Development of a novel behaviourally anchored instrument for the assessment of surgical trainees

Tzong-Yang Pan [ID],* Frank Piscioneri* and Cathy Owen†

*Canberra Hospital, Australian National University, Canberra, Australia and
†Medical School, Australian National University, Canberra, Australia

Correspondence

Dr Tzong-Yang Pan, Building 4, The Canberra Hospital, Hospital Rd, Garran, ACT, 2605, Australia.
Email: tzong.yang.pan@gmail.com

T.-Y. Pan MChD, BSc; **F. Piscioneri** FRACS, MBBS; **C. Owen** MD, FRANZCP.

Abstract

Background: The Royal Australasian College of Surgeons (RACS) created its competency framework in 2003 which initially consisted of nine competencies each regarded as equally important for a practising surgeon. The JDocs Framework is aligned to these competencies and provides guidance for junior doctors working towards the Surgical Education and Training program.

Methods: A novel assessment instrument was designed around the JDocs framework using 48 behaviourally anchored questions. The study was completed in 2020 across five public hospitals in the ACT and NSW. Participants were invited to complete the self-assessment form online.

Results: Thirty-six of 59 (61%) trainees participated in the study, with 67 of 68 (98.5%) supervisors having completed the assessment form. Trainee self-rating scores were lower than that of supervisor ratings across all competencies except communication. The self-rating scores were negatively correlated with the seniority of a trainee's level in all nine competencies. The years of post-graduate experience was positively correlated with seven of the nine competencies. For gender and International Medical Graduate status, correlation was only identified for health advocacy and medical expertise. There was no correlation identified with a trainee's age.

Conclusion: This pilot study has provided an opportunity to explore a new assessment instrument for surgical trainees that is aligned to the RACS competency framework using behaviourally anchored questions. Looking ahead, a better understanding of this instrument will potentially be helpful in early identification of underperforming trainees in order to facilitate early intervention, or its use as a selection tool for formal training programs.

Introduction

The Royal Australasian College of Surgeons (RACS) is the leading advocate for surgical standards in Australia and New Zealand. Ten core competencies required by a surgeon have been described by the college since 2003, encompassing the domains of Medical Expertise, Judgement – Clinical Decision-Making, Technical Expertise, Professionalism and Ethics, Health Advocacy, Communication, Collaboration and Teamwork, Management and Leadership, and Scholarship and Teaching. Each competency is deemed equally important and are assessed throughout surgical education and training by supervisors and examination boards. The JDocs Framework is aligned to these competencies and provides guidance

for junior doctors working towards the Surgical Education and Training program. Trainees who are awarded Fellowship of the Royal Australasian College of Surgeons (FRACS) are recognized as competent in all of these 10 domains and considered to be qualified to practice independently as a surgeon.[1]

Workplace based assessment is increasingly being utilized to assess surgical trainees regularly through their training. These assessments not only provide an opportunity for feedback and reflection, but can also identify struggling trainees so support can be provided early. These assessments can be extended beyond supervisors and include various colleagues or even patients in the form of a 360° assessment, which has been well validated in many specialities of medicine and surgery.[2–4] Assessment instruments

can be further improved by using behaviourally anchored rating scales as they lead to more consistent ratings due to their more clinically relevant descriptors, rather than descriptors of values and principles.[5,6]

The increasing acceptance of workplace-based assessments has naturally led to the emergence of new assessment instruments. However, many of these instruments are focused on a specific skill or domain. For example, there are workplace-based assessments which focus on 'communication in the operating theatre' or 'professional behaviour'. There remains a need for a comprehensive assessment instrument that encompasses all the skills required of a surgeon and is aligned to the RACS framework. The aim of this study was to develop a novel instrument for assessment of technical and non-technical skills for surgical trainees that is based off the nine RACS competencies utilizing the JDocs framework. The additional 10th RACS competency on Cultural competence and cultural safety, was implemented after this study was already undertaken so was not included in the design of the assessment tool (Data S1).

This pilot study had three primary questions:
(1) Does this novel instrument have adequate internal reliability to be used in more comprehensive assessments?
(2) Do supervisors have a tendency to provide ratings that are higher, or lower, than trainee self-ratings using this instrument and method of delivery?
(3) What variables of the trainee can be used to estimate their self-ratings for each competency?

Methods

A novel instrument was designed based off the RACS nine competencies utilizing the JDocs framework. Components of the instrument were reviewed by three independent senior surgical and medical educators and refined to produce a 48-item assessment tool covering all nine competencies.

Invitation to participate in the study was sent in April 2019 to all 59 trainees working in the ACT and South-East NSW health network, including prevocational surgical registrars, SET trainees and fellows. Background demographic data was collected which included age, gender, post-graduate year, level of training, and whether they were an International Medical Graduate (IMG). Participants then rated themselves in all 48 items on a Likert scale. Participants were asked to nominate two surgical supervisors to assess them using an analogous instrument modified for supervisors. All collected data was de-identified and the instrument was accessible for 12 weeks on SurveyMonkey for participants to complete.

Statistical analysis of the data was completed using SPSS 25. The reliability of the data for both trainees and supervisors were calculated using the Cronbach's Alpha test. The student's t-test (independent two-sample) was used to determine if there was a significant difference between the trainees' self-rating, to that of the averaged supervisor ratings. A multiple regression model using age, gender, post-graduate year, IMG status, and level of training as the variables was performed with backwards elimination, and pairwise comparisons made to identify the degree and direction of influence each variable contributed to trainee self-ratings.

Table 1 Demographic of trainees

Age	25–29	11 (32.4%)
	30–34	12 (35.3%)
	35+	11 (32.4%)
Gender	Female	12 (35.3%)
	Male	22 (64.7%)
PGY	3–4	11 (32.4%)
	5–6	10 (29.4%)
	7+	13 (38.2%)
Level of training	Prevocational	22 (64.7%)
	SET	9 (26.5%)
	Fellow	3 (8.8%)
IMG	No	29 (85.3%)
	Yes	5 (14.7%)

Results

During the survey period, 36 of 59 (61%) trainees completed the assessment. Two responses were grossly incomplete and were excluded from further analysis. From 25 nominated supervisors, 67 of 68 (98.5%) assessments were fully completed, 1 was partially complete. Five of these supervisors had completed the Foundation Skills for Surgical Educators course and all had completed the Operating With Respect course. The demographic data of the trainees is presented in Table 1. The Cronbach's Alpha reliability for each competency is presented in Table 2 and shows good consistency across all nine competency ratings for both trainee and supervisor responses with the exception of health advocacy for trainees. The student's t-test presented in Table 3 shows that for all competencies except communication, there is a statistically significant difference in self-rating scores compared with those of supervisors.

The multiple regression analysis undertaken looking at trainee self-rating scores is shown in Table 4. Trainees with more years of post-graduate experience were estimated to provide self-ratings that were higher across seven of the nine competencies compared to those with less years of experience. The post-graduate year of the trainee was identified as statistically significant in its correlation in the domains of communication, management and leadership, professionalism and ethics, scholarship and teaching, medical

Table 2 Internal consistency

Competency	Number	Trainees (α)	Number	Supervisors (α)
Communication	34	0.918	68	0.827
Collaboration and teamwork	34	0.806	68	0.702
Management and leadership	34	0.889	68	0.755
Professionalism and ethics	34	0.856	66	0.789
Health advocacy	34	0.447	68	0.713
Scholarship and teaching	34	0.904	67	0.819
Medical expertise	34	0.881	67	0.867
Judgement – clinical decision making	34	0.784	67	0.823
Technical expertise	34	0.832	67	0.845

Table 3 Student's *t*-test between trainee and supervisor ratings

Competency	Trainee self-rating	Averaged supervisor rating	P-value
Communication	4.218	4.387	0.097
Collaboration and teamwork	4.162	4.346	0.033
Management and leadership	3.891	4.239	0.000
Professionalism and ethics	3.861	4.308	0.000
Health advocacy	3.882	4.346	0.000
Scholarship and teaching	3.812	4.224	0.000
Medical expertise	3.960	4.334	0.000
Judgement – clinical decision making	4.140	4.340	0.029
Technical expertise	3.926	4.351	0.000

expertise, judgement and technical expertise. The level of training level was identified as statistically significant in its correlation with the competencies of communication, collaboration and teamwork, management and leadership, professionalism and ethics, health advocacy, scholarship and teaching, medical expertise, judgement and technical expertise. While for both gender and IMG status, correlation was only identified for health advocacy and medical expertise. There was no correlation identified for age.

The multiple regression model utilizes backward elimination. The baseline values for the estimate of parameter of each competency are presented with the estimated change in value for each subgroup.

Table 4 Multiple regression model for each of the nine RACS competencies

Competency	Terms		Estimate of parameter (se)		χ^2-statistics	df	P-value
Communication	Baseline value		4.039 (0.1066)				
	Accepted	Postgraduate year	PGY 3–4	0 (baseline)	17.55	2	<0.001
			PGY 5–6	+0.530 (0.1177)			
			PGY 7+	+0.779 (0.1705)			
		Training Level	Prevocational	0 (baseline)	29.39	2	<0.001
			SET	−0.556 (0.1839)			
			Fellow	−1.437 (0.2660)			
	Rejected	Gender			0.054	1	0.816
		IMG			0.067	1	0.796
		Age			0.755	2	0.755
Collaboration and teamwork	Baseline value		4.216 (0.0887)				
	Accepted	Training level	Prevocational	0 (baseline)	3409.1	3	<0.001
			SET	+0.006 (0.1387)			
			Fellow	−0.633 (0.2402)			
	Rejected	Postgraduate year			3.040	2	0.219
		Gender			0.007	1	0.933
		IMG			1.166	1	0.280
		Age			0.261	2	0.878
Management and leadership	Baseline value		3.740 (0.1423)				
	Accepted	Postgraduate year	PGY 3–4	0 (baseline)	7.873	2	0.020
			PGY 5–6	+0.424 (0.1571)			
			PGY 7+	+0.727 (0.2276)			
		Training Level	Prevocational	0 (baseline)	10.010	2	0.007
			SET	−0.606 (0.2454)			
			Fellow	−1.039 (0.3551)			
	Rejected	Gender			2.554	1	0.110
		IMG			2.638	1	0.104
		Age			1.885	2	0.390
Professionalism and ethics	Baseline value		3.766 (0.1329)				
	Accepted	Postgraduate year	PGY 3–4	0 (baseline)	6.747	2	0.034
			PGY 5–6	+0.188 (0.1467)			
			PGY 7+	+0.643 (0.2124)			
		Training Level	Prevocational	0 (baseline)	8.003	2	0.018
			SET	−0.483 (0.2291)			
			Fellow	−0.885 (0.3314)			
	Rejected	Gender			3.585	1	0.058
		IMG			3.491	1	0.062
		Age			0.278	2	0.870
Health advocacy	Baseline value		3.814 (0.1042)				
	Accepted	Training level	Prevocational	0 (baseline)	12.178	2	0.002
			SET	−0.122 (0.1500)			
			Fellow	−1.272 (0.3932)			
		Gender	Male	0 (baseline)	4.575	1	0.032
			Female	+0.317 (0.1482)			
		IMG	No	0 (baseline)	5.083	1	0.024
			Yes	+0.686 (0.3043)			
	Rejected	Age			0.075	2	0.963
		Postgraduate year			2.329	2	0.374
Scholarship and teaching	Baseline value		3.691 (0.1606)				
	Accepted	Postgraduate year	PGY 3–4	0 (baseline)	6.600	2	0.037
			PGY 5–6	+0.195 (0.1773)			

Table 4 Continued

Competency		Terms		Estimate of parameter (se)	χ^2-statistics	df	P-value
		Training level	PGY 7+	+0.753 (0.2568)	6.419	2	0.040
			Prevocational	0 (baseline)			
			SET	−0.492 (0.2769)			
			Fellow	−0.978 (0.4006)			
	Rejected	Gender			2.179	1	0.140
		IMG			1.649	1	0.199
		Age			0.303	2	0.860
Medical expertise	Baseline value			3.588 (0.1033)			
	Accepted	Postgraduate year	PGY 3–4	0 (baseline)	25.559	2	0.000
			PGY 5–6	+0.551 (0.1085)			
			PGY 7+	+0.814 (0.1567)			
		Training level	Prevocational	0 (baseline)	29.594	2	0.000
			SET	−0.458 (0.1657)			
			Fellow	−1.524 (0.2942)			
		Gender	Male	0 (baseline)	4.045	1	0.044
			Female	+0.226 (0.1123)			
		IMG	No	0 (baseline)	4.588	1	0.032
			Yes	+0.505 (0.2357)			
	Rejected	Age			0.599	2	0.741
Judgement – clinical decision making	Baseline value			3.955 (0.1099)			
	Accepted	Postgraduate year	PGY 3–4	0 (baseline)	15.920	2	0.000
			PGY 5–6	+0.378 (0.1214)			
			PGY 7+	+0.823 (0.1758)			
		Training Level	Prevocational	0 (baseline)	17.471	2	0.000
			SET	−0.541 (0.1895)			
			Fellow	−1.112 (0.2742)			
	Rejected	Gender			0.205	1	0.651
		IMG			0.197	1	0.657
		Age			3.742	2	0.154
Technical expertise	Baseline value			3.455 (0.1400)			
	Accepted	Postgraduate Year	PGY 3–4	0 (baseline)	23.296	2	0.000
			PGY 5–6	+0.768 (0.1545)			
			PGY 7+	+1.2 (0.2238)			
		Training level	Prevocational	0 (baseline)	14.471	2	0.001
			SET	−0.365 (0.2413)			
			Fellow	−1.322 (0.3492)			
	Rejected	Gender			2.787	1	0.095
		IMG			2.344	1	0.126
		Age			1.491	2	0.474

Discussion

As boards of surgical training programs and surgical institutions continue to explore the use of work-based assessments, it is becoming more important to understand the relationships and correlations between the instrument and trainees. This pilot of our instrument to trainees and supervisors was simple in its delivery. The response rates for both trainees (61%) and supervisors (98.5%) represent excellent participation in the survey which support the acceptability of the delivery method. The design we utilized can be easily expanded to include more supervisors, or ratings from colleagues in other work relationships such as nurses. The flexibility of this instrument means that it can be adapted to be used as a comprehensive 360° assessment, or a simple 1-on-1 supervisor feedback template.

In this pilot study, we have demonstrated excellent internal consistency of each competency in the instrument for both trainee and supervisor ratings, the exception being poor reliability of trainee self-ratings in health advocacy. Interestingly, supervisor ratings for this competency demonstrates good reliability. Whether this finding suggests that supervisors are truly able to provide more consistent evaluations of this competency than trainees, or are reliably providing the same rating to both questions as they have no idea how to answer, is a point of interest. Future iterations of this instrument can clarify this and improve the consistency in this competency by expanding on the current two questions to at least four–six questions to make it similar to the other competencies.

In the comparison of trainee self-ratings with averaged supervisor ratings, supervisors gave higher ratings across eight of the nine competencies. This finding is consistent with other instruments in the literature demonstrating that supervisor ratings are higher than self-ratings.[7] Trainees are thought to be more modest in self-ratings, while supervisors may feel that providing low scores may result in a plethora of further inquiries about the trainee's performance and the additional work to be a deterrent. This tendency to give higher ratings is a problem as it may not accurately identify trainees who may be 'slightly underperforming'. However, trainees who are 'severely underperforming' may still be given ratings far enough below the average to be identified as a struggling trainee. If a greater number of supervisors are used in this instrument, then it will improve the likelihood of identifying struggling trainees and provide the opportunity for early academic and personal support.

The data shows that trainees of a higher post-graduate year experience are estimated to rate themselves higher across seven of the nine competencies. The exceptions to this being 'collaboration and teamwork' and 'health advocacy' (Table 4). This implies that simply having more years of clinical experience does not directly result in trainees feeling more competent in these two areas. Targeted training in these two areas may be most valuable for trainees and surgical training programs can consider placing more time and emphasis on them in the future.

The model reveals an interesting finding about the level of training of trainees with their self-ratings in all nine of the competencies. Prevocational trainees are expected to rate themselves the highest, but as they progressed along their career into a SET or fellow, the self-ratings declined. It is important to emphasize that this model is a multiple regression model, so the post-graduate year has already been factored in. This finding suggests that those in SET and fellows have better insight and are more aware of their own weaknesses in all nine competencies. Trainees early in their career are unconscious of their incompetence and rate themselves higher than they feel they are, that is to say that they 'don't know what they don't know'. As a trainee then progresses in their training, they then reach a level of conscious incompetence. They start to understand the limits of their knowledge and skills, they 'know what they don't know'. The last two stages of this model are conscious competence and unconscious competence and are probably most likely to be seen at the consultant and senior consultant level. The instrument used in this study prompts trainees to provide ratings that they feel are relative to their level of experience, this is subjective in nature. However, it does suggest that there is an underlying element of increasing standards and expectations of oneself at higher levels of training. This progression can also be interpreted as improving insight and is likely a result of the actual training program and increasing responsibility in their roles, rather than a simply having more years of experience.

For two of the nine competencies 'health advocacy' and 'medical expertise' the model shows that females and IMGs rate themselves higher than their counterparts. However, the reliability of trainees self-rating themselves on 'health advocacy' is poor so it is difficult to interpret this further and requires further research to clarify this specifically. The finding that females are expected to rate themselves higher than males in 'medical expertise' is interesting, as it is not reflected in any of the other competencies. The same finding is also seen for IMGs rating themselves higher than non-IMGs. There is no literature to date that explains this and further research is required to clarify these findings and investigate the underlying reasons.

The pilot of this instrument has its limitations. Firstly, each trainee was only required to nominate two supervisors for rating and feedback. We encourage that further use of this instrument should utilize as many supervisors as possible in as it would help not only provide more volume and variation in feedback, but also identify any underperforming trainees. For an instrument to be reliably used as a 360° assessment, then more colleagues should be sought for assessment as literature suggests as many as 5–10 are required for reproducible results.[8] The

competency of 'health advocacy' did not have many sub-questions and can be further expanded to improve reliability. The sample size of Fellows was small and we did not stratify IMG by the country of their original training, nor whether the training was undertaken in an English-speaking country or not. Further research with more Fellows involved and data around IMGs to find if there are any correlations and allow us to draw stronger conclusions. There would also be value in comparing this new assessment tool against other validated assessment tools in the medical field.

This pilot has provided an opportunity to explore the use of a new assessment and feedback instrument for surgical trainees that is aligned to the RACS competency framework using behaviourally anchored questions and rating scales to maximize relevance and trainee reflection upon their skills. Looking ahead, a better understanding of this instrument will potentially be helpful in early identification of underperforming trainees in order to facilitate early intervention, or even its use as a selection tool in formal training programs.

Acknowledgement

Conflict of interest

None declared.

Author contributions

Tzong-Yang Pan: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; validation; visualization; writing – original draft; writing – review and editing. **Frank Piscioneri:** Conceptualization; data curation; formal analysis; investigation; methodology; project administration; supervision; validation; writing – review and editing. **Cathy Owen:** Conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; supervision; validation; visualization; writing – review and editing.

References

1. RACS. Training Standards Framework 2012.
2. Al Khalifa K, Al Ansari A, Violato C, Donnon T. Multisource feedback to assess surgical practice: a systematic review. *J. Surg. Educ.* 2013; **70**: 475–86.
3. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad. Med.* 2014; **89**: 511–6.
4. Stevens S, Read J, Baines R, Chatterjee A, Archer J. Validation of multisource feedback in assessing medical performance: a systematic review. *J. Contin. Educ. Heal. Prof.* 2018; **38**: 262–8.

5. Smith PC, Kendall LM. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *J. Appl. Psychol.* 1963; **47**: 149–55.
6. Bernardin HJ, Smith PC. A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). *J. Appl. Psychol.* 1981; **66**: 458–63.
7. Arora S, Miskovic D, Hull L *et al*. Self vs expert assessment of technical and non-technical skills in high fidelity simulation. *Am. J. Surg.* 2011; **202**: 500–6.
8. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. Resident-patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. *Acad. Med.* 1994; **69**: 216–24.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1 Supporting information