



2025

To Fall at the Final Hurdle? Unpacking Experiences of Ophthalmology Trainees Undertaking Final Summative Examinations in Australia and New Zealand

Belinda Jessup

University of Tasmania

Penny Allen

University of Tasmania

Melissa Kirschbaum

University of Tasmania

Santosh Khanal

Royal Australian and New Zealand College of Ophthalmologists

Victoria Baker-Smith

Royal Australian and New Zealand College of Ophthalmologists

Barnabas Graham

Royal Australian and New Zealand College of Ophthalmologists

Tony Barnett

University of Tasmania

Follow this and additional works at: <https://ajce.scholasticahq.com/>



© Copyright the authors.

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivative 4.0 International Licence](https://creativecommons.org/licenses/by-nc-nd/4.0/).

To Fall at the Final Hurdle? Unpacking Experiences of Ophthalmology Trainees Undertaking Final Summative Examinations in Australia and New Zealand

**Belinda Jessup,^{*} Penny Allen,[~] Melissa Kirschbaum,^{*} Santosh Khanal,[^]
Victoria Baker-Smith,[^] Barnabas Graham,[^] Tony Barnett^{*}**

^{*} Centre for Rural Health, University of Tasmania
[~] Rural Clinical School, University of Tasmania
[^] Royal Australian and New Zealand College of Ophthalmologists

Abstract

Background: Trainees of specialist medical colleges commonly experience sub-optimal achievement on final summative examinations.

Objective: To explore experiences of trainees sitting the Royal Australian and New Zealand College of Ophthalmologists Advanced Clinical Examination (RACE), and their supervising Fellows, to identify reasons for failure.

Methods: Mixed methods study involving trainee post-RACE feedback surveys 2013-2021 (written n=317, clinical n=418), and semi-structured interviews with trainees (n=19) who sat RACE 2017-2021, and supervising ophthalmologists (n=10).

Results: Four meta-themes were identified: 1) Examination difficulty: *"Oh, I read that paper, it's ... so hard"*, related to the perception that some examinations were disproportionately challenging; 2) Ambiguous questions: *"I don't know what you guys are asking me"*, encompassed concerns about understanding the content and volume of information required; 3) Not fit for purpose: *"The majority of competent general ophthalmologists, they would not pass"*, captured perceptions that some examinations reflected subspecialty rather than generalist knowledge; and 4) Time pressure: *"It is so time pressured"*.

Conclusion: Enhanced mechanisms are required to improve examination reliability, so as not to unfairly compromise the achievement of specific cohorts of trainees. Valid and defensible score adjustment processes are needed to account for variability in examination design and difficulty across semesters.

I INTRODUCTION

In Australia and New Zealand, training to become an ophthalmologist takes 10 to 20 years, commencing with undergraduate medical school and progressing through internship, residency, two years of pre-Vocational training, and five years of vocational training with the Royal Australian and New Zealand College of Ophthalmologists (RANZCO). The RANZCO Advanced Clinical Examination (RACE) is the final summative barrier examination of the vocational training program, which must be successfully completed before progression to the final year of training (The Royal Australian and New Zealand College of Ophthalmologists, 2018).

Final summative medical specialty examinations, such as the RACE, are notoriously difficult and are known to cause significant stress for trainees (Hoffman et al., 2022; Abhary et al., 2021; Royal Australian and New Zealand College of Ophthalmologists, 2022a). This is despite trainees often having committed years to preparation and having passed numerous formative and summative examinations in earlier years of training. As well as negatively affecting trainee wellbeing (Winter et al., 2021), examination failure causes delays to trainee career progression. This subsequently impacts the specialist healthcare workforce, which is currently in, or is projected to be in, undersupply for many medical specialties, including ophthalmology (Department of Health, 2018, and 2021).

High failure rates are common within Australian specialist medical colleges (Australian Medical Council, 2021) and in other developed countries (Henry, 2021; Willett et al. 2016). According to Department of Health Medical and Education Training (MET) data, 47% of RANZCO advanced trainees failed the written component of the RACE in 2019 and 41% failed the clinical component (Department of Health and Aged Care, 2020). MET data for the same year shows similar failure rates for the Australasian College of Emergency Medicine (42% failed the written examination and 41% failed the clinical examination) and the written component of the Royal Australian and New Zealand College of Obstetricians and Gynaecologists examination (40% failure rate).

In Australia, the Australian Medical Council (AMC) is responsible for the accreditation of specialist medical colleges and for setting standards to ensure that assessments are fair and evidence-based. Concerns about low examination pass rates have prompted the AMC to review college assessment procedures and to utilise academic research, trainee feedback and external consultation to establish and ensure the implementation of best-practice policies, the process of which is ongoing (Australian Medical Council, 2021).

There is relatively little published literature around reasons for postgraduate medical examination failure. Trainee-related factors, such as lack of preparation and difficulty balancing study with clinical work, have been reported (Ashokka et al., 2021; Henry, 2021). The validity, fairness, and transparency of the examination process have been questioned (Abhary et al., 2021; Alexander, 2015; Khanna et al., 2018; Sami, 2017), with suggestions that examiner bias, language barriers (Henry, 2021), as well as ambiguity and the subjectivity of questions (Alexander, 2015) contribute to poor examination outcomes. A study exploring the views of specialist trainees in Australia (Australian Medical Association Council of Doctors in Training, 2021) reported that trainees perceive success to be dependent on examination technique rather than knowledge, that content does not reflect clinical practice or the curriculum, and that marks may be influenced by examiner perceptions of trainees. Insufficient time to complete assessment was also thought to contribute to low pass rates.

We need to ensure that competent ophthalmology trainees progress into the specialist workforce in a timely manner, yet almost half of Australian and New Zealand ophthalmology trainees are failing the final speciality examination. The problem of high failure rates may rest with the examination itself, or it may lie elsewhere (such as trainee burnout, inequitable access to experienced supervisors and examination preparation activities, difficulties joining a study group, etc.) We wanted to know how trainees and supervisors (Fellows) viewed this issue. Our research questions were: 1) How do trainees and supervisors view the examination construct, design, and delivery? 2) What factors do trainees and supervisors perceive to inform low RACE pass rates? 3) What do trainees and supervisors see as potential solutions? The findings may assist specialist

medical colleges more broadly to improve the design and implementation of final summative examinations, thereby reducing the personal, social, and workforce consequences associated with specialist trainee failure.

II METHODS

An exploratory sequential mixed-methods design was used involving an initial survey component, followed by semi-structured interviews to further explain the survey findings (Cresswell & Cresswell, 2018). Ethics approval to undertake the study was provided by the Tasmania Health and Medical Human Research Ethics Committee (Project ID: 25018) and the RANZCO Human Research Ethics Committee (Project ID: 129.21).

A *Setting*

RACE is the final summative examination of the Royal Australian and New Zealand College of Ophthalmologists. Trainees must have completed three years of the vocational training program and have passed previous summative and formative assessments to be eligible to sit RACE (RANZCO, 2018; Jessup et al., 2022). The examination comprises a written (online) component and a clinical examination. The written component consists of two parts: Part A, nine Short Essay Questions (SEQs); and Part B, 30 Very Short Answer Questions (VSAQs). The questions are weighted for equal representation from curriculum performance standards for nine practice areas assessed by RACE (Box 1). The clinical component consists of an Objective Structured Clinical Examination (OSCE) with two stations assessing each of the nine clinical curriculum performance standards. The written and clinical components must be successfully passed before progression to the fifth and final year of training. At the time of the research, trainees were permitted to attempt the examination up to three times. A new policy will become effective in mid-2024, allowing unlimited attempts at RACE within the maximum 12-year training period.

Box 1.

Clinical Curriculum Performance Standards examined by RACE

- Cataract and Clinical Refraction
- Cornea and External Eye Disease
- Glaucoma
- Neuro-ophthalmology
- Ocular Inflammation
- Ocular Motility
- Oculoplastic and Orbit
- Paediatric Ophthalmology
- Vitreoretinal

B *Post-RACE feedback survey*

During the initial quantitative phase, RANZCO provided historical RACE feedback survey data for the period semester 1, 2013 to semester 2, 2021. This survey is administered by RANZCO and is sent to all RACE candidates following each written and clinical examination for the purposes of allowing RANZCO to gather feedback on aspects of trainee examination experience. Consent to participate in the survey is implied by completion. A total of 16 survey questions using four-point Likert scales (strongly agree, agree, disagree, strongly disagree) sought feedback on examination design, delivery and experiences. Free text boxes were also included for each question to capture qualitative responses. To protect the confidentiality of survey respondents, demographic data (e.g., age, gender, training network) is not collected in the survey.

Quantitative data were analysed using descriptive analyses. The percentage of respondents who agreed or strongly agreed with each survey question was calculated. Free text survey responses were imported into NVivo (QSR, 2022) qualitative data analysis software, with thematic analysis undertaken (BJ) (Bruan and Clarke, 2006) to identify codes directly related to difficulties with the examination construct, design, delivery, or experience. A second author (PA) independently read the free text responses and verified the accuracy and completeness of the codes. The survey data was then amalgamated with the interview data as described below.

C Semi-structured interviews

The qualitative component comprised of semi-structured in-depth interviews. Ophthalmology trainees or Fellows who had sat RACE within the past five years (2017-2021) (n=166); and RANZCO Fellows providing supervision to trainees who were soon to sit, or had recently sat RACE (n=127), were invited to participate. Participants were recruited by email invitation which included an attached Study Information Sheet. Prior to interview, all participants were asked if they had read the Study Information Sheet and whether they had any questions, before being asked to provide verbal consent to participate. Interviews were conducted via Zoom or telephone at a time convenient (average length 42 minutes, range 27 to 75 minutes). Guiding questions were used to elicit key information during the interview and varied depending on whether the participant had recently sat RACE or were in a position where they provided supervision to trainees (Appendix A). Different questions were asked of trainees and supervisors based on their roles within the training and examination program. Information was sought relevant to all the interview guide questions, although the order in which the information was elicited varied according to the interviewee's responses. All interviews were audio recorded and transcribed verbatim, with participants given the opportunity to check their transcript for accuracy and completeness and to make any changes where required, prior to data analysis. Once confirmed, transcripts were deidentified by allocating a numerical code and redacting all identifiers.

Interview data were thematically analysed (Braun & Clarke, 2006). Firstly, transcripts were reviewed multiple times by two members of the research team (BJ, PA). Using an inductive approach, both authors then independently coded transcripts using Nvivo (QSR, 2022). Codes were identified as they emerged, by selecting relevant text and creating a named code with a description. Codes were then hierarchically organised into broader sub-themes and themes as the coding progressed using the software. Iterative discussions between the research team helped to refine and confirm identified themes.

D Data integration

Convergent data analysis was undertaken, integrating the coded text from the survey to the interview data analysis. This process was used to identify areas of agreement, partial agreement, and dissonance between the survey and interview data (O'Cathain, Murphy & Nicholl, 2010). This process was achieved through consensus discussion with members of the research team (BJ, PA, TB) and saw the emergence of meta-themes, which cut across both quantitative and qualitative data sets and underpinned the overall findings of the study. Verbatim quotations were used from either the survey or interview data to exemplify findings given the agreement found.

III RESULTS

Between semester 2, 2013 and semester 2, 2021, a total of 317 survey responses were received from trainees following the clinical examination, and 418 responses were received following the written examination. Response rates for the written examination feedback survey ranged from 66% to 96%, depending on the semester, while response rates for the clinical examination feedback survey ranged from 39% to 94%. Survey responses were received from trainees from all training networks across Australia and New Zealand.

A total of 29 interviews were conducted (12 females and 17 males), including those who had sat RACE recently (n=19), together with those supervising RACE candidates (n=10). Interviewees were drawn from all seven RANZCO training networks across Australia and New Zealand.

Following integration of survey and interview data, four meta themes were identified related to difficulties with the examination construct, design, delivery or experience: “Oh, I read that paper, it’s ... so hard”; “I don’t know what you guys are asking me”; “The majority of competent general ophthalmologists, they would not pass”; and “It is so time pressured”.

A Examination difficulty: “Oh, I read that paper, it’s ... so hard”

Interviewees shared their beliefs that the RACE should set a high passing standard to ensure the safety, competency, and integrity of the ophthalmologists exiting the training program. However, interviewees also shared their concern that the examination was sometimes overly challenging, creating a barrier for trainees who were recognised as competent and safe in the clinical setting to achieve pass standards. Interviewees expressed concern about the written examination and its variable level of difficulty across semesters.

The written was a lot more challenging than expected and compared to past papers. I had done all of the past papers online since I think 2012 and I did all of the questions... maybe I’m biased but I think it was the most challenging one that had been up to that point. Some of the other people had also fed back, like other registrars that had sat just before us, they’re like, ‘oh, I read that paper, it’s ... so hard’ (Trainee interview #14).

Interviewee comments were supported by the survey data (Table 1 and Table 2). Across all semesters, an average of 75% of survey respondents agreed or strongly agreed that the SEQs were pitched at the appropriate level of difficulty (Table 1). However, in semester 2, 2019 and semester 2, 2021 this was as low as 58%. Semester 1, 2020 had the highest level of agreement, with 96% of respondents agreeing or strongly agreeing that the SEQs were pitched at the appropriate level of difficulty. Similarly, there was wide variation in the percentage of survey respondents who agreed or strongly agreed that the VSAQs were set at the appropriate level of difficulty, from a highpoint of 92% in semester 2, 2017 to a low of 54% among respondents in semester 2, 2020.

Table 1.

Written examination feedback survey, percentage of respondents who agreed or strongly agreed with statements (n=418)

Semester	Written Examination Overall %				Very Short Answer Questions (VSAQs) %			Short Essay Questions (SEQs) %			
	Pass rate %	Examination process went smoothly	Examination related closely to Curriculum Standards	Related closely to Clinic / Theatre Work	Examination pitched at appropriate Level of Difficulty	Easy to complete tasks in Allocated Time	Clearly Phrased and Unambiguous	Related Closely to Clinic / Theatre Work	Examination pitched at appropriate Level of Difficulty	Easy to complete in Allocated Time	Clearly Phrased and Unambiguous
s2 2013	n/a	100.0	90.5	38.1	38.1	23.8	95.2	85.7	76.2	14.3	66.7
s1 2014	n/a	76.9	83.3	*	59.0	59.0	82.1	51.3	59.0	23.1	66.7
s2 2014	n/a	100.0	93.3	60.0	86.7	73.3	100.0	80.0	73.3	33.3	66.7
s1 2015	n/a	90.9	100.0	*	87.9	51.5	97.0	81.8	90.9	12.1	78.8
s2 2015	n/a	95.5	*	54.5	72.7	59.1	95.5	81.8	81.8	18.2	81.8
s1 2016	n/a	96.9	80.0	56.3	71.9	78.1	87.5	46.9	65.6	34.4	59.4
s2 2016	n/a	86.7	100.0	33.3	*	66.7	100.0	66.7	66.7	33.3	80.0
s1 2017	n/a	97.5	95.0	45.0	72.5	82.5	85.0	67.5	72.5	62.5	*
s2 2017	30%	100.0	100.0	61.5	92.3	100.0	84.6	69.2	84.6	23.1	61.5
s1 2018	81%	67.2	96.9	54.7	87.5	90.6	90.6	90.6	81.3	53.1	65.6
s2 2018	73%	87.5	93.8	50.0	62.5	50.0	75.0	68.8	93.8	43.8	87.5
s1 2019	46%	92.0	72.0	40.0	68.0	80.0	96.0	68.0	72.0	44.0	60.0
s2 2019	67%	78.9	84.2	57.9	89.5	63.2	68.4	73.7	57.9	31.6	73.7
s1 2020	63%	100.0	92.0	60.0	88.0	56.0	92.0	84.0	96.0	40.0	84.0
s2 2020	44%	75.0	83.3	37.5	54.2	33.3	66.7	75.0	75.0	25.0	79.2
s1 2021	n/a	94.3	80.0	*	*	85.7	97.1	54.3	*	34.3	*
s2 2021	n/a	91.7	83.3	*	58.3	66.7	83.3	41.7	58.3	25.0	75.0
Mean	-	90.1	89.2%	49.9	72.6	65.9	88.0	69.8	75.3	32.4	72.4
SD	-	10.0	8.5%	9.9	16.1	19.9	10.4	14.2	12.2	13.4	9.2

*Question not asked in survey that semester. SD: Standard deviation. n/a denotes examination results not available for semester.

Given the variability in perceived difficulty, some interviewees shared their belief that an element of luck contributed to their passing achievement, since their examination was felt to be easy relative to others. Interviewees believed that challenging examinations with a high failure rate had to reflect, in part, deficiencies in the rigour of the examination development process to ensure a consistent level of difficulty. This was noted as unfairly impacting candidates sitting during years when the examination was considered more difficult than usual.

So I think 80% of the exams are generally quite good. But every few years, you just get this one that's just, 'what on earth were they thinking?' That's the response. When I talked to my candidates the day after that exam, I was just shocked. We were all shocked. Even the people who had just sat six months previously were shocked. So that's clearly a failure of the system to have not picked that up at the time that the exam was being set (Supervisor interview #5).

B Ambiguous questions: “I don’t know what you guys are asking me”

The theme “I don’t know what you guys are asking me” related to the ambiguity of some examination questions and encompassed both challenges with understanding what and how much information to provide. Again, this issue varied in prevalence from semester to semester and was more evident in the written examination survey responses. For example, an average of 85% (range 67%-100%) of respondents across all years agreed or strongly agreed that the clinical examination tasks were clearly phrased and unambiguous (Table 2). A comparatively lower average of 72% (range 59%-88%) of candidates agreed or strongly agreed that the SEQs in the written examination were unambiguous, although the feedback on the VSAQs was similar to the clinical examination (average 88% range 67%-100%) (Table 1 and Table 2). Despite the survey findings, many interviewees described how poorly worded written examination questions left them unsure of what exactly was being asked of them. Consequently, interviewees shared experiences of failing questions on the written component that they could have answered correctly had they understood what was being asked. Interviewees described that understanding marking allocations for each question may also be helpful in clarifying how much information to provide in response to an ambiguous question.

Basically, there's a 'guess what I'm thinking' and 'I don't know what you're asking for', so you either get lucky or you get unlucky. And it shouldn't be that way. I guess the questions, you shouldn't be feeling 'I don't know what you guys are asking me' (Trainee interview #19).

There was so much ambiguity! Need to show perspective, but things could be interpreted in many ways which means we might fail even with ample knowledge (Survey respondent semester 1, 2021).

In contrast to the written exam, interviewees highlighted that an important advantage of the clinical examination was that it allowed for prompting and feedback to be provided concurrently by the examiners. This was seen as critical to success, as trainees were able to be redirected if they had misunderstood the question asked or had gone off on a tangent not desired by the examiner. On average across all years of the survey, 93% (range 77%-100%) of respondents agreed or strongly agreed that examiner communication during the clinical examination helped them demonstrate their knowledge (Table 2). However, interviewee experiences highlighted that individual examiners varied in their ability and willingness to ‘guide’ trainees throughout the scenario to keep them focused. Trainees were concerned that performance was more related to examiner ability than their own lack of knowledge.

The written has one key difference to the clinical and that's there's no one to right the ship if it's going awry. It's a true thing. Like, in the clinical ... if you knew what you're talking about but you were just heading off into the woods, they were very quick to redirect you, bring it back, keep you on point, keep you on topic and get you answering the questions appropriately. In the written, it's a free for all. If you go off into the woods, you are going to spend the entire 9 or 10 minutes writing that essay on the wrong thing and then they'll just look at it and say, 'wow, [they] didn't even answer the question. Fail' (Trainee interview #3).

There was a lot of variability with examiners – some guided really well, but I felt that others did not and in their stations, it was difficult to know what the aim of the station was (Survey respondent semester 1, 2020).

Table 2.
Clinical examination feedback survey, percentage of respondents who agreed or strongly agreed with statements (n=317)

Semester	Pass rate	Examination process ran smoothly	Images were good quality	Examiners polite and non-threatening	Examiner communication helped demonstrate knowledge	Examination related closely to Clinical Curriculum Standards	Examination pitched at appropriate level of difficulty	Tasks closely to clinic/ theatre experience	Tasks clearly phrased and unambiguous	Tasks effective at testing ability to evaluate cases and present assessment and management	Easy to complete tasks in time
s2 2013	n/a	100.0	72.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	72.7
s1 2014	n/a	96.6	72.4	96.6	96.6	96.6	89.7	69.0	72.4	93.1	37.9
s2 2014	n/a	100.0	36.4	90.9	100.0	90.9	100.0	90.9	81.8	90.9	54.5
s1 2015	n/a	100.0	97.0	100.0	87.9	97.0	97.0	90.9	90.9	87.9	33.3
s2 2015	n/a	100.0	100.0	100.0	100.0	100.0	91.7	100.0	100.0	100.0	50.0
s1 2016	n/a	100.0	76.9	96.2	96.2	100.0	92.3	80.8	80.8	88.5	53.8
s2 2016	n/a	100.0	88.9	100.0	88.9	100.0	94.4	88.9	83.8	100.0	88.9
s1 2017	n/a	96.9	84.4	96.9	93.8	96.9	96.9	90.6	93.8	96.9	87.5
s2 2017	71%	100	42.9	100.0	85.7	100.0	100.0	85.7	71.4	100.0	85.7
s1 2018	88%	100	74.1	96.3	92.6	100.0	100.0	85.2	92.6	96.3	74.1
s2 2018	56%	84.6	53.8	69.2	76.9	92.3	69.2	69.2	76.9	84.6	53.8
s1 2019	58%	96.8	71.0	83.9	80.6	96.8	90.3	83.9	74.2	87.1	67.7
s2 2019	67%	94.7	73.7	84.2	84.2	94.7	89.5	73.7	78.9	89.5	63.2
s1 2020	80%	90.9	90.9	100.0	100.0	90.9	100.0	72.7	90.9	90.9	27.3
s2 2020	90%	83.3	100.0	100.0	91.7	83.3	100.0	83.3	66.7	100.0	58.3
s1 2021	n/a	100	93.8	100.0	100.0	93.8	100.0	100.0	100.0	100.0	75.0
s2 2021	n/a	77.8	88.9	100.0	100.0	100.0	88.9	77.8	88.9	100.0	100.0
Mean	-	95.4	77.5	94.9	92.6	96.1	94.1	84.9	84.9	94.4	63.8
SD	-	7.0	18.9	8.5	7.3	4.7	7.8	10.2	10.7	5.6	20.5

SD: Standard deviation

C Not fit for purpose: “The majority of competent general ophthalmologists, they would not pass”

The theme “the majority of competent general ophthalmologists, they would not pass” related to the opinion that some of the questions in the examinations were not reflective of cases typically and more regularly encountered in clinical practice and that in turn, a subspecialist rather than generalist knowledge standard was being evaluated. The clinical examination was largely felt to reflect clinical curriculum standards, with an overall average of 96% (range 83%-100%) of respondents agreeing or strongly agreeing with the statement (Table 2). However, there was greater variation in relation to the written examination, with an overall average of 89% (range 72%-100%) of survey respondents agreeing (Table 1) or strongly agreeing that the written examination related closely to curriculum standards. Survey respondents also expressed varying levels of agreement that examinations related closely to clinical/theatre experiences. Again, greater agreement was generally seen for the clinical examination, with an overall average of 85% (range 69%-100%) agreeing or strongly agreeing that the clinical examination tasks were closely related to clinic/theatre experiences during training (Table 1). However, the relevance of the VSAQs to clinic/theatre experiences was the second most consistently problematic component of RACE with an average of 50% (range 38%-62%) of survey respondents reflecting that the VSAQs related closely to their clinic/theatre experiences (Table 1).

Real life training as a registrar in ophthalmology is VERY different to the expectations of RACE (Survey respondent semester 1, 2019).

Questions did not seem to test ability as a general ophthalmologist, many were subspecialty approaches (Survey respondent semester 2, 2019).

Interviewees shared that the clinical examination typically better reflected training experiences given that they included more ‘bread and butter’ cases. However, interviewees did still share criticisms regarding the clinical examination, with some exams felt to present a disproportionate number of rare cases not encountered in training. There was also concern that subspecialty examiners were responsible for assessing stations in their own area of expertise, which trainees believed led to an expectation that responses needed to be well above that expected of a generalist ophthalmologist. Several interviewees believed that the written examination failed to assess candidates at the standard of a competent general ophthalmologist. Interviewees described that in recent years, the examination has increasingly focused on esoteric conditions or complex cases which required subspecialist ophthalmic knowledge. Trainees highlighted that many of the rare conditions presented in the written exam had not been encountered in the training setting, and therefore, the exam was felt to be testing their approach to hypothetical patients requiring ophthalmic care rather than their real-world clinical knowledge.

If you learnt it from the perspective of ‘what would you do as a competent general ophthalmologist?’, that’s not what’s being tested in RACE. I guarantee you if you showed that exam to the majority of competent general ophthalmologists, they would not pass (Trainee interview #17).

I think it has to come back to the core competencies and not try and play clever with esoteric, abstract, academic knowledge that none of us even have. So I think it needs to get back to basics and it needs to be following the same ethos that the OSCE does, which is you are not being examined to be the next chair of Harvard Medical School, right. You are being examined to be a competent suburban ophthalmologist and that’s the standard to which you will be assessed (Supervisor interview #5).

D Time pressure: “It was extremely time-pressured”

Finally, the theme “it was extremely time pressured” referred to the message from survey respondents and interviewees regarding time pressures during the examinations. These pressures were more prevalent in some semesters, with wide variability in the percentage of survey respondents by semester/year who agreed or strongly agreed that the clinical tasks were easy to complete within the allocated time. Overall, an average of 64% (range 27%-100%) of survey respondents agreed or strongly agreed with the question for the clinical examination (Table

2). The lowest percentage was in semester 1, 2020 where only 27% of respondents agreed or strongly agreed with the question. Notably, this aligned with the first virtual clinical examination that was impacted by COVID-19 pandemic restrictions.

To more accurately assess a candidate's ability to properly assess and manage, definitely need more time per station. Six minutes (inclusive of describing further examination) feels like a rapid dash to the finish rather than a considered discussion (Survey respondent, semester 1, 2020).

There was also wide variation in the percentage of respondents who agreed or strongly agreed that it was easy to complete the VSAQs within the allocated time (Table 1). Overall, an average of 66% (range 24%-100%) of respondents agreed with this question, with the lowest percentage (24%) in semester 2, 2013 and the highest (100%) in semester 2, 2017. The time constraints of the SEQ component were the most consistently problematic component of RACE between 2013 and 2021. Overall, an average of 32% (range 14%-63%) of respondents agreed or strongly agreed that the SEQs were easy to complete within the allocated time. The lowest percentage was in semester 2, 2013, when only 14% of respondents agreed with the question.

Given the inherent time constraints, there were concerns expressed by interviewees who had recently sat the examination in an online format. Interviewees described the impact of both typing speed and accuracy, as well as technological glitches, in being able to respond to written examination questions within the short time frame provided. As such, some felt that the examination was more a test of typing proficiency as opposed to ophthalmic knowledge.

I wasn't very good on the computer, typing. So the first thing I did was actually a typing course ... I think that made a big difference, because there's a huge time pressure and if you're not a touch typist, I don't know how you would do it (Trainee interview #2).

The overall (for all available years) mean feedback scores for the same questions asked across the clinical and written examination surveys were investigated to identify the components that have most frequently caused difficulties for candidates (Table 3). This identified the time constraints, particularly of the SEQ component, but also the clinical examination, as the most consistently problematic components of RACE.

Table 3.
Overall comparative means for survey questions in the clinical examination and written examination feedback survey

Semester	Clinical Examination Mean (SD)	Written Examination Overall Mean (SD)	Very Short Answer Questions (VSAQs) Mean (SD)	Short Essay Questions (SEQs) Mean (SD)
Examination process ran smoothly	95.4 (7.0)	90.1 (10.0)	n/a	n/a
Examination related closely to Clinical Curriculum Standards	96.1 (4.7)	89.2 (8.5)	n/a	n/a
Examination pitched at appropriate level of difficulty	94.1 (7.8)	n/a	72.6 (16.1)	75.3 (12.2)
Tasks closely to clinic/theatre experience	84.9 (10.2)	n/a	49.9 (9.9)	69.8 (14.2)
Tasks clearly phrased and unambiguous	84.9 (10.7)	n/a	88.0 (10.4)	72.4 (9.2)
Easy to complete tasks in time	63.8 (20.5)	n/a	65.9 (19.9)	32.4 (13.4)

SD: Standard deviation

IV DISCUSSION

Previous MET data has identified high rates of failure on this examination among ophthalmology trainees (Department of Health and Aged Care, 2020). This problem is not unique to ophthalmology. High failure rates have also been described among trainees sitting summative examinations for the Royal Australian College of General Practitioners (Parker et al., 2021), the Royal Australasian College of Physicians (Wilkinson et al., 2008) and the Australasian College for Emergency Medicine (Rogers et al., 2004). International research has identified high failure rates among registrars sitting summative specialist examinations for the UK Royal College of Physicians (Unwin et al., 2018) and a trend of decreasing pass rates among residents sitting the American Board of Surgery Certifying Examination (Falcone, 2014). Our research sought to identify the reasons for examination failure, by exploring trainee and supervisory experiences of final summative assessment for ophthalmology in Australia and New Zealand, the RACE. The analysis of survey data from sitting candidates between 2013 and 2021, interview data from candidates who had sat recently (>2017) and ophthalmology Fellows who support candidates to prepare for the examination, identified four main themes relevant to the examination construct, design, delivery that may explain examination failure.

The first theme, “I read that paper, it’s so hard”, related to the common perception among survey respondents and interview participants that both the clinical, and more so written examinations, were sometimes excessively difficult. Inconsistent examination difficulty is concerning as it suggests the examination lacks reliability in testing curriculum knowledge (defined in this context as the consistency with which an examination conducted on different occasions assesses clinical knowledge/skills). This is not unique to RACE, for example, there were concerns about a clinical examination investigated by the Royal Australasian College of Physicians (Wilkinson et al., 2008). Reliability issues have also been described in relation to the recently implemented ‘Hot Case’ final summative examination within the College of Intensive Care Medicine (Hoffman et al., 2022). Examination constructs with poor reliability are unfair for candidates sitting during semesters when the examination has a higher level of difficulty. Training colleges can assess the reliability of examinations using statistical methods such as the reliability coefficient or the Standard Error of Measurement, which measures the collective performance of the items used in an examination. However, these methods may not be appropriate in colleges with small numbers of examination candidates each year. Reliability can be improved by increasing the number of items in an examination (Tighe et al., 2010), although this may be prohibitive when the examination already includes a large number of items. It may also not be feasible due to financial costs, time constraints and opportunity costs. In addition to reliability analyses, the Angoff standard setting method can be used to ensure the standard of difficulty of an examination is the same every time it is held. This method is used by training colleges in Australia and New Zealand (Karcher, 2009) and the United Kingdom (McAleer & Grieve, 2011), and has recently been implemented by RANZCO, to address the issue of variable examination difficulty and to improve fairness.

The second main theme, “I don’t know what you guys are asking me”, encompassed the ambiguity of some questions and uncertainty about the level of detail to provide in answers. Again, this was perceived to be more of an issue within the written examination given that the clinical examination provides the opportunity for candidates to receive some guidance from examiners regarding the appropriateness and adequacy of their answers. RANZCO has sought to address concerns about ambiguous questions by implementing a thorough review of each examination question prior to its appearance in the actual exam. First, examiners write their questions and have them reviewed by a second examiner. Second, examiners then meet in small groups to discuss, review and edit each question with an emphasis upon asking “is this question clearly phrased or is it ambiguous?” This group consists of examiners who are not sub-specialists in the particular topic of the question, with an aim of ensuring the question is set at the appropriate level of general ophthalmologists. Third, an additional group of examiners (at least 5) who were not part of the review process in the second stage are asked to independently review and Angoff rate each question; any lingering concerns about the presence of ambiguous wording can be

addressed at this step. Fourth, the Chair of the Examining committee reviews all questions for the exam prior to their publication in the examination and will make any final necessary adjustments.

Specialist medical colleges internationally have implemented continuous quality improvement measures to ensure the rigour of final summative examination questions. Lee et al. (2009) reported that the Joint Faculty of Intensive Care Medicine holds an annual written examination workshop, where Short Answer Questions (SAQs), are discussed together with model answers and structured marking grids, with esoteric questions removed from consideration. The Federation of Royal Colleges of Physicians (United Kingdom) convenes question-writing workshops to generate new questions for the fellowship examination question bank (Twort & Mucklow, 2011). The workshops provide examiners with detailed guidance on question design, format and style, to ensure rigour and comprehension. Draft questions undergo face-to-face peer review, further scrutiny and standard setting prior to being included in the examination.

Pre-examination information sessions can help address candidate uncertainty about the level of detail to provide in answers. Providing candidates with the scoring rubrics that will be used by examiners can reassure trainees about the breadth and depth of information required to pass high-stakes examinations. It also serves to improve the transparency of the examination process (Hoffman et al., 2022) and may reduce failure rates. The increasing expectation of transparency related to marking is evidenced by a recent petition by trainees to the Royal College of Surgeons in the UK, calling for greater transparency after changes were made to the examination pass mark for the final summative examination (Munro, 2020).

The third theme, 'The majority of competent general ophthalmologists, they would not pass', related to the perception that some of the questions in the examinations assessed subspecialist rather than generalist knowledge. This is despite RACE being designed to assess knowledge of the general ophthalmology curriculum (RANZCO, 2022b). This may further explain why candidates perceive some examinations to be more difficult than others, especially if the number of subspecialty questions is greater. Interestingly, there was a high level of agreement (89%) that the clinical examination was closely related to clinic/theatre experience, but this was not the case for the VSAQs (50%), with the SEQs rated mid-way between the other components (70%). This suggests that the clinical examination more consistently assesses general ophthalmology knowledge. RANZCO is addressing this concern in the four-tier process described above. The second stage of the four-tier review process specifically requires examiners who are not subspecialists in the question topic to be part of the question's review in an attempt to bring the level of each question to that of generalist.

The fourth theme, 'It is so time pressured', related to the common perception that both components of the examination were excessively time pressured. Short essay or case report style questions posed the most time-related difficulty as only 32% of survey respondents described the SEQs as easy to complete in the allocated time. This finding echoes the literature among other medical specialties. In the UK, Yeung et al. (2011) found 27% of survey respondents felt there was sufficient time to complete the long cases reporting section of the Fellowship of the Royal College of Radiologists 2B examination, yet 84% reported there was sufficient time to complete the rapid reporting section. Similarly, in their analysis of examination feedback from candidates and stakeholders of the Royal Australian and New Zealand College of Psychiatrists modified essay question exam, Parker et al. (2022) found time management during the examination was one of five key themes related to examination success. The authors noted that candidates 'should actively practise "moving on" or leaving questions when they have reached the time limit for marks allocated'. Specialist medical colleges could therefore help prepare candidates for time-pressured summative examinations by organising mock written examinations, with a focus on the essay style components, in parallel to more common mock clinical examinations.

This research utilised post-examination surveys from trainees sitting the examination. Unfortunately, supervisors of sitting candidates are not currently surveyed post-examination. Supervisors may be able to contribute valuable feedback if they are provided with the questions post-examination. Colleges may wish to include this group in future feedback seeking efforts.

The convergent data analysis identified high levels of agreement across the survey data and interview data. However, there were some areas of partial agreement. For example, there was a strong theme in the interview data that the examination did not reflect the ophthalmology generalist curriculum and was biased towards sub-speciality knowledge. However, this was less prominent in survey data with 95% and 84% respectively of the survey respondents agreeing or strongly agreeing that the clinical examination and written examination were closely related to curriculum standards. However, in the free-text comments, some survey respondents noted that the examination did not seem to reflect the generalist curriculum.

There was one example of disagreement between the survey data and the interview data in the convergent analysis. A high percentage of survey respondents agreed or strongly agreed that the examination questions were clearly phrased and unambiguous, however, the interview data revealed a strong theme of the examination containing ambiguous questions. This indicates that the post-examination survey may need to include an additional item asking respondents to identify ambiguously worded questions.

A Limitations

Both the historical RACE candidate survey data and interview data were derived from motivated respondents who self-selected to participate. The findings of this study therefore cannot claim to be representative of all trainees who have sat RACE in the past five years (or longer for survey data) or supervisors who have supported them to prepare and succeed across Australia and New Zealand. Further, it may be that those trainees who had a negative RACE experience were more motivated to participate in an interview or to return a RACE survey, in turn skewing the results negatively. Increased participation may have provided further insight into other issues that were not identified as part of this research. The findings of the study, particularly the examination experiences of trainees and supervisors elicited during interviews, may not be generalisable to other specialist examinations in Australia or internationally. However, the main findings are likely to have value for medical colleges seeking to improve examination processes.

Another limitation of this study was that survey data was historically collated by RANZCO, and it was noted that over time, changes have been made to RACE candidate feedback surveys. In some cases, previous survey questions have been removed as they were considered no longer relevant, and in other instances questions have been added, leaving gaps in available data over the time period studied.

V CONCLUSION

Examination of trainee experiences sitting RACE have identified possible examination deficiencies that may be contributing to low passing rates. The written component of RACE in particular was perceived to have varying levels of difficulty over time, ambiguously phrased questions, requiring subspecialist knowledge rather than generalist ophthalmology knowledge, and unreasonable time constraints. However, it also should be noted that some participants considered the examinations to be fair.

Concerns about the examinations may be addressed through the implementation of continuous improvement processes. Reliability testing of individual examination items, and the overall examination construct, alongside Angoff standard setting can reduce variations in the level of difficulty of the examination. Greater transparency, including examination information sessions where candidates are provided scoring rubrics and detailed advice about the required answer style and content will address concerns about the ambiguity of questions. Examiner training in question design and the implementation of formalised question validation processes can reduce the risk of ambiguous questions in future examinations. While trainee concerns about examination time restrictions are understandable, it may not be possible to address this by re-designing the examination as this may reduce validity and reliability. Instead, colleges can provide training to candidates and organise mock written examinations so that trainees can learn the best approach

to answering questions expeditiously. Investment in these processes at college level is needed to ensure fairness across different cohorts of trainees and training networks.

References

- Abhary, S., Botti, M., Dhulia, A., Tham, C., Loh, E., & Catford, J. (2021). Factors impacting health and well-being and the utilisation of supports among Australian doctors in medical specialty training. *BMJ Leader*, 5, 26-30.
- Alexander, H. (2015). General practitioners exam: More than half the candidates fail. *The Sydney Morning Herald*. <https://www.smh.com.au/national/nsw/general-practitioners-exam-more-than-half-the-candidates-fail-20150507-ggwgsa.html>
- Ashokka, B., Lee, T. L., & Verstegen, D. M. (2021). 'Exam preparedness': Exploring non-academic predictors of postgraduate exam success. *The Asia Pacific Scholar*, 6(4), 80-91.
- Australian Medical Association. (2022). *2021 AMA Specialist Trainee Experience Health Check*.
- Australian Medical Association Council of Doctors in Training. AMACDT Trainee Forum on College Assessment, April 2021. <https://www.ama.com.au/articles/amacdt-trainee-forum-college-assessment-april-2021>
- Australian Medical Council. (2021). *Effecting reforms to Australia's specialist medical training and accreditation system post COVID-19. Report 4: Changes in assessment in specialist medical programs - opportunities for system improvement*. <https://www.amc.org.au/wp-content/uploads/2021/12/2021-11-AMC-Project-Report-4-Changes-in-assessment-in-specialist-medical-programs-Final.pdf>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research In Psychology*, 3(2), 77-101.
- Dave, S., Chakravorty, I., Menon, G., Sidhu, K., Bamrah, J., Mehta, R. (2020). Differential attainment in summative assessments within postgraduate medical education & training. *Sushruta J Health Policy & Opin*, 13(3). <https://doi.org/10.38192/13.3.15>
- Department of Health. (2018). *Australia's Future Health Workforce - Ophthalmology*. <https://www.health.gov.au/resources/publications/ophthalmology-australias-future-health-workforce-report?language=en>
- Department of Health and Aged Care. (2020). Medical Education and Training (MET) 5th edition. <https://hwd.health.gov.au/met-primary/index.html>
- Department of Health. (2021). *National Medical Workforce Strategy 2021–2031*. <https://www.health.gov.au/our-work/national-medical-workforce-strategy-2021-2031>
- Falcone, J. L. (2014). A longitudinal study of the American Board of Surgery examination index: Can residents pass both the qualifying examination and certifying examination the first time? *Journal of Surgical Education*, 71(4), 455-458.
- Henry, E. (2021). What are the factors affecting performance in high-stakes postgraduate medical examinations? *International Conference of Education, Research and Innovation 2021 Proceedings*, 9025 DOI:10.21125/iceri.2021.2074
- Hoffman, K. R., Nickson, C. P., Ryan, A. T., & Lane, S. (2022). Too hot to handle?: Assessing the validity and reliability of the college of intensive care medicine "hot case" examination. *Critical Care and Resuscitation*, 24(1):87-92.
- Jessup, B., Allen, P., Khanal, S., Baker-Smith, V. Graham, B., & Barnett, T. (2022). *Promoting Success in the RANZCO Advanced Clinical Examination (RACE)*. Centre for Rural Health, University of Tasmania. <https://hdl.handle.net/102.100.100/502836>
- Karcher, C. (2019). The Angoff method in the written exam of the College of Intensive Care Medicine of Australia and New Zealand: Setting a new standard. *Critical Care and Resuscitation*, 21(1), 6-8.

- Khanna, R., Matta, G., Visa, B., & Ryan, A. (2018). Demographics and study practices of psychiatry trainees sitting the RANZCP written examinations. *Australasian Psychiatry*, 26(5), 551-555.
- Kleiven, H. M., Tejani, N., & Sullivan, L. (2016). What is the viva experience of phase 2 Radiation Oncology examination candidates? Survey and advice for future candidates. *Journal of Medical Imaging and Radiation Oncology*, 60(3), 428-432.
- Lee, R. P., Venkatesh, B., & Morley, P. (2009). Evidence-based evolution of the high stakes postgraduate intensive care examination in Australia and New Zealand. *Anaesthesia and Intensive Care*, 37(4), 525-531.
- McAleer, J. J., & Grieve, R. J. (2011). The role of the final FRCR examination in specialist training in clinical oncology. *Clinical Oncology*, 23(8), 499-502.
- Munro, C. (2020). Trainees call for transparency over raised pass marks for Royal College of Surgeons exam. *British Medical Journal*, 25, 371.
- O’Cathain, A., Murphy, E., & Nicholl, J. (2010). Three techniques for integrating data in mixed methods studies. *British Medical Journal*, 17, 341.
- Parker S (2021). Failure rate of doctors sitting the GP fellowship exams and how to beat the statistics. *Modmed*. <https://www.modmed.com.au/articles/failure-rate-of-doctors-sitting-the-gp-fellowship-exams-and-how-to-beat-the-statistics/>
- Parker, S., Freier, K., Gallagher, E., Cutbush, J., Suetani, S. (2022). Working towards success in the Royal Australian and New Zealand College of Psychiatrists Modified Essay Question Exam. *Australasian Psychiatry*, 30(1), 126-31.
- QSR International. (2022). Nvivo Qualitative Data Analysis Software, release 1.7.1.
- Sami, M. B. (2017). Does the CASC examination meet the standard? Assessing utility for a high stakes postgraduate psychiatry clinical assessment. *Kent Journal of Psychiatry*, 1(3). <https://www.kjpsych.co.uk/september-2017/does-the-casc-examination-meet-the-standard-assessing-utility-for-a-high-stakes-postgraduate-psychiatry-clinical-assessment>
- The Royal Australian and New Zealand College of Ophthalmologists. (2018). *Trainee Progression Policy* version 3. https://ranzco.edu/wp-content/uploads/2018/11/Trainee-Progression-Policy_May2018.pdf
- The Royal Australian and New Zealand College of Ophthalmologists. (2022a). *RANZCO Annual Evaluation Report*. https://ranzco.edu/wp-content/uploads/2022/02/RANZCO-Annual-Evaluation-Report_Education.pdf
- The Royal Australian and New Zealand College of Ophthalmologists. (2022b). *Vocational Training Program Curriculum Overview*. <https://ranzco.edu/wp-content/uploads/2022/12/1.-Vocational-Training-Program-Curriculum-v.-16.12.2022.pdf>
- Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP (UK) examinations. *BMC Medical Education*, 10(1), 1-9.
- Unwin, E., Potts, H. W., Dacre, J., Elder, A., & Woolf, K. (2018). Passing MRCP (UK) PACES: A cross-sectional study examining the performance of doctors by sex and country. *BMC Medical Education*, 18, 1-9.
- Wilkinson, T. J., Campbell, P.J., & Judd, S.J. (2008). Reliability of the long case. *Medical Education*, 42, 887-893.
- Willett, L. L., Halvorsen, A. J., Adams, M., Chacko, K. M., Chaudhry, S., McDonald, F. S., Oxentenko, A. S., Swenson, S. L., Zaas, A., & Arora V. M. (2016). Factors associated

with declining residency program pass rates on the ABIM certification examination. *The American Journal of Medicine*, 129(7), 759-65.

Winter, R., Norman, R. I., & Patel, R. (2021). A qualitative exploration of the lived experience of GP trainees failing to progress in training. *Education for Primary Care*, 32(1), 10-18.

Yeung, A., Booth, T.C., Jacob, K., McCoubrie, P., & McKnight, L. (2011). The FRCR 2B examination: A survey of candidate perceptions and experiences. *Clinical Radiology*, 66(5), 412-419.