionally seek clarification on a candidate's answer from their co-examiner but from the vivas observed there was no evidence of examiners discussing their judgment on candidates' performances or the marks they awarded. Open and closed questions were used appropriately. The venue for the viva, while spacious, had poor acoustics and noise levels were high.

Statistical analyses

Reliability coefficients and kappa statistics were not routinely calculated for this component, but analysis of the November diet was undertaken to inform this evaluation. The Cronbach's alpha across the 4 case reports and the 4 viva scores was 0.59, once pairs of examiner scores and domain scores were averaged to avoid inflation of the coefficient.

Weighted Cohen's kappa values were calculated as a measure of inter-rater reliability for both case report and viva scores (Table 5). This value is expressed on a -1 to 1 scale. According to Landis and Koch (1977) the values can be interpreted as follows: below 0 =poor agreement, 0-0.2 =slight agreement, 0.2-0.4=fair agreement, 0.4-0.6=moderate agreement, 0.6-0.8=substantial agreement and >0.8=almost perfect agreement.

The overall kappa value for the case reports was negative (-0.25) indicating poor agreement. Domains 1 (Relevance of the chosen case) and 4 (Structure, Organisation and Presentation) had the lowest kappa values. The kappa value for the viva element (0.33) indicated fair agreement overall but inter-rater reliability was poor across domain 1 (Critical appraisal), slight across domains 2 (Risks and complications) and 3 (Treatment) and substantial across domain 4 (Evidence).

The kappa statistic for examiner pairs across the entire component (case reports and vivas) was -0.08 indicating poor agreement. Values differed for each pairing falling within the range -0.09 to 0.29 commensurate with poor to fair agreement. A matrix (Table 6) summarises the marks assigned by examiner pairs. Of 192 marking events, a mark of 1 was awarded on a total of 13 occasions with no agreement between examiner pairs. A score of 4 was awarded 66 occasions with examiner pairs agreeing on only 3 occasions.

Five of the ten candidates presenting for this component of the March diet were awarded a pass. All 6 candidates for the November diet passed the component with scores ranging from 70.7% to 78.5% (pass mark > 62.5). Three candidates were resitting the examination.

OSCE component

Format and scope

This component is described in the Guide to Candidates, (RCSEd, RCSEng & RCPSG, 2012b), as being of 150 minutes duration. The Board Chair's and educationalist's reports on the March diet describe an OSCE comprising 9 stations: 4 "single" stations of 10 minutes duration (3 minutes for preparation and 7 minutes examining), 4 "double" of

20 minutes duration and one 30 minute “triple” station. Each station employed two examiners. The November diet was observed to consist of 10 stations (8 single stations each of 10 minutes duration and two 30 minute triple stations) and used a single examiner per station. No item specifications were available to describe the relative proportions of topics, learning outcomes or skills to be covered in the OSCE component of each diet

Marking schemes and standard setting

A checklist based approach was used for scoring in both 2013 diets. The score sheets for the March diet allowed for both the examiner and the patient actor to provide a global score on a 1-4 scale aligned to the descriptors fail, borderline fail, borderline pass, and pass. The educationalist report from this diet explains that the global scores were intended to be used for standard setting with the borderline regression method. The approach was abandoned in favour of a modified Angoff method for the November diet. The modified Angoff method was conducted as described for the November 2013 written paper but with panel members estimating the score that a minimally competent candidate would be expected to achieve for each station. The average number of checklist items was 33 (range 12-42) for stations used in the March diet and 31 (range 17.5 – 52) for the November diet. None of the stations in the March diet were reused for the November diet. The standards set for the March and November diets were 63.0% and 65.9% respectively.

Delivery and administration

Information on the delivery and administration of this component of the examination is based on the November diet only and was derived through observation and interview

with the examination administrators. No protocols or standard operating procedures were available.

Examiners reviewed their stations against the scenario information prior to delivery and minor amendments were made where required. Examiners were briefed by the Exam Board Chair just prior to the OSCE. Arrangements for rotations, recording of marks and good practice in conducting an OSCE were covered. Examiners were permitted to discuss candidates' performances with the actor for their station but the actor did not contribute directly to the marking. Candidates were briefed by the Board Chair on the format of the exam and the role and conduct required of all participants. The Exam Board Chair provided a briefing for the actors who subsequently discussed their role with the examiner for their station in detail with the actor.

Stations were split across two venues, one of which had poor acoustics and where noise levels were high. No external assessor was present. Feedback was routinely collected from each examiner on the fidelity and content of their station. Examiners reported at the examination debrief session that they considered actor performance to be consistent across candidates.

Review of OSCE cases

Case information for the 19 OSCE stations delivered across both diets was analysed. This included the case author's judgement as to the difficulty (difficult, moderate, easy) and importance (essential, important, supplementary) levels that the scenario was pitched at and their description of the skill (clinical, communication, management and leadership) and topic areas the scenario was designed to assess. Candidate instructions, actor brief, kit list and station set up instructions and marking rubric were also detailed.

One station in the November diet addressed operative skill, one addressed managerial skills and another knowledge recall with the remainder focusing on clinical knowledge, skills and judgment. In the March diet one station addressed technical skills and another clinical knowledge with the remainder focusing on history taking, clinical knowledge, skills and judgment. Fourteen stations were blueprinted as covering communication skills – imparting and/or gathering information. In 4 of these stations, marks were awarded for how the information was communicated (e.g. avoiding technical language, displayed empathy) while in the remainder marks were awarded only for checklist items addressing the content of the communication (e.g. correct diagnosis, appropriate treatment plan, asking the patient appropriate questions).

Statistical analyses and outcome

Estimates of reliability and station metrics were not routinely generated for 2013 diets. However, reliability estimates were calculated from the November diet data to inform the evaluation. Cronbach's alpha across all stations based on raw scores r= -0.23, and giving each equal weighting (by calculating a percentage score per station) resulted in r= -0.65.

Eight candidates presented for the March diet, three of whom passed. All six candidates for the November diet (three of whom were resitting the component) achieved the pass standard with marks ranging from 67.1% to 74.2%.

DISCUSSION

This section comprises an analysis of examination quality illustrated against Kane's validity framework and two additional criteria (educational impact, and acceptability) from the utility framework described by van der Vleuten (1996). Evidence supporting the quality of the examination for each criterion and assumption (scoring, generalisation, extrapolation and implication) are highlighted and aspects requiring further development identified.

Examination purpose

An understanding of the purpose of an examination is required before judgement can be made as to the quality of its design and delivery. While the role of the MOralSurg in managing progression decisions in training is made clear in the curriculum document, there is some ambiguity as to its purpose. The regulations and candidate guide describe the exam as testing knowledge and its application relating to the specialty, yet the blueprint suggests that it can be used to assess technical, operative, managerial, research and interpersonal skills. The syllabus maps the exam to the basic sciences, clinical governance, research, professional practice and evidence based practice.

Scoring

For this assumption to be met, the process by which an observation of a trait or skill is converted to a score, and the methodology, must be justified and quality assured. In support of this assumption, the MOralSurg written component which employs questions in true / false, SBA and EMQ formats could be deemed suitable for assessing knowledge and its application. It has been reported that while true / false type questions lend

themselves to factual recall, context rich and well written one-best answer questions are more appropriate for knowledge application (Norman et al, 1996; Case and Swanson, 2002; Schuwirth, 2004). The scoring for selected response papers is not subject to rater variance which should promote standardization across candidates. The post-hoc practice of reviewing, and removing flawed questions or those with a low p value (<0.25) should increase item quality.

For the case report component, the calibration exercise for the viva will aid consistency between examiners. The more favourable kappa coefficients in the viva component compared with the report component may be in part attributable to the calibration exercise. A structured viva approach has been used successfully elsewhere for measuring clinical reasoning (Anastakis et al, 1991; Wakeford et al, 1995). Use of anti-plagiarism software would provide some reassurance of case report attribution. The decision to discard the previous complex and unconventional marking rules and move to better aligned descriptors should be beneficial in promoting standardisation and justifying score interpretation (Streiner et al, 2014).

For the OSCE, actor training and the detailed brief in the scenario information should promote reproducible and appropriate performances (Boulet et al, 2003). If the practice of using different cases for new diets is confirmed as policy, this should allay concerns about station exposure (Whelan et al, 2005). Routine examiner feedback on station performance will be a useful quality assurance measure.

The current policy for double checking data following transfer should go some way to reducing the likelihood of transcription errors in all components but may not be sufficient to ensure the integrity of the data. The examiner selection process and provision of thorough examiner training covering item writing and review, examining in

structured orals and OSCEs should promote item quality and a standardised approach to exam delivery (Downing and Haladyna, 1997).

Threats to the scoring assumption for the written component include the degree of question reuse and the paper format which could cue candidates and promote learning to the test (Haladyna et al, 2002). Around a third of the questions used in the MOralSurg had a facility of 1. While these questions may cover important areas of the syllabus and thus be worthy of inclusion in the paper the high facility may be also be a consequence of question over exposure or bank security. Written formats would not be suitable for assessing technical and operative competence as described in the blueprint. Technical flaws identified could compromise validity (Downing and Haladyna, 1997). The ability to track question edits, their exposure and assigned categories is limited by the current question banking process.

If the aim is to assess clinical reasoning, having 50% of the marks available for the case component awarded for choice of case and structure and format of the report, and full compensation across domains, is problematic. The complexity and nature of the cases chosen as stimulus material is governed largely by the candidates and can differ from candidate to candidate. Should a candidate choose an inappropriate case, it would be reasonable to assume that the marks assigned for all domains are not really relevant, and thus validity might be compromised. It is unclear the degree to which case reports reflect candidates' own work and this is difficult to verify with any degree of certainty. This case report method would not be first choice for addressing technical and operative competency, a role for this component suggested in the blueprint (Miller, 1990; Wass et al, 2001). There are no domains in the marking scheme to address these areas. The lack of a calibration exercise for the case report element, and the poor kappa coefficients,

would suggest different raters are applying different criteria, weighting or focusing on different aspects of performance. The majority of scores given are concentrated in the middle of the scale. An expanded scale, if appropriately annotated could increase discrimination and reliability (Streiner et al, 2014). Examiner training could include benchmarking exercises using case reports from previous diets and video recording of viva sessions and examiners could be given feedback on their rating behaviours (Raymond and Reid, 2001)

For the OSCE component there was incongruity between checklist items and what stations purported to measure. The use of a checklist based approach for candidates at an advanced stage in training is inappropriate. It has been said to risk atomisation of skills (Norman et al, 1991; Schuwirth and van der Vleuten, 2003) and the award of marks for trivial tasks, thus compromising validity. Station metrics (e.g. station total correlations, coefficients of determination, reliability if station was removed, failure rate) could be routinely used to improve station quality although with small candidate cohorts some metrics could be misleading (Pell et al, 2010).

Across all components, the absence of records as to what training has been delivered and to whom, could result in inadequately prepared examiners and potentially biased results. No pilot studies were undertaken prior to the first delivery of the examination. The venue used for both performance assessments, (viva and OSCE), was problematic with noise levels potentially distracting candidates. In the OSCE, the possibility of overhearing questions in forthcoming stations when an existing station had been completed could also compromise validity.

Generalization

For this assumption to be met, a representative, and sufficiently large, sample of assessment tasks must be drawn from the universe of possible tasks to ensure a stable score. In support of this, the number of items in the MOralSurg written paper exceeds the minimum suggested by Norman et al (1996) and is broadly in keeping with the 3-4 hour testing time suggested by van der Vleuten and Schuwirth (2005) to achieve acceptable reliability. The Cronbach's alpha of 0.84 exceeds the threshold of 0.8 often quoted in the literature as being the minimum required for a high-stakes written exam. That said, with small cohorts and criterion referenced exams, the SEM can be a more appropriate statistic for considering stability (Norcini, 1999).

For the case reports, testing time can be difficult to estimate and compared against suggested thresholds. Here, content specificity may well be an issue. Norcini et al (1990) describes achieving an r = 0.63 - 0.73 with twelve 15 minute essays and using three physician raters. Both elements of this component use the same 4 case reports and two raters per case, with poor inter-rater reliability. Thus, the extent of sampling across two major sources of measurement error (rater and case) may well be insufficient. This is borne out by the Cronbach's alpha for this component (0.59) although estimates will be subject to error due to the small cohort size (n=6).

Nine and ten stations were used in the OSCE component for the two diets and the examination time was between 2 and 2.5 hours. The skills and topics covered were diverse and as with the written component, there was no item specification to guide sampling. Cronbach's alpha was negative although the small cohort size may have been an issue.

Extrapolation

This assumption requires that a candidate's performance in the assessment accurately reflects that in the workplace. Written formats and structured orals underperform in this regard as the testing conditions lack authenticity (Hawkins and Swanson, 2008). Questions pitched at knowledge recall, including those with true/false formats, do not replicate real world tasks. Where papers consist of questions which are context rich and designed to simulate real clinical tasks, the latter concerns can be countered (van der Vleuten, 1996; Schuwirth and van der Vleuten, 2003). The written component of MOralSurg includes questions of both types. A strength of the case reports component is that it includes real clinical cases. They are, however, selected by the candidate and may represent what a candidate can best achieve given ample time, rather than a typical performance. The OSCE component, by employing scenarios developed and reviewed by senior clinicians, and involving standardized patients, should accurately reflect the workplace and, as such, support the extrapolation assumption. That said, checklist based scoring may fail to capture the important elements of a task especially where technical and procedural knowledge are not the focus of the component (Regehr et al, 1998). Candidates may adopt test taking strategies that will yield additional marks but not accurately reflect the important aspects of what they would do in practice. There have been no studies undertaken to provide empirical evidence of the association between MOralSurg examination scores and performance in the workplace, stage of training or other tests of competence. While studies aimed at providing such evidence can be time-consuming and logistically challenging, examples of good practice exist and could be emulated (Tamblyn et al, 2002; 2007).

Interpretation

This assumption requires evidence of appropriate interpretation and use of examination scores. The use of established criterion-referenced standard setting methods and processes across all components of the MOralSurg supports this assumption. Use of a modified Angoff, rather than a BLR approach, for the OSCE as originally planned is appropriate bearing in mind the number of candidates required to generate a stable standard using BLR. Anchoring the rubric for the case reports on the concept of the minimally competent candidate should help support pass/fail decisions. Use of a contrasting group method standard setting process could further improve rigour by ensuring cut-score decisions were independent of judgements on individual candidate performance.

While an excess number of conjunctive components can increase the failure rate for an examination, allowing compensation across diverse skills and topics can render a score or outcome difficult to interpret (Haladyna et al, 1999; van der Vleuten and Schuwirth, 2005). The November 2013 diet of the MOralSurg had only three components but did allow compensation across broad topic and skill areas. For example, the written paper includes small numbers of questions on biostatistics and legislative knowledge. If these are important areas of practice, they are underrepresented and might be better assessed elsewhere (e.g. within the Intercollegiate Surgical Fellowship Examination or using WBA). At present, with no topic-level feedback to candidates, it is not clear what the inclusion of these questions can achieve. The case report component allows compensation between written communication skills/choice of case and clinical reasoning. Similarly, the OSCE includes stations aimed at assessing clinical, managerial,

technical, communication and operative skills. This can confound or limit interpretation of candidates' scores.

Although low candidate numbers can lead to error prone reliability estimates and thus error prone SEM estimates, the SEM was calculated for the written component of the March diet. This allowed confidence intervals to be generated around candidate scores and revealed that one candidate's score fell within two SEMs, another candidate within 1 SEM. Pass/fail classification errors could be minimised by increasing the number of questions, including a greater proportion of items with a high discrimination index.

The absence of any limit on the number of attempts at the examination that a candidate can make could theoretically increase the probability of a false positive outcome (passing a failing candidate) by chance (McManus et al, 2012).

Reporting of scores from the case report component as a percentage is potentially misleading being that a 1-4 scale is used for the marking scheme and thus a score below 25% is not possible. Clarification of the purpose and scope of each component and publication of item specifications should help with score interpretation. Rules for rounding scores and passing standards are not clear, and could lead to pass/fail classification errors.

Educational Impact

Educational impact can be explored from the perspective of what candidates do as they prepare to take the examination and the catalytic effect as how they might respond to the outcome. The guide to candidates provided on the College websites includes information on duration and format that should assist candidates in their preparation for the examination, although sample questions are not included. The use of true / false

and recall level one best answer type questions is likely to encourage rote learning and cramming when learning over time is more desirable. Using case reports for a high-stakes examination may encourage undue focus on perfecting four cases to the detriment of other areas of the curriculum. Ambiguity between candidate guide, regulations and curriculum could confuse candidates seeking information to direct and focus their studies.

The more detailed feedback available to failing candidates, although constrained in the interests of item security, can provide valuable information on candidate strengths and weaknesses. Similar feedback could also be of use to passing candidates. Effective blueprinting could allow more detail to be given on the broad topic areas where marks were lost and gained, and inform future study but this should be balanced against the need for test security. Domain level scores or sub-scores could be calculated but advice on their interpretation would need to be provided, especially if reliability was low.

Acceptability

The acceptability of the examination to stakeholders could be compromised by failure to address any of the issues arising from its design, delivery and quality assurance as described above, but also in relation to how the exam is perceived and its feasibility. The anonymised candidate feedback routinely collected for the MOralSurg should help in identifying issues from candidates' perspectives. Aspects of the case report component may compromise acceptability. Cases submitted for the report are prepared over many months, sometimes years. A fail outcome, where a candidate has chosen an inappropriate case, would require preparation of a new case report before the exam could be retaken. This could be problematic for candidates who are no longer in training programmes. Candidates resubmitting a case that had been given a passing score could

find that the case is given a failing score in any resit. While this can be acceptable from a psychometric perspective (e.g., some variation in examiner judgements would be expected on a case basis) it could be perceived as unfair by candidates.

RECOMMENDATIONS

The MOralSurg is a relatively new examination which has undergone a number of developments since inception. While there is considerable evidence as to its quality, some additional changes as described below could further improve its utility and defensibility. The governing body should consider these in the context of available resources.

Documents describing the purpose and scope of the exam should be reviewed with a view to improving their alignment and providing a more detailed rationale for the exam design. This would assist the future development of the exam, better inform candidates how to prepare and stakeholders on how to interpret the scores reported.

The written paper should focus on the application of knowledge rather than factual recall. Expanding the use of context-rich questions would better simulate the workplace, cover a broader knowledge base, and address some compensation issues Being that a higher proportion of EMQs than SBAs from the 2013 diets were associated with application of knowledge, a review of item writing practice and training may be warranted to encourage development of more context-rich questions. Questions should be routinely reviewed for technical flaws and performance. Facility data can be pooled from different diets. Discrimination indices could be used, but where based on small diets, caution should be exercised. In this case, data cannot be pooled as each diet uses a different sample but scores from different diets could possibly be equated by taking advantage of item overlap. Investment in question bank software would facilitate question and option randomisation (Haladyna et al, 2002), monitoring of question exposure, and assist with blueprinting. Combined with optical scanning or computer-

based delivery, this could reduce administrative burden and the likelihood of transcription errors.

If a further measure of clinical reasoning that addresses longitudinal management of cases is warranted, then consideration should be given to replacing the case report and viva component (e.g. with a simulated case based viva component, covering more cases). This should improve defensibility on the exam based on scoring, generalization, extrapolation, interpretation, educational impact and acceptability.

For the OSCE component, assessing a narrower range of skills (e.g. restricted to communication skills) and/or increasing the testing time should improve the argument supporting the generalisation assumption and should help with interpretation. One should also consider if some skills might be better assessed elsewhere. For example, operative skill at the level of a specialist can be difficult to recreate *in vitro* but relatively easy to assess in the workplace. Domain-based scoring for the OSCE component could be considered if a greater focus on interpersonal skills is sought. The use of standardized patient scores for communication skill would improve extrapolation evidence. Quieter and better standardized examination venues should be sought.

Scores from the viva and OSCE should be reported on the same scale as that used to score them rather than as a percentage. Use of external assessors and provision of structured feedback for examiners on their performance should be routine. Examiner training could be enhanced by using video footage of candidate performance to benchmark. If the case report element is retained, submissions from previous diets could be used for benchmarking. Use of a standard setting method that separates standard setting decisions from examiner judgments of performance would improve the

scoring and interpretation arguments. Records of data from each standard setting session should be kept and collated real data used to inform the process.

Where the size of cohort allows, reliability coefficients should be calculated. Where examiner pairs are employed, inter-rater agreement should be estimated. Generalizability theory could be useful to explore the sources of measurement error and their contribution to overall variance but cohort size and its effect on error estimates should be taken into account. Corrected station total correlations, station failure rate and r-coefficients if station was to be deleted should be routinely generated and aggregate data used where appropriate (Pell et al, 2010). Records should be kept of training provided and attendance to ensure that all examiners have the necessary knowledge and skills to fulfil their role

Empirical studies should be undertaken to evidence the extrapolation assumption (Downing, 2003). This might involve comparing scores of candidates at different levels of training or those completing appreciably different training programmes. Performance indicators in real practice (e.g. appraisal records, complaints, treatment failure rate) could be used to determine whether the examination scores and outcomes predict future performance. Scores on the MOralSurg could be correlated with those of the Intercollegiate Surgical Fellowship Examination although the skills assessed differ.

Each component should have an item specification, sampling plan or blueprint indicating the relative proportion of items or stations that will be used across all iterations of the examination. This will ensure consistent sampling from diet to diet and provide information of relative importance of topics or skills addressed.

Longitudinal studies could be undertaken to explore the performance of candidates resitting one or more components of the examination. The results could then be used to inform policy relating to the number of allowable attempts.

Table 1. A comparison of criteria comprising published assessment quality frameworks

Kane (1994)	Messick (1995)	van der Vleuten (1996) / Norcini (2011)	Baartman (2006)
Scoring	Structural Substantive	Validity	Fairness Comparability
Generalization	Content Generalization Substantive	Validity Reliability Equivalence	Reproducibility
Extrapolation	External	Validity	Authenticity Cognitive complexity
Interpretation	Consequential	Acceptability	Meaningfulness Directness
		Educational effect Catalytic effect	Educational consequences
		Acceptability	Transparency Fairness
		Cost	Costs and efficiency

Table 2: Threats to validity for each inference comprising Kane's validity framework[a]

Inference	Threats to validity
Scoring	Overexposed item/station bank Insecure item/station bank Poor quality Items/station Lack of or inappropriate scoring rubrics Inappropriate examination venue Lack of/poor examiner training Inappropriate examiner selection processes Rapid examiner turnover Poor quality score recording processes Poor score analysis procedures Lack of/poor standardized patient/actor training Candidates scored on different samples Methodology misaligned to trait
Generalization	Sample of items too small Sample of items unrepresentative of wider domain/s Poor blueprinting Sample of examiners too small Sample not guaranteed to be the candidate's own work Sample inappropriately selected by candidate Inappropriate weighting placed on items
Extrapolation	Exam conditions don't reflect the clinical environment (e.g. time provided, degree of atomisation) Assessment content doesn't capture important aspects of practice Assessment misaligned with curriculum aims
Decision/Interpretation	Excessive number of conjunctive components Policy allowing unlimited attempts at the examination Inappropriate standard setting method chosen Inappropriately trained or selected panel

	Inappropriate use of the standard error of measurement Lack of discrimination around the cut-point Unclear reporting of test outcomes Inappropriate score reporting format

a. The threats listed take into account a broad interpretation of Kane's (1994) validity framework that encompasses the criteria described by Messick (1995) and the work of Clauser (2008), Cook et al (2015) and Hatala (2015) in applying Kane's framework to assessment tools and approaches.

Table 3: Information sought and sources used to evaluate assessment quality against framework criteria[a]

Criterion	Information sourced via:		
	Documentation	**Observation (O)/ Discussion (D)**	**Analysis**
General	Exam purpose (GtC, ER, OSC) Component purpose (GtC, ER)		
Scoring	Item exposure (WP, OCS) Item quality/ flaws (WP) Rubric quality (MR) Examiner selection criteria (JD) Examiner turnover (Da) Score calculation (Da) Sampling frames (GtC, ER) Exam administration (EdR)	Exam venue (O) Exam administration (O, D) Candidate briefing (O) Actor briefing (O) Actor performance (O,D) Examiner training (O) Calibration session (O)	Inter-rater reliability (kappa) Item analysis
Generalization	Duration of component (GtC, ER) Number of marking events (Da) Sampling policy - content (Da) (EdR) Sampling policy – examiners (Da) Item weighting (Da)	Sampling policy content (D)	R-coefficients
Extrapolation	Congruity of item content with professional practice (GtC, OSC, WP, OCS, EF) Congruity of exam format with professional practice (GtC, ER)	Item writing training (O) Scenario review (O) Item review (O)	

Decision/Interpretation	Compensation across components (GtC, ER) Candidate feedback format (CF) Standard setting panel – number (Da) Standard setting calculation (Da) (EdR) Information for stakeholders (ER) Resit policy (ER)	Standard setting method (O) Standard setting policy (D) Standard setting process (O) Standard setting training (O) Candidate feedback policy (D)	
Educational Impact	Information for candidates (GtC) Item format (WP) Resit policy (ER) Feedback format (FL)		
Acceptability	Candidate (CF) Other stakeholders (C)		

a. The information sought is informed by a broad interpretation of Kane's (1994) validity framework that encompasses the criteria described by Messick (1995) and takes into account appropriate sources of evidence described by Clauser (2008), Cook et al (2015) and Hatala (2015) in applying Kane's framework.

Table 4: Description of formats for each component of the MOralSurg (2013)

	MOralSurg Component	
	March 2013	**November 2013**
1	Written paper (SBA, EMI)	Written paper (SBA, EMI)
2a	Case report reviews	Case report reviews and vivas
2b	Viva on case reports	
3	OSCE	OSCE

Table 5: Kappa coefficients for examiner ratings by domain relating to the case review and viva component of the MOralSurg (November, 2013)

Section	Domain	Kappa
Report	Relevance	-0.39
	Information gathering	-0.13
	Understanding	0.09
	Presentation	-0.45
Viva	Critical appraisal	0.14
	Risks and complications	0.11
	Treatment	0.13
	Evidence	0.69

Table 6: Cross-tabulation of scores from examiner pairs relating to the case review and viva component of the MOralSurg (November, 2013)

		Examiner 2				
		1	2	3	4	Total
Examiner 1	1	0	0	1	0	1
	2	0	15	18	12	45
	3	6	14	75	32	127
	4	6	1	9	3	19
	Total	12	30	103	47	192

REFERENCES

Academy of Medical Royal Colleges. (2009) Improving Assessment, [online]. Available: http://www.aomrc.org.uk/committees/academy-assessment-committee.html. [14 Sep, 2015]

Academy of Medical Royal Colleges. (2015). Guidance in standards for candidate feedback: Summative postgraduate medical examinations in the United Kingdom. [online] Available at: http://www.aomrc.org.uk/doc_download/9872-guidance-in-standards-for-candidate-feedback-summative-postgraduate-medical-examinations-in-the-uk.html. [31 Jan, 2016]

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014). *Standards for educational and psychological testing* [7th edition]. Amer Educational Research Assn.

Anastakis, D. J., Cohen, R., & Reznick, R. K. (1991). The structured oral examination as a method for assessing surgical residents. *The American journal of surgery*, *162*(1), 67-70.

Anastasi, A. (1968). *Psychological testing*. Oxford. England. MacMillan.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, *32*(2), 153-170.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality assurance methods for performance-based assessments. *Advances in Health Sciences Education*, *8*(1), 27-47.

Boulet J.R., and McKinley, D. (2013). Criteria for a Good Assessment. In McGaghie, W.C. ed. *International Best Practices for Evaluation in the Health Professions*. London, England. Radcliffe.

British Medical Association Central Consultants and Specialists Committee. (2008). The Role of the Consultant. Available at http://bma.org.uk/developing-your-career/medical-student/the-role-of-the-doctor/role-of-the-consultant [Dec, 2015]

Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational measurement: issues and practice*, *23*(4), 31-31.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological methods*, *17*(1), 31.

Clauser, B.E., Margolis, M.J., and Swanson, D.B. (2008). Issues of validity and reliability for assessments in medical education. In: Holmboe, E.S, Hawkins, R.E, eds. *Practical Guide to the Evaluation of Clinical Competence.* Philadelphia, PA: Mosby/Elsevier.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*. Routledge.

Coles, C. R., & Grant, J. G. (1985). Curriculum evaluation in medical and health-care education. *Medical Education*, *19*(5), 405.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical education*, *49*(6), 560-575.

COPDEND. The Dental Gold Guide: A Reference Guide for Postgraduate Dental Specialty Training in the UK, June 2013 [online]. Available: http://www.copdend.org//data/files [6 Sep 2015]

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New directions for testing and measurement*, *5*(1), 99-108.

Cronbach, L. J. (1988). Five perspectives on validity argument, in: H. Wainer & H.I. Braun (Eds) *Test validity*, pp. 3-17. Hillsdale, NJ. Erlbaum.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in education*, *3*(3), 265-286.

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical education*, *36*(9), 800-804.

Dingwall, H. M. (2005). *A Famous and Flourishing Society: The History of the Royal College of Surgeons of Edinburgh, 1505-2005*. Edinburgh University Press.

Dolmans, D. H., Wolfhagen, H. A., & Scherpbier, A. J. (2003). From quality assurance to total quality management: how can quality assurance result in continuous improvement in health professions education? *Education for Health (Abingdon, England)*, *16*(2), 210-217.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, *38*(9), 1006-1012.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61-82.

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

Edwards, J. (1991). *Evaluation in adult and further education.* Liverpool, England. Workers Educational Association.

General Dental Council. (2015) Standards for specialty education; standards and requirement for providers. [online]. Available: http://www.gdc-uk.org/dentalprofessionals/specialistlist/documents. [18 Oct 2015]

General Medical Council (2010). Standards for curricula and assessment systems, [online] Available: http://www.gmc-uk.org/education/postgraduate. [16 Oct 2015]

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Psychology Press.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and evaluation in Counselling and Development*, *36*(3), 181-192.

Haladyna, T., & Hess, R. (1999). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment*, *6*(2), 129-153.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333.

Hatala, R., Cook, D. A., Brydges, R., & Hawkins, R. (2015). Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): A systematic review of validity evidence. *Advances in Health Sciences Education*, *20*(5), 1149-1175.

Hawkins, R.E., and Swanson, D.B. (2008). Using written examinations to assess medical knowledge and its application. In: Holmboe, E.S, Hawkins, R.E, eds. *Practical Guide to the Evaluation of Clinical Competence.* Philadelphia, PA: Mosby/Elsevier.

Higgins, R., Gallen, D., & Whiteman, S. (2005). Meeting the non-clinical education and training needs of new consultants. *Postgraduate medical journal*, *81*(958), 519-523.

Hutchinson, L., Aitken, P., & Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical education*, *36*(1), 73-91.

Ibbetson R. (2012). Dental Deans Update - progress on many fronts. Surgeons News, [Online]. Available: http://www.surgeonsnews.com/dental/dental-deans-update [6 Sep 2015]

Ibbetson, R. (2013). The new tri-collegiate specialty dental membership examinations–an update. *Bulletin of The Royal College of Surgeons of England*, *95*(1), 17-17.

Iedema, R., Degeling, P., Braithwaite, J., & Chan, D. K. Y. (2004). Medical education and curriculum reform: putting reform proposals in context. *Med Educ Online*, *9*, 17.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & the Health Professions*, *17*(2), 133-159.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice*, *18*(2), 5-17.

Kane, M. T. (2006). Validity. In R.L. Brennan (Ed), Educational Measurement (4th edition). Washington DC: National Council on Measurement in Education and American Council on Education.

Karle, H. (2006). Global standards and accreditation in medical education: a view from the WFME. *Academic medicine*, *81*(12), S43-S48.

Kusters, C.S.L. et al. (2011). *Making evaluations matter: A practical guide for evaluators.* Centre for Development Innovation, Wageningen University & Research centre, Wageningen, The Netherlands

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher*, *20*(8), 15-21.

Linn, R.L., and Gronlund, N.E. (1995). *Measurement and Assessment in Teaching. (Eighth Edition)*. Upper Saddle River. NJ. Prentice-Hall.

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*(1), 1.

McManus, I. C., Woolf, K., Dacre, J., Paice, E., & Dewberry, C. (2013). The Academic Backbone: longitudinal continuities in educational achievement from secondary school

and medical school to MRCP (UK) and the specialist register in UK medical students and doctors. *BMC medicine*, *11*(1), 1.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, *23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice,* 14, 5-8

Miller, G. E. (1990). The assessment of clinical skills/ competence / performance. *Academic medicine*, *65*(9), S63-7.

Munro, N., Rughani, A., Foulkes, J., Wilson, A., & Neighbour, R. (2000). Assessing validity in written tests of general practice–exploration by factor analysis of candidate response patterns to Paper 1 of the MRCGP examination. *Medical education*, *34*(1), 35-41.

Murray, E., Gruppen, L., Catton, P., Hays, R., & Woolliscroft, J. O. (2000). The accountability of clinical education: its definition and assessment. *Medical Education*, *34*(10), 871-879.

Norcini, J. J., Diserens, D., Day, S. C., Cebul, R. D., Schwartz, J. S., Beck, L. H., ... & Elstein, A. (1990). The scoring and reproducibility of an essay test of clinical judgment. *Academic Medicine*, *65*(9), S41-2.

Norcini Jr, J. J. (1999). Standards and reliability in evaluation: when rules of thumb don't apply. *Academic Medicine*, *74*(10), 1088-90.

Norcini,J.J. (2003), Setting standards on educational tests. Medical Education, 37(5), 464-469.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... & Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical teacher*, *33*(3), 206-214.

Norman, G. R., Vleuten, C. P. M., & Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical education*, *25*(2), 119-126.

Norman, G. R., Swanson, D. B., & Case, S. M. (1996). Conceptual and methodological issues in studies comparing assessment formats. *Teaching and Learning in Medicine: An International Journal*, *8*(4), 208-216.

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics–AMEE guide no. 49. *Medical teacher*, *32*(10), 802-811.

Pierce, J., Kavanagh, P., and Fyfe, D. (2014) Reviewing Regulation of the Specialties. General Dental Council, [Online]. Available: http://www.gdc-uk.org/Aboutus/Thecouncil [6 Sep 2015]

Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational measurement: Issues and practice*, *16*(2), 9-13.

Popham W. J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders.* Needham. MA. Allyn and Bacon.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In: Cizek, G.J. ed. *Setting performance standards:*

Concepts, methods, and perspectives. Laurence Erlbaum Associates. Mahwah, NJ. pp119-157.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, *73*(9), 993-7.

Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., & Tiller, D. (2008). Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Medical education*, *42*(4), 396-404.

Royal College of Surgeons of Edinburgh, Royal College of Surgeons of England and Royal College of Physicians and Surgeons of Glasgow (2012a). Regulations relating to the Tricollegiate Diploma in Oral Surgery (October 2012). [online]. Available: https://www.rcsed.ac.uk/examinations.aspx [January, 2013]

Royal College of Surgeons of Edinburgh, Royal College of Surgeons of England and Royal College of Physicians and Surgeons of Glasgow (2012b).Guidance to Candidates – Tricollegiate Diploma of Membership in Oral Surgery (October 2012). [online]. Available: https://www.rcsed.ac.uk/examinations.aspx [January, 2013]

Schuwirth, L. W., & Van der Vleuten, C. P. (2003). The use of clinical simulations in assessment. *Medical Education*, *37*(s1), 65-71.

Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical Education*, *38*(9), 974-979.

Schuwirth, L. W., & van der Vleuten, C. P. (2012). Programmatic assessment and Kane's validity perspective. *Medical education*, *46*(1), 38-48.

Southgate, L., Cox, J., David, T., Hatch, D., Howes, A., Johnson, N., ... & Turner, J. (2001). The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's Performance Procedures. *Medical Education*, *35*(s1), 2-8.

Specialty Advisory Committee for Oral Surgery (2010). Specialty Training Curriculum: Oral Surgery, May 2010. [Online]. Available: http://www.baos.org.uk/userfiles/files/OralSurgerytraining_2010.pdf. [6 Sept, 2015]

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). *Health measurement scales: a practical guide to their development and use*. Oxford University Press 309-317.

Tamblyn, R., Abrahamowicz, M., Dauphinee, W. D., Hanley, J. A., Norcini, J., Girard, N., ... & Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *JAMA*, *288*(23), 3019-3026.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., Wenghofer, E., Jacques, A., Klass, D., ... & Du Berger, R. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *Jama*, *298*(9), 993-1001.

Van Der Vleuten, C. P. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, *1*(1), 41-67.

Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. *Medical education*, *39*(3), 309-317.

Van Zanten, M., Boulet, J. R., & McKinley, D. W. (2003). Correlates of performance of the ECFMG Clinical Skills Assessment: influences of candidate characteristics on performance. *Academic Medicine*, *78*(10), S72-S74.

Van Zanten, M., Boulet, J. R., McKinley, D. W., DeChamplain, A., & Jobe, A. C. (2007). Assessing the communication and interpersonal skills of graduates of international medical schools as part of the United States Medical Licensing Exam (USMLE) Step 2 Clinical Skills (CS) Exam.*Academic Medicine*, *82*(10), S65-S68.

Wakeford, R., Southgate, L., & Wass, V. (1995). Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP. Royal College of General Practitioners. *BMJ: British Medical Journal*, *311*(7010), 931.

Walshe, K. (2002). The rise of regulation in the NHS. *BMJ: British Medical Journal*, *324*(7343), 96.

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet*, *357*(9260), 945-949.

Westerman, M., Teunissen, P. W., van der Vleuten, C. P., Scherpbier, A. J., Siegert, C. E., van der Lee, N., & Scheele, F. (2010). Understanding the transition from resident to attending physician: A transdisciplinary, qualitative study. *Academic Medicine*, *85*(12), 1914-19.

Whelan, G. P., Boulet, J. R., McKinley, D. W., Norcini, J. J., van Zanten, M., Hambleton, R. K., ... & Peitzman, S. J. (2005). Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA®). *Medical teacher*, *27*(3), 200-206.

WFME. (2003a) World Federation for Medical Education. Basic Medical Education. WFME Global Standards for Quality Improvement. Copenhagen; http://www.wfme.org

WFME (2003b) World Federation for Medical Education. Postgraduate Medical Education. WFME Global Standards for Quality Improvement. Copenhagen: http://www.wfme.org.

Yudkowsky, R., Downing, S. M., & Sandlow, L. J. (2006). Developing an institution-based assessment of resident communication and interpersonal skills. *Academic Medicine*, *81*(12), 1115-1122.