Evaluating the Validity of a National Multi-Assessment System in Postgraduate Surgical Training

By
DARA ANN O'KEEFFE
MB BCh BAO, University College Dublin, 1998
BMedSci, University College Dublin, 1998

THESIS

Submitted as partial fulfilment of the requirements
for the Degree of Master of Health Professions Education
in the Graduate College of the
University of Illinois at Chicago, 2019

Chicago, Illinois

Defense Committee:
Yoon Soo Park, Chair and advisor
Ara Tekian
Oscar Traynor, Royal College of Surgeons in Ireland

I dedicate this thesis to my late father, Peter O'Keeffe (1946-2019) and to my husband Ian and daughters Lily, Grace and Scarlett, whose understanding and support helped make this possible.

AKNOWLEDGEMENTS

I would like to thank my committee members Yoon Soo Park, Ara Tekian and Oscar Traynor for all their help and support in the preparation of this thesis. Particular thanks go to Oscar Traynor and Paula Mansell in the Royal College of Surgeons in Ireland for their assistance with understanding the evolution of our assessment system and gathering the data required for this study.

TABLE OF CONTENTS

CHAPTER PAGE

I. INTRODUCTION..... 1

II. METHODS……. 3
A. Study design……. 3
B. Study setting and subjects……...… 3
C. Methods of assessment data collection……. 4
D. Methods for validity analysis…… 6
1. Methods for Content evidence collection and analysis……. 7
2. Methods for Response Process evidence collection and analysis……. 7
3. Methods for Internal Structure data collection and analysis……. 8
4. Methods for Relations to Other Variables evidence collection and analysis…….. 8
5. Methods for Consequences of Testing evidence collection and analysis…….. 8

III. RESULTS……..…….. 9
A. Content…….…….. 9
B. Response process…… 13
C. Internal structure……. 17
D. Relations to other variables…… 25
E. Consequences of testing…….. 29

IV. DISCUSSION……...……. 31

REFERENCES……..... 34

VITA..... 36

LIST OF TABLES

TABLE PAGE

I - STRUCTURE OF THE MULTI-ASSESSMENT SYSTEM…….. 5

II - DESCRIPTIVE STATISTICS…….……..…… 14

III - COMPONENT WEIGHTING AND RELIABILITY…….……...…. 17

IV - GENERALIZABILITY STUDY: VARIANCE COMPONENTS AND RELIABILITY…….. 22

V - CORRELATION BETWEEN ASSESSMENT MARK AND COMPONENT MARK / CORRELATION BETWEEN COMPONENT MARK AND OVERALL MARK…….…….. 25

VI - CORRELATIONS BETWEEN PERCENTAGE GRADE IN EACH ASSESSMENT……..… 26

LIST OF FIGURES

FIGURE PAGE

1 - Projections in reliability by assessment component: decision study…….. 19

2 - Composite score reliability by weight…… 20

LIST OF ABBREVIATIONS

RCSI	Royal College of Surgeons in Ireland
CST	Core Surgical Training
NSCSC	National Surgical and Clinical Skills Centre
OSCE	Objective Structured Clinical Exam
WBA	Workplace-based Assessments
MMI	Multiple Mini Interview

SUMMARY

Trainees on the National Core Surgical Training program administered by the Royal College of Surgeons in Ireland undergo multiple assessments over the first two years of postgraduate training. These assessments are collated into a high-stakes summative assessment at the end of their second year which determines their progression into higher surgical training in the specialty of their choice.

This study evaluated the validity evidence supporting the use and interpretation of this multi-faceted assessment system during surgical training. This was a national study using data collected over 34 months from two cohorts of the entire population of postgraduate year 1one and two trainees in the Republic of Ireland (N=114). Trainee assessments were categorised as Workplace-based (WBA), Structured assessments performed by the Royal College of Surgeons in Ireland (RCSI) and Multiple Mini Interview (MMI). Validity evidence was examined using Messick's unified validity framework in the following domains: Content, Response process, Internal structure, Relations to other variables and Consequences of testing.

Results revealed that best practice standards for educational testing were adhered to in the majority of steps in the assessment system, providing a large body of validity evidence to support the use of this process. Composite score reliability of the assessment was 0.89 which demonstrates a highly reliable process. Correlation between workplace-based assessments and standardised tests performed in the simulation setting was also very high (0.93).

In conclusion, The Royal College of Surgeons in Ireland's assessment process for Core Surgical Training is statistically highly reliable and is supported by a large body of validity evidence.

I. INTRODUCTION

Surgical training is undergoing a paradigm shift in its evolution from the traditional apprenticeship model towards competency-based education (Sonnadara et al, 2014). This new model is dependent on frequent and rigorous assessments to determine competency in a variety of domains relevant to the practice of surgery. The Royal College of Surgeons in Ireland (RCSI) is responsible for the training of all surgical residents (known locally as trainees) in the Republic of Ireland. As such, we allocate and administer in-hospital training posts and certify all aspects of surgical training and assessment.

In 2014, a new postgraduate program for surgical training was instituted in the Republic of Ireland. In this program, trainees enter at year one, specialize in year two and apply directly for their specialty senior training starting in year three. The two initial years of training represent what is known as ‘Core Surgical Training’ (CST) in Ireland.

The majority of training occurs during hospital-based rotations through clinical experience and feedback from expert faculty trainers. In addition, formal teaching days occur in the National Surgical and Clinical Skills Centre (NSCSC) where technical and non-technical skills are taught using various instructional formats in the classroom, skills laboratory and simulation centre.

As part of this training, we designed and implemented a multi-faceted assessment system over the course of the first two years of their training. Trainees are summatively assessed on multiple occasions using various formats. These include in-hospital assessments, trainer evaluations, logbook data of their operative experience, online course work and objective

structured assessments in technical and non-technical skills. The final assessment is in the form of a Multiple Mini Interview (MMI) (Dore et al, 2010).

This assessment system differs from the common North American system which includes in-service exams of clinical knowledge and workplace-based assessment of various competencies. Our system, administered over the first two years of training, results in a high stakes summative measurement which can have profound consequences on the career path of the trainee surgeon. As these stakes are high, it was important to carry out a comprehensive review of the validity evidence supporting this complex assessment system. To date, there is limited published literature describing a comprehensive review of validity evidence supporting assessment systems in surgery, particularly for assessments carried out during surgical training (Louridas, 2016). Those that do describe assessments during training tend to focus on single assessments and most commonly those that assess technical skills.

We examined the validity evidence of our multi-faceted national assessment system in all five categories of Messick's unified validity framework and present the findings here. This was a national study encompassing assessment data from all surgical trainees in the Republic of Ireland. The full assessment data from two consecutive cohorts of surgical trainees across the first two years of their training was included.

II. METHODS

A. Study design

This study was a prospective validity review of our existing assessment process over a 34-month period from July 2014 to April 2017, spanning the cycle of two consecutive cohorts of trainees completing the Core Surgical Training program in Ireland. Statistical analysis of all assessment results collected over that period formed an integral part of the study. As mentioned above, the conceptual framework utilized for the review was Messick's unified validity framework (Messick, 1989; Kane, 2006). We collected evidence of the five major sources of test validity outlined in this framework: Content, Response process, Internal structure, Relations to other variables and Consequences of testing (Messick, 1989; Downing, Yudkowsky,2009).

B. Study setting and subjects

The study took place in the National Surgical and Clinical Skills Centre in the Royal College of Surgeons in Ireland. The subjects were year 1 and 2 Core Surgical Trainees (CST 1 and 2; N = 114). As the validity of the assessment process as a whole was being studied, we included data from all trainees entering the training program in 2014 and 2015, including those who had not completed all aspects of their training and assessment by 2017. Those with missing data due to missed assessments or early exit from the program were included and missing data was scored as 0 for the purposes of statistical analysis.

C. Methods of assessment data collection

Assessments for trainees on the Irish National Core Surgical Training program are categorized as Workplace-based (WBA), RCSI-based (administered by the academic department) or interview (MMI). Table I shows a detailed description of the breakdown of these assessments.

TABLE I: STRUCTURE OF THE MULTI-ASSESSMENT SYSTEM

Assessment category	**Assessment title**	**Description**	**Number of assessment time-points**
Workplace-based Assessments (WBA)	**Structured Surgical Assessment of Operative Performance** (SSAOP)	Direct observation of trainee by faculty trainer while performing operative procedure in vivo at appropriate level of competence.	9 assessments over 3 rotations
	Structured Clinical Assessment (SCA)	Direct observation of clinical practice (interaction with patient).	9 assessments over 3 rotations
	Logbook of operative experience	Detailed electronic logbook analysis for each of three rotations.	Assessment of case volume at three separate time points
	Trainer reports	Trainee evaluation by senior trainer on completion of clinical rotation.	3 assessments over 3 rotations
RCSI assessments	**Case-based online course work**	Participation in online discussion and analysis of clinical cases. 10 cases per rotation.	3 assessments of 10 cases each
	Technical Skills OSCE	Objective structured technical skills assessment in the skills laboratory using simulated models.	2 multi-station OSCEs, 9 months apart
	Non-technical skills OSCE	Objective structured non-technical skills assessment using simulated patients.	2 multi-station OSCEs, 9 months apart
Interview	**Multiple Mini Interview (MMI)**	Five station MMI: 1. Quality and Safety in Surgical Healthcare 2. Commitment to Academic Advancement and Lifelong Learning 3. Knowledge of Current Issues Relevant to Surgical Practice 4. Decision Making in Surgery 5. Professionalism and Probity in Surgical Practice	5 MMI stations with 2-3 independent raters per station.

The workplace assessments are administered by trainers during clinical rotations and are based on direct observations of patient interactions and operative performance. Locally developed rating tools are used to score these assessments. Summary reports of each trainee's global performance over the course of each rotation are also submitted, along with an operative logbook detailing the trainee's operative experience during that rotation. For the purpose of this review, individual trainee reports including qualitative comments were not reviewed. Only the scores allotted for those assessments were reviewed in an anonymized format with aggregated data for the whole group.

The RCSI assessments consist of online case-based course work and objective structured assessments (OSCEs) of technical and non-technical skills. For the technical skills assessments we use a combination of locally developed task-specific checklists and a validated and widely published global rating tool, Objective Structured Assessment of Technical Skills (OSATS) (Martin et al, 1997) to reduce the variance related to using the checklist format alone. Non-technical skills are assessed in a multi-station OSCE format using a modified Calgary-Cambridge communication skills checklist (Kurz et al, 2003) combined with a locally developed case-specific checklist.

D. Methods for validity analysis

The construct that our validity study is based on is that our assessment process measures the level of competency required for our trainees to progress from year two to year three of the surgical training programme. The most commonly used benchmark for best practice in

educational assessment are the ‘Standards for Educational and Psychological Testing’ produced by the American Educational Research Association *(AERA, APA, & NCME, 2014).*

These are the standards we measured our assessment process against. These standards include best practice recommendations for each of the five areas Messick’s unified validity framework as described below.

1. Methods for Content evidence collection and analysis

Examining the process for the Content section of an assessment does not require traditional statistical analysis. A review of the educational methods used to create the assessment and an analysis of whether best practice *(AERA, APA, & NCME, 2014)* was adhered to in designing the assessment was performed. This included review and appraisal of how the content for each assessment was chosen and how it mapped to the content taught and the level of competency of the learner. We also examined the process of developing the assessment tools and those for training and maintaining standards of the assessors.

2. Methods for Response Process evidence collection and analysis

Response Process analysis involved a critical appraisal of the methods used to gather the assessment data. This included examining the quality assurance process used to ensure data is collected with minimal chance of error; a description of the methods used to combine scores and create the composite score on which the educational decision is made; appraisal of the standard setting process and pass/fail decision rules; and examining the process whereby results are interpreted and reported to the learners.

3. Methods for Internal Structure data collection and analysis

Internal structure was assessed through statistical analysis of the assessment results to examine the reliability of individual assessment types and composite reliability of the assessment as a whole. Numerical results were de-identified and analysed as described below. Individual test and composite reliability were analysed (Stata 14 software, College Station, TX). Generalizability theory was used to further analyse the reliability of this complex measurement system and identify the multiple sources of variability and potential sources of error using the urGENOVA software package (Iowa City, IA). This method has been shown to be appropriate in measuring multiple performance tests delivered across an extended timeframe (Bergus, Kreiter, 2007).

4. Methods for Relations to Other Variables evidence collection and analysis

For this section, we examined the internal correlation of all the individual assessments within our process. A correlation matrix was created to examine the types of assessment that correlate well or poorly with each other.

5. Methods for Consequences of Testing evidence collection and analysis

The final category of validity evidence reviewed examined the consequences of the assessment. This area covers the personal or societal consequences of the results obtained. We appraised the appropriateness of the process and data analysis to the level stakes applied to the assessment. High stakes consequences inform the level of standard setting and statistical analysis required for pass/fail decisions.

III. RESULTS

A. Content

The content of any given test and its relationship to the construct being measured is a vital element in assessment validity (Downing, Yudkowsky, 2009). We examined this for our system by looking at how the content of each test within the assessment system was chosen.

A curriculum map was created based on the syllabus for Core Surgical Training and the assessments were mapped to this curriculum. This allowed us to match the content of the assessment to the curriculum objectives and to the appropriate level of competency. As an internationally designed and recognised examination is a prerequisite to progression within surgical training in the UK and Ireland (the Membership exam of the Royal College of Surgeons or MRCS), we mapped the domains of content covered by this exam and concentrated our assessment system on areas that were relatively underrepresented by the MRCS. This led us to design an assessment based predominantly on skills and performance and less on basic science and clinical knowledge as these domains are extensively represented in the MRCS exam.

Training on our program occurs both in the hospital setting and in the simulation centre, so separate test methodologies were used in each setting. The main competencies identified were broadly grouped into technical proficiency and other non-technical and clinical skills. It was agreed via expert discussion within the committees of the surgical training governance structure, that tests to measure competence within these groups of skills be administered in

both settings. This was based on evidence for best practice in the published literature (Sugden, Aggarwal, 2010). An example of this is that tests of technical proficiency and non-technical skills are measured both in the simulation centre and via observation of operative practice.

Rating tools for the standardised tests administered in OSCE format in the simulation centre were a combination of locally designed task specific checklists and internationally published validated rating tools as described in the methods section above. A scientifically robust process of checklist development is a vital component of test validity and the method of checklist development is often not reported in the literature (Yudkowsky et al, 2014).

Our locally developed technical skills checklists were modified from published checklists of surgical tasks and modified by a minimum of four experienced surgical faculty members. New checklists were then examined for construct validity via blinded testing using videos of novice and expert performance and for inter-rater reliability. This follows best practice standards for checklist development. However, over the years of development of multiple checklists, minor modifications were made in certain checklist items without revalidating the modified checklist and this may have led to potential error in construct under-representation. The alterations made included removal or addition of certain checklist items or changes to improve ambiguous wording for example.

Workplace-based rating tools were used to measure intraoperative performance (SSAOP), clinical consultation skills (SCA) and a global rating of performance in the training post in the

form of an evaluation from the trainee's attending-level trainer. The tools designed to capture scores in these areas were developed locally based on previously published Direct Observation of Practice (DOPS) forms. Although each of the forms was not independently analysed for construct validity before use, we have shown that the workplace-based assessments as whole are extremely reliable (0.80). This will be discussed in more detail below in the results of the internal structure of the assessment.

The area of technical competency was also measured via a scoring rubric applied to the documented real-life operative experience of each trainee as represented by their logbook. The logbook data was collected electronically by trainee self-report. The trainee's supervising surgeon must then validate these entries as a true reflection of the trainee's experience. The algorithm which produced a score from the logbook experience was developed in RCSI using data modelling from historical data of an existing cohort of trainees at a similar level. This data informed a decision study to inform the allocation of scores which adequately represented a fair weighting based on the procedure complexity and the trainee's level of involvement in the surgery. This was analysed and agreed by a team consisting of experienced surgeons and data analytical experts. Results were then reviewed externally by an independent group of experienced surgeons.

Finally, we employed rigorous methods of examiner training and standardisation for the RCSI based assessments. For the workplace-based assessments, on-site training was provided to trainers in the clinical sites on how to complete the rating scales. However, this training was

not mandatory and some trainers may not have received it prior to implementation of the assessments. All content was approved by the Core Surgical Training Committee prior to implementation.

B. Response process

Validity in this category is concerned with how we can ensure that the process for collection and compiling data is as free from error as possible. We utilise a process map checklist developed by our Quality Assurance team in conjunction with faculty and administrators, all of whom have responsibility for verifying that all steps of the process are complete.

At the start of the review period we collected examiner data on paper sheets and the data was transferred to electronic spreadsheets for analysis. Blinded double entry and checks of data were performed which revealed numerous potential areas for data error. We progressed during this time period to electronic entry of all data points to attempt to eliminate many of the potential sources for error, including missing data and errors when data was transcribed. OSCE scores are currently collected using the Qpercom software package (www.qpercom.com) and workplace-based assessments are uploaded directly to our locally developed application, MSurgery. However, we still maintain the quality process checklist to ensure all aspects of potential error are double checked prior to reporting of final scores.

Table II shows the descriptive statistics for the trainee group (n=114). Over the two-year period, each trainee was assessed 24 times in the clinical environment, seven times in the academic department and at five stations during the MMI. The sample size for the MMI was 74. This is because not all of the trainees progressed to the interview stage of the assessment system, either because they withdraw from the program or they had not passed the MRCS examination which is a pre-requisite for progression.

TABLE II. DESCRIPTIVE STATISTICS (N = 114 LEARNERS)

Component	Assessment	Rotation / Station	Mean	SD	Min	Max
Workplace Assessment (390 points)	SSAOP (15 points x 9)	Rotation 1	35.82	4.82	20.4	44.50
		Rotation 2	33.70	9.28	0.0	44.60
		Rotation 3	34.52	10.99	0.0	45.00
		Total	104.03	20.07	20.4	129.60
	SCA (10 points x 9)	Rotation 1	24.49	3.64	6.0	30.00
		Rotation 2	23.25	6.51	0.0	30.00
		Rotation 3	23.48	7.90	0.0	30.00
		Total	71.23	13.81	14.4	88.20
	E logbook (30 points x 3)	Rotation 1	17.14	6.51	0.0	30.00
		Rotation 2	20.74	6.55	0.0	30.00
		Rotation 3	17.00	9.77	0.0	30.00
		Total	54.88	17.49	7.0	86.50
	Trainer Reports (25 points x 3)	Rotation 1	19.46	2.97	12.8	25.00
		Rotation 2	19.36	4.70	0.0	25.00
		Rotation 3	19.33	6.86	0.0	25.00
		Total	58.16	11.19	17.1	74.00
	Total		288.30	55.10	80.7	365.50
RCSI (260 points)	Online Cases (60 points)	Cases 1-10	19.12	1.57	14.0	20.00
		Case 11-20	18.81	2.41	4.0	20.00
		Case 21-30	17.53	5.18	0.0	20.00
		Total	55.46	7.46	18.0	60.00
	TS OSCE (100 points)	OSCE 1	36.50	11.24	0.0	45.98
		OSCE 2	34.55	12.58	0.0	46.19
		Total	71.05	19.00	0.0	90.54
	NTS OSCE (100 points)	OSCE 1	38.13	7.86	0.0	48.38
		OSCE 2	31.26	13.52	0.0	48.26
		Total	69.39	18.48	0.0	93.31
	Total		195.89	39.66	18.0	236.46
MMI (350 points)	Total		145.41	125.05	0.0	317.33

Part of this QA process is to have our standard setting and statistical analysis of OSCE scores managed by an external department in RCSI, the Quality Enhancement Office. They use the Borderline Regression method to determine a cut score for each OSCE station and for the overall exam. Examiners record the task specific score and also make a global judgement of whether the candidate is ‘competent’, ‘borderline’ or ‘not competent’ at the task. Candidate scores for each station are regressed against the global ratings and the general linear model is used to determine the score for each station as the point at which the plotted line intersects the 'Borderline' global rating category. This process results in a variable pass score depending on the performance of the group,

An academic exam board group meets to ratify all OSCE scores before they are included in the final overall assessment score. Pass / fail decision rules in the OSCEs apply only to remediation of failing trainees and as yet do not necessarily prevent progression. However, there is a further pass / fail decision which does prevent progression in the form of a ‘Satisfactory’ or ‘Not Satisfactory’ decision at a Competency Assessment and Performance Appraisal (CAPA) review. Senior faculty review all the assessment data on a biannual basis and meet with trainees to mentor them on their progression and potential areas for remediation. Final results are uploaded to the MSurgery application in the form of a percentage score for each section converted to an equivalent score out of the marks allocated to that test. For example, if a trainee scores 50% in a technical skills OSCE which has a mark allocation of 50, the trainee will receive 25 marks for that test.

Non-OSCE assessments do not have a pass mark assigned as they are measured as an overall score. The number of assessment time-points in each category and the weighting of the various components in the assessments was decided via a panel of experts and ratified at committee level.

Final RCSI scores are uploaded to MSurgery and there they are combined with the WBA scores which are uploaded directly to the application. The result is a composite score out of 650 marks. Trainees have access to these scores online and can review their progress in comparison to the anonymised cohort of their peers over the continuum of their two years of training. Trainees can query any of their scores and discuss them with faculty at their CAPA session if they have a concern over any score they achieved. Academic faculty within the department will also meet trainees individually and pull their full test sheets if they wish to have specific feedback on their performance in any test.

Following this process, final scores are confirmed and the trainees are called for the Multiple Mini Interview in their chosen specialty. Scores obtained at interview are analysed and quality checked before being combined with the pre-interview score. This provides a ranking within each specialty which determines the trainees which will progress to their specialty of choice and continue their training.

C. Internal structure

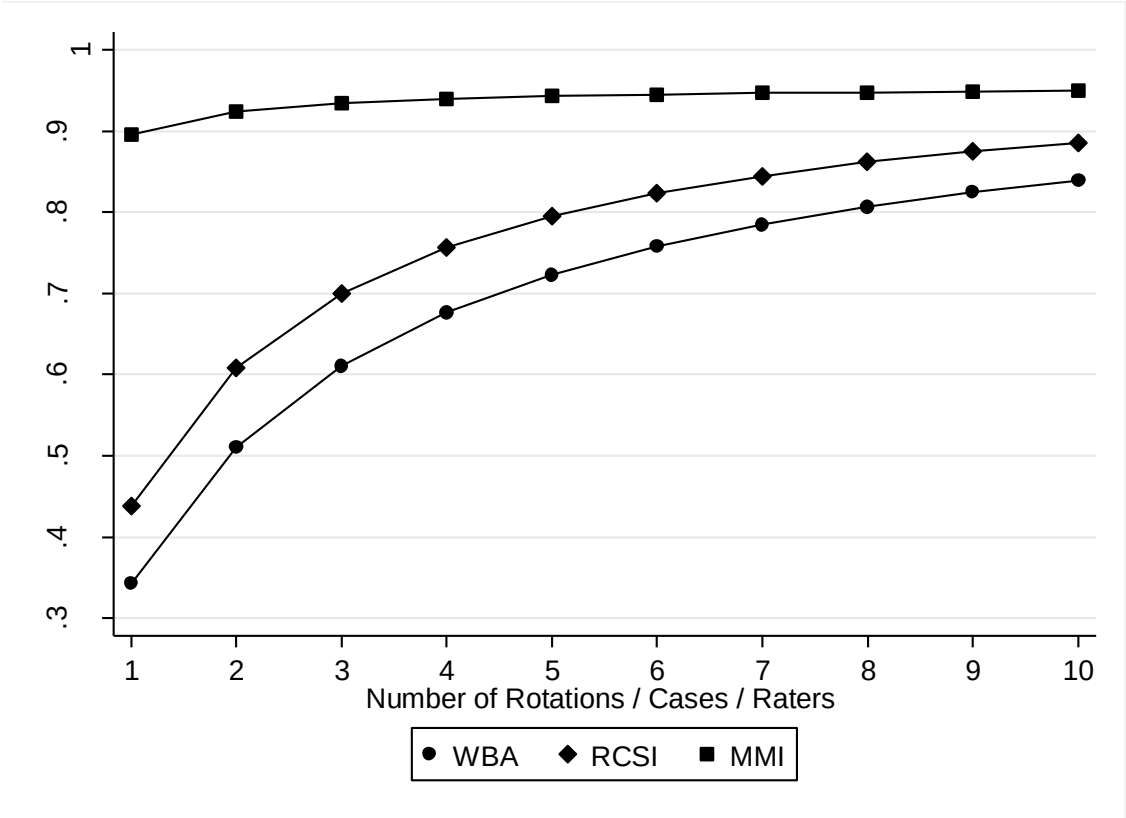
The main component of internal structure that we examined was the reliability. Reliability describes the degree to which we can be confident that the same person undergoing this assessment would get the same result on multiple occasions. Therefore, high reliability equates with a low chance that the results are due to random error (Downing, 2004). Reliability is measured on a scale from 0 to 1, where 1 is the best reliability an assessment can achieve (Kane, 2004). The composite reliability for our assessment system as measured by a Generalizability study is 0.89. This is a high level of reliability suitable for a moderate to high stakes exam. Higher than 0.9 would generally only be expected for licensure type assessments (Downing, Yudkowsky, 2009). The weighting and reliability of each component is listed in Table III.

TABLE III. COMPONENT WEIGHTING AND RELIABILITY

Assessment	Weight	Reliability
Workplace	39%	0.80
RCSI	26%	0.61
MMI	35%	0.69

The weighting of the WA, RCSI and MMI assessments are 39%, 26% and 35% respectively. The weighting of individual components has an effect on overall reliability results (Kreiter et al, 2004). This weighting was initially decided by a panel of experienced surgeons and ratified at committee level. However, we analysed the impact of the weighting and the effect that the structure of the assessment system had on the overall results after we had collected our first year of data. We used this data to perform a Decision (D) study. This analysis demonstrated that increasing the number of assessments in the WBA or RCSI categories would confer a minimal reliability improvement (0.7 to 0.89) in what were already highly reliable assessments. For the MMI, there was no reliability advantage to increasing the number of raters (Figure 1). The D study also showed that the existing weighting allowed for excellent composite score reliability (0.891) and that this reliability did not significantly change if the weighting was adjusted. (Figure 2).

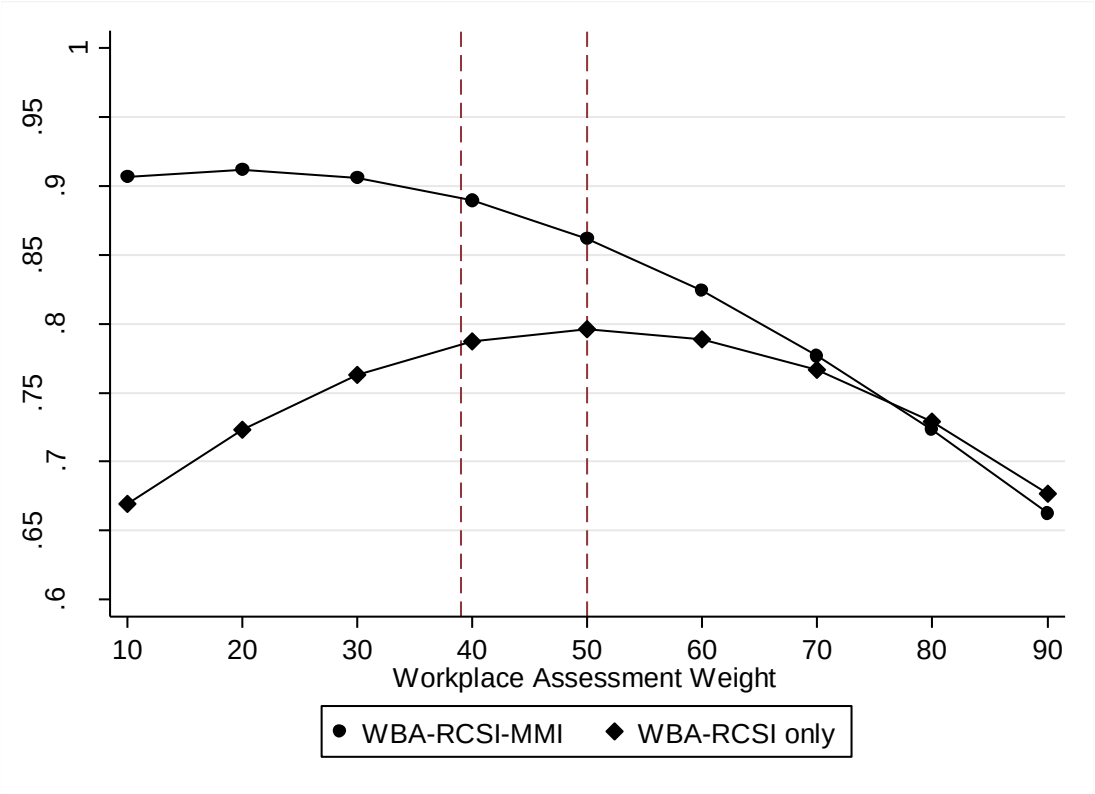
***Figure 1*. Projections in Reliability by Assessment Component: Decision Study**



Note:

1. Workplace assessments (WBA): X-axis refers to the number of rotations
2. RCSI assessments: X-axis refers to the number of cases or components
3. MMI: X-axis refers to the number of raters

***Figure 2*. Composite Score Reliability by Weight**



Note:

1. The "composite score reliability" based on the current weight configuration is .891.
2. The "WBA-RCSI-MMI" used weights based on workplace assessment weight, RCSI assessment weight [= (WBA weight + 10%) / 2], and MMI weight (remaining weight).
3. The "WBA-MMI" used equal weights.

Generalizability Theory (GT) allows us to examine where the variability and potential sources of error lie within this complex assessment. This is particularly useful where there are multiple different assessment formats and multiple independent raters for any one learner and is commonly used in performance exams (Bergus and Kreiter, 2007). The results of a Generalizability analysis performed on our assessment are contained in Table IV.

TABLE IV. GENERALIZABILITY STUDY: VARIANCE COMPONENTS AND RELIABILITY

Component	Effect	Description	VC	%VC	G-Coefficient	Φ-Coefficient
Workplace Assessment	person (*p*)	True variance: how well the assessment discriminates between high and low performers	0.018	30.9%	0.87	0.80
	assessment (*a*)	Difference in difficulty between assessments in this category	0.007	11.8%		
	p x *a*	Assessment specificity: variability in performance of learners across different assessments in this category	0.000	0.0%		
	rater : a	Rater severity: variability in ratings by raters across the same assessment type	0.001	2.1%		
	residual variance	Unexplained variance	0.032	55.2%		
RCSI Assessments	person (*p*)	True variance: how well the assessment discriminates between high and low performers	0.016	25.4%	0.804	0.613
	assessment (*a*)	Difference in difficulty between assessments in this category	0.017	26.3%		
	case (*c*) : *a*	Case variance: variability in cases within the same station	0.003	4.9%		
	p x *a*	Assessment specificity: variability in performance of learners across different assessments in this category	0.000	0.0%		
	Residual variance	Unexplained variance	0.028	43.4%		
MMI	person (*p*)	True variance: how well the assessment discriminates between high and low performers	0.004	22.7%	0.704	0.687
	station (*s*)	Difference in difficulty between stations	0.001	3.6%		
	rater : *s*	Rater severity: variability in ratings by raters within the same station	0.000	1.2%		
	p x *s*	Station specificity: variability in performance of learners across different stations	0.006	35.2%		
	Residual variance	Unexplained variance	0.007	37.4%		

The high percentage variability attributed to the *person* across all three assessment components (30.9%, 25.4% and 22.7%) implies that the assessment system accurately differentiates between trainees regardless of other variables. This is what is known as true variance: how well the assessment discriminates differences between high and low performers.

For the workplace-based component, variability for the *assessment* category was low to moderate (11.8%), indicating that the difficulty across different types of workplace-based assessments was similar. Inter-rater consistency within the same type of assessment was good, as indicated by low variability in the *rater : a* category. There was no perceptible variability in the *p x a* category, indicating that learners' performance did not vary by assessment and were highly correlated.

For the RCSI component, variability for the *assessment* category was high (26.3%) indicating that there was a range of difficulty between the various types of assessment. This would be expected in this category because of the nature of the three assessment types. The online case-based discussions require a basic level of clinical knowledge to achieve a relatively high score, compared to the OSCEs. Also, the scores in the technical skills OSCE tend to be higher across the board than in the non-technical skills OSCE. Case variance within each type of assessment was low however (case (*c*) : *a = 4.9%)*, which means that there was little variability between stations in each OSCE for example.

As seen in the WBA component, there was again no perceptible variability in the *p x a* category. So that, although some types of assessment were more difficult for the group as a whole, those who performed well or poorly in one type did so across all of the RCSIs assessments.

In the MMI component, the high variability (35.2%) in the *person x station* category implies station specificity, i.e. the same trainee may not do as well in all stations. For the MMI, this can be explained by the varied content of each station across the structured interview **(**1. Quality and Safety in Surgical Healthcare; 2. Commitment to Academic Advancement and Lifelong Learning; 3. Knowledge of Current Issues Relevant to Surgical Practice; 4. Decision Making in Surgery; 5. Professionalism and Probity in Surgical Practice). The variability for the *station* category and the rater 'nested' in station (*rater : s*) are both low (3.6% and 1.2% respectively). This indicates that the stations are of similar difficulty and that raters at the same station tend to mark similarly (high inter-rater reliability).

D. Relations to other variables

The relationship between different assessments within our system was statistically analysed and is presented here as correlations, with 1.0 being the highest possible correlation score. Table V shows the correlations between the individual assessment mark and the overall mark for that component. Individual assessments within each of the three components (WBA, RCSI and MMI) correlated highly with the overall mark in that component (range 0.81 to 0.94). Table 5 also outlines the correlation between each component mark and overall mark achieved by the trainee. Of these, the Multiple Mini Interview correlated most highly (0.91) with the overall mark achieved by the trainee.

TABLE V. CORRELATION BETWEEN ASSESSMENT MARK AND COMPONENT MARK / CORRELATION BETWEEN COMPONENT MARK AND OVERALL MARK.

Component	Assessment	Item-Total Correlation
Workplace Assessments	SSAOP	.94
	SCA	.93
	E logbook	.81
	Trainer Reports	.83
RCSI Assessments	Online Cases	.83
	TS OSCE	.90
	NTS OSCE	.89
Workplace Assessments		.79
RCSI Assessments		.78
MMI		.91

Table VI Shows the correlations between all the individual elements of the assessment. This analysis allowed us to look for strong correlations between tests measuring the same or similar constructs (convergent correlation) and also to look for weak or negative correlations between tests that measured different constructs (divergent).

TABLE VI. CORRELATIONS BETWEEN PERCENTAGE GRADE IN EACH ASSESSMENT

	SSAOP	**SCA**	**Logbook**	**Trainer**	**Online**	**TS_OSCE**	**NTS_OSCE**	**MMI**
SSAOP	1.00	0.91	0.61	0.75	0.64	0.70	0.60	0.42
SCA	0.91	1.00	0.63	0.74	0.67	0.77	0.69	0.45
Logbook	0.61	0.63	1.00	0.54	0.61	0.44	0.52	0.45
Trainer	0.75	0.74	0.54	1.00	0.62	0.65	0.63	0.42
Online	0.64	0.67	0.61	0.62	1.00	0.69	0.66	0.41
TS OSCE	0.70	0.77	0.44	0.65	0.69	1.00	0.62	0.43
NTS OSCE	0.60	0.69	0.52	0.63	0.66	0.62	1.00	0.46
MMI	0.42	0.45	0.45	0.42	0.41	0.43	0.46	1.00

One form of convergent correlation was the relation between the SSAOP (measuring operative skill in vivo) and the technical skills OSCE (measuring technical skill in the simulation setting). It was interesting to see that these two assessments correlated well with each other (0.7). This reassured us of the generalizability of the results of the skills assessment in the simulation setting to clinical practice. A similar level of correlation was seen between the SCA (observation of a clinical encounter) and the non-technical skills OSCE (simulated clinical encounter). However, a much stronger correlation was seen between the SCA and the SSAOP (0.91) where we might have expected more divergence. This would imply that highly performing trainees performed well in both of these assessments in the hospital. However, in the simulated equivalent OSCE format, the correlation was lower (0.62). This may be explained by less objectivity in the workplace-based assessments as the assessor knows the trainee and may experience halo or other bias.

The weakest correlations were seen between the MMI and the other individual assessments (range 0.41 to 0.46). This is most likely explained by the interview format being most divergent from the other forms of assessment, but does beg the question of whether the interview is measuring the same construct as the other assessments. Weak correlation in the same range (0.44) was seen between the Logbook score and the technical skills OSCE, perhaps implying that operative experience did not translate to better technical skills as measured in the simulation setting. However, the correlation was somewhat higher when the logbook score was compared to the operative skill as observed in vivo (0.61). This correlational analysis provides a rich source of data showing the internal relationships between assessments within our system. This data should be generalizable to other cohorts of surgical trainees as many training programs employ some or many similar type assessments.

Looking at external correlations with existing measures of surgical competence is a more difficult process. The main convergent assessment in use in the UK and Ireland is the aforementioned MRCS examination. While clinical knowledge is the core competency assessed in the MRCS, elements of technical and non-technical skills are also examined. This is the current benchmark for surgical competency in Europe and is usually taken by trainees in their first or second year of training. Because passing this exam is a pre-requisite to sitting the MMI in our system, we could not objectively compare success in the MRCS exam to high performance in our assessment. All of the trainees who were eligible for the additional points obtained in the MMI would have had to have been successful at the MRCS, so a high correlation would have been inherently biased. However, the lack of analysis of the relationship between our assessment and external data measuring similar constructs (the competency of junior surgical trainees) is a deficit in this study. From 2019 forward, we are collecting actual scores in the MRCS of our trainees and will be correlating them to the scores in our assessment in the future. It would also be desirable to perform correlational analysis with benchmark assessments in other jurisdictions.

E. Consequences of testing

This category of validity evidence can seem more nebulous than the other categories but is nevertheless important. The principle of discussing the consequences of an assessment is to ensure that the testing is justified and at the very least that positive outcomes outweigh negative ones (Downing, Yudkowsky, 2009). The evidence should demonstrate the effect and consequences of the assessment on learners, assessors and potentially other groups within society.

For our assessment system, the consequences of poor scores in this assessment may be career limiting. The competitive nature of the progression to higher training means that some trainees will not progress to their specialty of choice if they are outperformed by the majority of trainees. This is a norm-referenced system occurring after the criterion-referenced competencies are achieved and initial scores are compiled. Those who are not successful will either have to reapply after a year of additional clinical experience or move to another jurisdiction outside of Ireland to apply for higher surgical training there.

In addition to these consequences for the trainee, an inadequate assessment system that does not identify weaker trainees may result in a reduced competence level of surgical trainees entering year three of training. Although they will still be trainees, at that level there will be increased autonomy in dealing with patients in certain clinical circumstances without supervision and this has obvious patient and societal consequences. Because of these personal career consequences and potential societal consequences, this assessment is determined as a moderately high-stakes assessment.

This classification means that there are higher standards for reliability as discussed above. It also impacts the pass/fail decision rules and Standard Error of Measurement criteria. In our assessment, we do not apply absolute pass/fail consequences for any of the individual assessments as it would be difficult to stand over failing a trainee because of one fail in the whole system (1 out of 36 assessments). Instead, fail results as measured by the Borderline Regression method described above, in key assessments such as technical skills are flagged and the trainee is remediated and reassessed. In this cohort, 4 trainees were remediated in year one and 5 in year two. The CAPA process described above is the tool used to identify either outlying areas for development or a pattern of poor performance which may lead to an unsatisfactory CAPA result. Such a result will apply a barrier to progression which ensures a basic level of competency appropriate to the year of training.

For the standardised assessments performed in RCSI, each cohort's results are analysed within two weeks of the assessment and decisions about the application of the Standard Error of Measurement (SEM) are made by a group of senior faculty members. If the reliability of that iteration of the exam is below what is expected or the SEM is unexpectedly high, then the pass mark is adjusted based on the SEM so that a fair representation of the group performance is demonstrated in the numbers passing or failing.

All of these methods are used to ensure a fair process in recognition of the potential prevention of a trainee's progression to higher surgical training.

IV. DISCUSSION

The assessment of clinical competence in any area of healthcare is a complex problem. Surgery presents unique challenges in addition to those faced by other specialties in that procedural competence is added to the broad range of other clinically significant competencies required (Louridas, 2016). Our specialty has historically suffered from a propensity to predominantly measure technical skill competency with lesser emphasis on other essential clinical skills. Now in this era of competency-based progression, we need a wider scope of competency measurement. However, we know that no single assessment can measure the full range of competencies required (Auewarakul, 2005) and so increasingly complex assessments with varied methods used will become the new norm.

There are many references in the literature pertaining to selection of trainees into residency programs (Schaverien, 2016) but most discuss selection directly from medical school or intern year. Our own institution has previously published on processes for selection to higher surgical training specialties (Gallagher, 2014) based on a previous system for progression. The assessment system analysed here occurs earlier along the continuum of surgical training.

The lack of comprehensive psychometric analyses of methods used in surgical training assessments was highlighted recently by Evgeniou et al (Evgeniou et al, 2013). Contemporary validity frameworks such as the one described by Messick (Messick, 1989; Kane, 2006; AERA, APA, & NCME, 2014) tell us that a body of evidence should be gathered in various categories, supporting the level of construct validity of a given assessment. There are examples in the literature of the use of Messick's validity framework in other areas of medical education (Auewarakul et al, 2005), however, contemporary validity frameworks are rarely used in the

surgical training literature (Borgersen, 2018). Most studies published in this area use outdated models of validity or do not describe a conceptual framework at all (Borgersen, 2018). Terms such as face or concomitant validity, for example, are now considered outdated. The systematic review by Borgersen et al, showed the use of contemporary frameworks in only 6.6% of 498 papers on surgical simulation. This was also described in a recent review of assessment in surgical skills (Ghaderi I, et al. 2015). This study revealed that most centres had not validated their tools or system of assessment according to the most up to date frameworks.

Using contemporary frameworks, we have shown here that a large body of validity evidence exists to support the use of our current national surgical training assessment system. This review and analysis also allowed us to identify areas for potential improvement and further study. Going forward, we intend to apply more rigorous standards for individual checklist validation and to examine the relationship of our assessment to other established measures of surgical competence. One limitation of this study is that the review was performed internally. This is the current standard in the published literature with most centers reporting the validity of their own assessments. However, our training program including the assessment system has also undergone external review and accreditation by the Irish Medical Council.

This is the first study to use national data on surgical trainees and to present validity evidence for a high-stakes assessment system during surgical training. Crucially important amongst the validity evidence gathered on this occasion is the high composite reliability achieved. Reliability can be affected by multiple factors but it will always be improved by increasing the number of timepoints at which a learner is assessed and by increasing the number of assessors. Individual assessments within a complex system may have low to moderate reliability, but the most important reliability is the one on which a decision is made. As we all move towards fully

competency-based training, it is important that high stakes competency decisions are never made by one person or one assessment. In that way they will be fully defensible.

Being a procedural specialty has historically predisposed the surgical community to favour measurement of technical skills. It is more difficult to measure surgical performance as a whole than individual competencies and therefore complex assessment systems looking at multi-factorial measurements of performance are less common. We feel the strength in our assessment system is its unique multi-faceted design, using many timepoints, assessors and methods of assessment. However, this design is generalizable to any training program in surgery worldwide as many jurisdictions and training programs already employ some of the individual types of assessment described here. As assessment systems are designed or modified it is essential that appropriate analysis of their validity is performed. There is a need within the education and training community in surgery to embrace formal and transparent review of our assessment processes to ensure an equitable and defensible system of competency assessment for all trainees.

REFERENCES

AERA, APA, & NCME, 2014. The Standards for Educational and Psychological Testing. AERA Publications.

Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. *Medical Education* 2005; 39: 276–283

Bergus GR, Kreiter CD. The reliability of summative judgements based on objective structured clinical examination cases distributed across the clinical year. *Medical Education* 2007: 41: 661–666

Borgersen NJ, Naur TMH, Sørensen SMD, Bjerrum F, Konge L, Subhi Y, Thomsen A. Gathering Validity Evidence for Surgical Simulation: A Systematic Review. *Ann Surg.* 2018 Jun;267(6):1063-1068.

Dore KL, Kreuger S, Ladhani M, Rolfson D, Kurtz D, Kulasegaram K, Cullimore AJ, Norman GR, Eva KW, Bates S,Reiter HI. The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. *Acad Med.* 2010 Oct;85 (10 Suppl):S60-S63.

Downing, SM. Reliability: on the reproducibility of assessment data. *Medical Education*. 2004; 38: 1006–1012

Downing, Steven M. & Yudkowsky R. (Eds.) (2009) Assessment in Health Professions Education. New York & London: Routledge.

Evgeniou E, Peter L, Tsironi M, Iyer S. Assessment methods in surgical training in the United Kingdom. *J Educ Eval Health Prof.* 2013; 10: 2.

Gallagher AG, O'Sullivan GC, Neary PC, Carroll SM, Leonard G, Bunting BP, Traynor O. An objective evaluation of a multi-component, competitive, selection process for admitting surgeons into higher surgical training in a national setting. *World J Surg.* 2014 Feb;38(2):296-304.

Ghaderi I, Manji F, Park YS, Juul D, Ott M, Harris I, Farrell TM. Technical skills assessment toolbox: a review using the unitary framework of validity. *Ann Surg*. 2015 Feb;261(2):251-262.

Kane M, Case SM. The reliability and validity of weighted composite scores. *Appl Meas Educ*. 2004;17(3):221-240.

Kane M (2006). Validation. In RL Brennan (ED.), Educational Measurement (4th Ed., p.17-64). New York: American Council on Education and Greenwood.

Kreiter CD, Gordon JA, Elliott S and Callaway M. (2004). Recommendations for assigning weights to derive an overall course grade. *Teaching and Learning in Medicine*. 16(2), 133-138.

Kurtz S, Silverman J, Benson J, et al. Marrying content and process in clinical method teaching: enhancing the Calgary-Cambridge guides. *Acad Med* 2003;78: 802-809.

Louridas M, Szasz P, de Montbrun S, Harris KA, Grantcharov TP. International assessment practices along the continuum of surgical training. *Am J Surg.* 2016 Aug;212(2):354-360.

Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997 Feb; 84(2):273-278.

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

Schaverien MV, 2016. Selection for Surgical Training: An Evidence-Based Review. *J Surg Educ.* 2016 Jul-Aug;73(4):721-729.

Sonnadara RR, Mui C, McQueen S, Mironova P, Nousiainen M, Safir O, Kraemer W, Ferguson P, Alman B, Reznick R. Reflections on competency-based education and training for surgical residents. *J Surg Educ.* 2014 Jan-Feb;71(1):151-158.

Sugden C, Aggarwal R. Assessment and Feedback in the Skills Laboratory and Operating Room. *Surg Clin N Am.* 2010; 90: 519–533.

Yudkowsky R, Park YS, Janet Riddle J, Palladino C, Bordage G. Clinically Discriminating Checklists Versus Thoroughness Checklists: Improving the Validity of Performance Test Scores. *Acad Med*. 2014; 89:1057–1062.

VITA

NAME: Dara Ann O'Keeffe

EDUCATION: MB BCh BAO, University College Dublin, 1998
BMedSci, University College Dublin, 1998
MHPE, University of Illinois, Chicago

ACADEMIC EXPERIENCE: Simulation Lead in Postgraduate Surgical Education, Royal College of Surgeons in Ireland, 2015- date

Assistant Director for Simulation-Based Learning,
STRATUS Centre for Medical Simulation, Brigham and Women's Hospital, Boston, MA. 2013 – 2015

Instructor, Harvard Medical School, 2014 – 2015

Special Lecturer in Surgical Education, 2006 - 2013
National Surgical Training Centre, Royal College of Surgeons in Ireland.

CLINICAL EXPERIENCE: Surgical residency and research, Republic of Ireland 1998 – 2005

PROFESSIONAL MEMBERSHIP: Member of the Royal College of Surgeons, 2001
Surgical Education Research Fellowship, Association for Surgical Education, 2010

PUBLICATIONS:

O'Keeffe DA, Nugent E, Neylon K, Conroy RM, Neary P, Doherty E. Use of a novel measure of non-technical skills in surgical trainees: Is there an association with technical skills performance? *J Surg Ed*, 2019 Mar - Apr;76(2):519-528.

Duignan S, Ryan A, O'Keeffe D, Kenny D, McMahon CJ. Prospective Analysis of Decision Making during Joint Cardiology Cardiothoracic Conference in Treatment of 107 Consecutive Children with Congenital Heart Disease. *Pediatric Cardiology*. 2018 Oct;39(7):1330-1338.

Harrington CM, Kavanagh DO, Quinlan JF, Ryan D, Dicker P, O'Keeffe DA, Traynor O, Tierney S. Development and Evaluation of a Trauma Decision-making Simulator in Oculus Virtual Reality. *Am J Surg*. 2018 Jan;215(1):42-47.

Davis WA, Jones S, Crowell-Kuhnberg AM, O'Keeffe D, Boyle KM, Klainer SB, Smink DS, Yule S. Operative team communication during simulated emergencies: Too busy to respond? *Surgery*. 2017 May;161(5):1348-1356.

O'Keeffe DA; Bradley D; Evans LA; Bustamante ND; Timmel M; Akineni R; Mulloy DF; Goralnick E; Pozner CN. Ebola Emergency Preparedness: Simulation Training for Frontline Healthcare Professionals. *MedEdPORTAL*. 2016; 12: 10433.

De Blacam C, O'Keeffe DA, Nugent E, Doherty E, Traynor O. Are Residents Accurate in their Assessments of their own Surgical Skills? *Am J Surg*. 2012 Nov; 204(5): 724-31.

Boyle E, O'Keeffe DA, Naughton PA, Hill AD, McDonnell CO, Moneley D. The importance of expert feedback during endovascular simulator training. *J Vasc Surg*. 2011 Jul;54(1):240-248.

Doherty E, O'Keeffe DA, Traynor O. Developing a human factors and patient safety programme at the Royal College of Surgeons in Ireland. *The Surgeon*. 2011 Vol 9, Suppl 1: S38-S39.

Boyle E, Kennedy AM, Doherty E, O'Keeffe DA, Traynor O. Coping with Stress in Surgery: The difficulty of measuring non-technical skills. *Ir J Med Sci*. 2011 Mar;180(1):215-20.

O'Keeffe DA, Hill AD, Sheahan K, Ryan F, Barton D, Fitzgerald R, McDermott E, O'Higgins N. RET Proto-oncogene analysis in Medullary Carcinoma of the Thyroid Gland. *Ir J Med Sci*. 1998 Oct-Dec;167(4):226-30.