The development of an evidence-based, practical and contextualised workplace-based assessment framework for general practice training and education

GPEX

In collaboration with

GP training

Flinders University

general practice training tasmania

MURRAY CITY COUNTRY COAST GP TRAINING

NORTHERN TERRITORY General Practice Education

RVTS REMOTE VOCATIONAL TRAINING SCHEME LTD

WAGPET

GP SYNERGY

GPTQ

MODMED INSTITUTE

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

GPEx would also like to acknowledge:

- Assoc Prof Jill Benson AM (GPEx) for leading the research team who completed this project.
- Prof Lambert Schuwirth (Flinders University) for providing academic guidance and input into all components of this project.
- The GPEx research team members: Dr Emily Kirkpatrick (GPEx), Mr Shaun Prentice (GPEx), Dr Taryn Elliott (GPEx), Ms Michelle Pitot (GPEx) and Ms Kiara Beens (GPEx).
- The collaborating Australian general practice training organisations: Eastern Victoria GP Training (EV), General Practice Training Tasmania (GPTT), Murray City Country Coast GP Training (MCCC), Northern Territory General Practice Education (NTGPE), Western Australian General Practice Education and Training (WAGPET), Remote Vocational Training Scheme (RVTS), GP Synergy and General Practice Training Queensland (GPTQ) for their support and contributions to the research and the final Framework.
- ModMed Institute for supporting and contributing to this project through their authorship of Steam 3c: Evaluation of Entrustable Professional Activities.
- EV for their authorship of Steam 3a: GP Registrar, Supervisor and Medical Educator use of Consultation Observation as an Educational and Assessment Tool.
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.

GPEx

In collaboration with

GP training · Flinders University · general practice training tasmania · MCCC Murray City Country Coast GP Training · Northern Territory General Practice Education

RVTS Remote Vocational Training Scheme Ltd · WAGPET · GP Synergy · GPTQ · ModMed Institute

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Contents

Executive Summary

Workplace-based assessments (WBAs) have proliferated in medical education with the premise that, through their use, trainees will learn to provide safe and effective health care. This premise rests on the unique advantage WBAs have over other assessment types: they permit the direct assessment of a trainee's actual performance rather than inferring this through knowledge or simulation. This report details a comprehensive project commissioned by the Royal Australian College of General Practitioners (RACGP) to produce a Workplace-based Assessment Framework for Australian General Practice Training.

For the purpose of this framework the following definition was used:

"*WBAs are the tools and the processes used in the collection of data about the trainee's performance in the workplace, and the judgement of their competence by an assessor, for a range of summative and/or formative purposes."*

In order to produce the WBA Workplace-based Assessment Framework for Australian General Practice Training, we embarked on a series of simultaneous research projects, designated as 'streams'. Concurrently an environmental scan was completed to identify: other national and international WBA Frameworks and models, and the relevant standards and policy context within Australian General Practice training. Both the research streams and the environmental scan were used to guide the Framework design.

The project was governed by a Steering Group including representatives from Eastern Victoria GP Training, General Practice Training Tasmania, Murray City Country Coast GP Training, Northern Territory General Practice Education, Western Australian General Practice Education and Training, Remote Vocational Training Scheme, GP Synergy and General Practice Training Queensland. Prof Lambert Schuwirth from the Prideaux Institute at Flinders University provided expert guidance and input across all project streams. Prof Cees Van der Vleuten provided feedback and support to the project.

The six research streams completed concurrently are described below.

Stream 1 involved a hermeneutic review of the postgraduate medical education literature in order to understand why WBAs are viewed as important forms of assessment; the attributes, perceived acceptability, and effectiveness of WBAs; the role of assessment literacy for trainees, supervisors, assessors and training organisations; and the interactions between the WBA tools and the trainee, supervisor, medical educator (ME) and training organisation.

Stream 2 comprised an audit of the WBAs in Australian General Practice Regional Training Organisations and the RVTS (RTOs) - who uses them; how often they are used; for what purpose; the challenges associated with their use from the perspective of the registrar, the supervisor and the RTO; the specific roles of the ME and training coordinator that facilitate and monitor their use; and how the WBAs map to the RACGP Domains of General Practice and Core Skills.

Under ***Stream 3***, four distinct primary studies were conducted:

- ***Stream 3a*** was a qualitative investigation into consultation observation looking at what enables and impedes direct observation of registrar consultations as a WBA; the contextual factors that impact on its utility; and how direct observations should be framed and supported in order for them to be a successful educational tool as well as an informant of registrar progress.

- In ***Stream 3b***, focus groups and interviews were conducted with registrars, supervisors and MEs in all participating RTOs (n= 127) to explore their attitudes and perceptions of WBAs; the challenges and opportunities WBAs provide for assessment and learning; how RTO structures and characteristics impact the utility of WBAs; and whether WBAs are useful as part of programmatic assessment.
- ***Stream 3c*** focussed on entrustable professional activities (EPAs) as tools for assessment, self-reflection, teaching, learning and progression. Quantitative and qualitative data was analysed from mid and end-term assessments in the three RTOs who have used EPAs; and focus groups and interviews with registrars, supervisors and MEs explored how EPAs were being utilised, their barriers and facilitators, how entrustment decisions were made, and whether they were useful for learning and teaching.
- Finally, ***Stream 3d*** aimed to explore workplace-based assessments as predictors of exam performance and remediation. A review of published and unpublished Australian research revealed limited available data to explore this question. Therefore the Steering Group adapted the focus to explore associations between flagging through the use of WBAs against exam performance, which was able to be explored through a combination of meta-analytic and qualitative techniques. Hence this project stream examined the association between exam performance and the different models of flagging in Australian RTOs. The quanitative data from seven RTOs about flagged registrars mapped to RACGP exam results was collated and analyse. Interviews with key personnel from each participating RTO about their flagging and remediation processes combined with data from Stream 3b focus groups and interviews was analysed to build a rich picture of flagging and remediation processes in participating RTOs. From this picture, elements were identified which were similar and different across each RTO flagging model. These elements were used to inform statistical analysis to examine the strengths of different flagging models.

The environmental scan used six key documents from within the postgraduate medical training environment from Australia, Canada, New Zealand and the United Kingdom that assisted to inform appropriate structure and content for the development of the Framework. The environmental scan of relevant standards and policy reviewed a range of relevant RACGP documents including the RACGP Standards for Training Providers (2nd edition), RACGP Curriculum (2016) and the RACGP Competency Profile. Relevant documents from the Australian Medical Council and the Australian College of Rural and Remote Medicine were also reviewed. The Working Group determined relevant information from the document review which should be integrated into or referenced within the WBA Framework.

From these research streams, and the environmental scan, we developed the Workplace-based Assessment Framework for Australian General Practice Training. The basis of this Framework follows the themes of the WBA tools, the RTO context, the assessors and the registrars.

WBA tools have core elements such as the use of narrative, entrustment scales, and a clear and standardised purpose. They should allow for intuitive decisions and comments to be made by assessors, and progress should be able to be tracked by all stakeholders. Feedback is essential and should be timely for all face-to-face WBAs and then involve follow-up if necessary. It should be documented online in an easy-to-use and reliable system that is accessible to the registrar, supervisor and ME. All WBAs should be followed by registrar self-reflection and action. This should be documented as part of the learning log. The purpose of the tools (summative or formative) should be clearly defined and the process for their implementation, feedback, follow-up and consequences documented and communicated to all those involved.

A programmatic assessment approach should be utilised with collation of outcomes and opinions about WBAs in order to map progress and identify registrars in difficulty. WBAs that are mandatory should have follow-up by the training coordinator and accountability for satisfactory completion. Registrars in difficulty who require a higher level of supervision will require more WBAs.

Multiple assessors should be used – some with an ongoing relationship with the registrar, and some without such a relationship. High-stakes assessments should be undertaken by an external assessor and not by the registrar's supervisor or someone of their choice.

There needs to be a clear process in place to determine and document flags appropriately, which is transparent and communicated to the registrar, supervisor, ME and training coordinators. Assessors should be trained in the possible reasons for flagging and the pathways and outcomes expected in personal, practice and professional spheres. The ME assigned to follow the longitudinal pathway of the registrar should be responsible for the communication, monitoring and implementation of any intervention plan associated with flagging with the support of templates and of a Training Review Committee of senior MEs.

The Training Organisation should ensure there is an ME and training coordinator assigned to each registrar preferably for their entire training time. These personnel are responsible for communication, monitoring, signing-off and collation of WBAs. They are integral to liaison with practices and supervisors and to the flagging process. A culture of excellence and extension of high performing registrars should also be encouraged by the RTO. Supervisors and registrars should have quarantined funded teaching time as part of the training program. RTOs should ensure that supervisors who are providing a higher level of support for registrars are remunerated and supported appropriately.

The Assessors should be trained to ensure a high quality and consistent approach to the delivery of WBA assessments and feedback. 'Communities of practice' for both supervisors and MEs will serve the purposes of mutual support, 'tips and tricks', as well as for formal training. MEs should have higher medical education qualifications and be encouraged to support supervisors in their external clinical teaching visits. Supervisors should be trained in giving constructive, outcome-based feedback, how to support struggling as well as excellent registrars, the purpose and principles of WBAs and programmatic assessment, and in the use of particular tools such as direct observation, random case analysis and EPAs.

Registrars should have a structured and contextualised orientation to training and to each practice that includes the purpose, timelines, implementation, monitoring, accountability and escalation process for WBAs. An initial four to eight-week safety assessment to assess the level of supervision needed to ensure patient safety will be followed by regular assessments using entrustment scales, most of which are formative as part of the registrar's learning pathway, but all of which can trigger a flag if the registrar is struggling. Part of registrar training will be in the skills of asking for feedback, accepting feedback, self-reflection and life-long learning. Completion of training will include a Statement of Awarded Responsibility (STAR) assessment and a final sign-off of the WBA portfolio.

The final WBA Framework is evidence-based, practical and contextualised. The Framework rationale is clearly outlined and the recommended tools and processes are described based on the project findings. The Framework is fit for purpose and has flexibility to meet the local contextual needs of the RTO and practice. It draws on existing, updated and new WBA tools and processes, and incorporates the latest evidence to provide recommendations for use.

Introduction

Dr Taryn Elliott (GPEx)
Assoc Prof Jill Benson AM (GPEx)
Dr Emily Kirkpatrick (GPEx)
Mr Shaun Prentice (GPEx)
Prof Lambert Schuwirth (Prideaux Centre, Flinders University)

Project context

GPEx was awarded a special Education Research Grant by the Royal Australian College of General Practitioners (RACGP) to design a Workplace-based Assessment (WBA) Framework for use within Australian General Practice training. This is the first WBA Framework developed for use within Australian General Practice training and as such a robust, evidence-based and consultative approach was required. The chapters in this report describe the project literature and background, the research and consultation undertaken and the process used to inform the design of the Framework.

A historical view of assessment within medical education

Up until the 1970s, assessment of clinical performance was based mainly on ad-hoc and ward-based clinical examinations. Although the authenticity and real-practice context of such assessments were valued, they were also seen as subjective and unreliable, and sometimes as biased. This led to the development of the objective structured clinical examination (OSCE). The OSCE aimed at counteracting the subjectivity and unreliability in three ways.

1. By keeping the stations short, many different tasks could be performed during an examination and that way a broad sampling of various domains could be obtained, which counteracted the problem of domain specificity.
2. By requiring the candidates to rotate from station to station, they would be judged on each station by a different assessor in a so-called nested design, thus increasing the sample of assessors, adding further to the reliability.
3. By structuring the assignments and the score rubrics in the original case checklists, it was hoped to counteract subjectivity, hence the word 'objective' and the acronym.

OSCE development led to a range of research, especially into factors that enhance or limit the reliability or generalisability of clinical examination scores (eg. Harden & Gleeson, 1979, Swanson & Norcini, 1989). Key findings from this research, which then informed the development of workplace-based assessment (WBA) are as follows:

- ***Assessment is never objective***
  The awareness emerged that assessment is never objective, and nor should it be. Any type of assessment involves the collection of information about the trainee's performance or competence and a value judgement. Even decisions around what information to collect, how to interpret that information, how to summarise that information, and when to collect the information, are processes of human judgement, which will always have a subjective element. So even what can be seen as the most 'objective' type of assessment, the multiple-choice test, is the culmination of a whole series of subjective judgements made by those who have written the exam questions.
- ***Generalisability does not rely on inter-observer reliability***
  Generalisability studies showed that domain specificity, or inter-case unreliability, impacts much more on the generalisability of scores than inter-observer unreliability.
- ***Standardisation did not improve generalisability***
  Given the previous two points, structuring and standardisation did not do much to increase generalisability, but were seen to have a huge negative impact on (face) validity. Studies showed, for example that more global rating scales produced better reliability of total scores than more detailed checklists.

These findings have formed the basis for the introduction of WBA as a gold-standard for assessment within medical education.

Workplace-based assessment in medical education

The first example of the reintroduction of WBA after the initial development of the OSCE, was the mini-CEX by John Norcini (1995). There were two perceived advantages of this reintroduction. The authenticity of the context would be better than the artificial nature of the performance in the OSCE context, and WBA seems to allow for a more holistic and integrated assessment of medical competence. Miller's (1990) pyramid may explain this. The top level 'does' is generally what assessment programs are aiming to assess. This ability of the top-level could of course be inferred from a student's performance on the other three levels ('shows how', 'knows how' and 'knows'), but that is a rather indirect assessment approach. WBA in the authentic context allows for a more direct assessment of this top-level, and from this, it can be inferred that the lower levels must be satisfactory as well.

Since the introduction of the mini-CEX many different WBA tools have been developed – such as objective structured assessment of technical skills (OSATs), 360° feedback (e.g. multisource feedback), direct observation of procedural skills (DOPS) etc. The common feature of all these instruments is a direct observation of a candidate's performance and a judgement on various aspects of that performance using a rubric.

Initial research into WBA has focused on construct validation and reliability from a more or less traditional psychometric framework. However, since the mid-1990s awareness has arisen that the assessment quality of WBA lies not so much in the rubric or the method, but that the assessor plays a vital role, because of their content and assessment expertise and experience. One of the first researchers to study this was Marjan Govaerts (2007, 2011, 2013), who made the link between the cognitive activities of diagnosing a disease and diagnosing 'competence'. Her studies showed that script theory applies to the cognitive processes of the assessor as it does to the cognitive processes of a diagnostician. Weller et al. (2014) demonstrated that by changing the WBA instrument to better match the expertise and experience of the assessor resulted in dramatic improvement of the psychometric properties of the WBA.

Since then, research has tried to better understand the interaction between the assessor, the trainee and the WBA tool as a mediator. Currently, validity of WBA is generally seen as the quality of the interplay between these factors rather than as a quality of the WBA tool in isolation. Therefore, to understand WBA all perspectives must be considered; the perspective of the *assessor*, the perspective of the *trainee*, the perspective of the *tool* and how well it is embedded in the *context* in which the assessment takes place.

Workplace-based assessment in Australian General Practice Training

Australian General Practice Training (AGPT) is currently delivered through a network of training organisations. The collaborating training organisations included eight of the nine regional training organisations (RTOs) and the Remote Vocational Training Scheme (RVTS). This regionalised approach takes into account the significant unique regional variations in Australia including, but not limited to, differing geography, communities and health systems.

To ensure a consistent quality in general practice training, each training organisation is accredited by both the Royal Australian College of General Practitioners (RACGP) and the Australian College of Rural and Remote Health (ACRRM) to deliver training according to the Standards set by the College/s. Training must also be delivered according to the AGPT program policies. However, the training models used by each training organisation vary considerably in order to cater for particular features of their region (e.g. urban/rural/remote), their allocated budget, and the available staffing. This has resulted in a variety of different models for the delivery of WBAs. Anecdotally there are differences and similarities between training organisations regarding:

- The types of WBAs used.
- The purposes for which WBAs are used.
- The processes used for support, monitoring and regulation of WBAs.
- The context within which the WBA is administered.
- The training and support provided to assessors.

There have been a number of research projects focussing on various aspects of WBA across training organisations, both published and unpublished, but no comprehensive synthesis of the literature within this context.

There is currently no shared understanding about what the most appropriate WBA Framework for general practice training should contain. As we progress towards a transition to College-managed delivery of the AGPT program, it is important to understand the current models of WBA used in Australian general practice, in order to plan for the future.

The gap

An environmental scan of WBA within Australian General Practice Training shows there is currently no shared understanding of the most appropriate WBA system. The development of an evidence-based WBA Framework would fill this gap.

However, in order to develop an evidence-based Framework, which is also feasible and acceptable, further research is required to fill the gaps in our understanding of the literature and the environment. The gaps are as follows:

- Most WBA research has been theoretical, with many experimental or medical school-based studies with only a few exploring the issue at the coalface by studying what actually happens in the real-life WBA situation.
- Those studies that do try to capture the real-life situations often focus mainly on the instrument and see the users as a factor that merely has to be instructed how to use the instrument.
- A smaller number of studies focus predominantly on the assessment literacy of the user without taking into account the other assessor qualities that are important.
- There have not been any studies within WBA which have captured input from registrars, medical educators and supervisors from across the whole of Australia.
- While there have been a number of studies focusing on aspects of WBA across training organisations, there has been no published synthesis of this literature.

Our project will aim to develop an evidence-based, practical and contextualised Workplace-based Assessment Framework for Australian General Practice Training.

The development of this Framework will be underpinned by a series of research projects which, address each of these identified gaps. The research will be authentic and practical and aimed towards the interplay between the assessor, the trainee, the context and the tools.

The project

Background

In November 2018, the RACGP awarded GPEx a Special Education Research Grant with the aim to design a WBA Framework for use within Australian general practice training.

The grant required the findings from each project stream to be used to inform the development of an evidence-based, practical and contextualised Workplace-based Assessment Framework for Australian General Practice Training (the Framework) which:

- Is fit for purpose (meets local contextual needs)
- May be existing, updated or new (based on findings).
- Clearly outlines the rationale for the framework and processes based on findings.

This report describes each of the project streams, the results from each stream, and the process through which these results were synthesised to develop the Framework.

Assumptions

There is a wide regional and contextual variation in the use of WBAs across Australian general practice training organisations. Therefore, an important underpinning assumption for the Framework is that it advocates for the quality of WBAs to be standardised across Australia rather than standardisation of the method in which assessments are performed. Not every registrar needs the same set of WBAs but every registrar is entitled to the same quality of assessment. There are other assumptions, which have been adopted to inform the Framework development, based on the synthesis of project results. The primary purpose of WBAs should be to give feedback and encourage self-reflection and learning. A secondary purpose should be to flag registrars who are struggling to meet the expected standard. WBAs should allow supervisors, MEs and the training organisation to assess the level of supervision needed to ensure registrar and patient safety within a larger programmatic model. The RACGP Curriculum and Standards and AGPT policy form the backbone for the design of this contextualised Framework, and must be considered in the implementation of a WBA system within any regional context.

Overall project method

Project governance

This project was completed in collaboration with Flinders University, Eastern Victoria GP Training (EV), General Practice Training Tasmania (GPTT), Murray City Country Coast GP Training (MCCC), Northern Territory General Practice Education (NTGPE), Western Australian General Practice Education and Training (WAGPET), Remote Vocational Training Scheme (RVTS), GP Synergy and General Practice Training Queensland (GPTQ).

The project was governed by a Steering Group, including representatives from all collaborators and the GPEx research team, which met at least monthly during the project period. Internationally recognised experts in the field of medical education and WBA, Professor Lambert Schuwirth (Prideaux Centre, Flinders University) and Professor Cees Van der Vleuten (Maastricht University) provided advice and input to the project.

Framework creation

The final deliverable from this project, the Framework, was informed by the following inputs:

- ***Literature review:*** A hermeneutic review of the existing published literature on WBAs.
- ***WBA audit:*** An audit of WBA tools currently used in Australian RTOs and RVTS.
- ***Primary research:*** A series of mixed-methods investigations focussed on the effectiveness of WBA tools and the nexus between the tool, the users and the context.
- ***Environmental scan:*** A scan which identified other relevant WBA Frameworks and models, and the relevant standards and policy context within Australian General Practice Training (AGPT).

Overall this project aimed to review the international literature about WBAs; collate how WBAs are currently being used in Australian RTOs and the RVTS; map the tools to the RACGP domains of general practice and core skills (RACGP, 2015); listen to the experiences and perspectives of registrars, supervisors and medical educators (MEs); examine in more detail direct observation visits (DOVs) and entrustable professional activities (EPAs) as assessment tools; and explore the process of flagging in each RTO and whether flagging maps to the RACGP exam results. As such these are the six project research streams:

- Stream 1: Literature review, this included current and past research into competency-based assessment in a workplace-based setting, with a focus on post-graduate medical education.
- Stream 2: Australian general practice training organisation workplace-based assessment audit, incorporating a summary of the types of tools currently in use within the collaborating training organisations footprints. Each tool mapped against the RACGP domains.
- Stream 3a: GP registrar, supervisor and ME use of consultation observation as an educational and assessment tool
- Stream 3b: A qualitative investigation of acceptability of WBAs in AGPT, incorporating a review of the perceived effectiveness of current WBAs from the perspective of MEs, supervisors and registrars.
- Stream 3c: Evaluation of EPAs, including the use of EPAs as a measure of progress.
- Stream 3d: An exploration of registrar flagging models and their association with the RACGP exam performance

The final Framework is a compilation of the above six streams of research, and the environmental scan, that make up this comprehensive project. Further information regarding the method for synthesising the results from the project streams is included in *Chapter 8- Designing a Workplace-based Assessment Framework for Australian General Practice Training.*

Framework recommendations

A series of recommendations were synthesised from the triangulation table, which will be included within the Framework. These recommendations map across the Framework elements: tools, assessors, registrars and context. A complete list of recommendations can be found in Appendix C.

An overview of the report

This report provides an overview of the results from each stream of research, the environmental scan, and details of how the results from each of the inputs were synthesised using a triangulation process to produce the Framework.

Some important definitions to note, which influence the reading of this report are as follows. Firstly, the definition of WBA used for the purpose of this project was:

"Workplace-based assessments are the tools and the processes used in the collection of data about the trainee's performance in the workplace, and the judgement of their competence by an assessor, for a range of summative and/or formative purposes."

Secondly, the literature on assessment often makes a distinction between 'assessment of competence' and 'assessment of performance'. 'Assessment of competence' generally means a form of assessment in which the candidate knows that he or she is being assessed and consequently will perform at the maximum of their ability. 'Assessment of performance' is aimed more at grasping how the candidate will perform on a day-to-day basis rather than on a specific occasion. In workplace-based assessment, this distinction is not as sharp. Candidates may understand that they are being assessed and will behave accordingly – for example in direct observation visits (DOVs)- but sometimes they may be less aware – e.g. with multi-source feedback. There is also a spectrum of levels of authenticity of the context of the assessment – with, for instance, DOVs being highly authentic and learning plans being less authentic. Therefore, in the remainder of this report we will be using the term 'competence' but we are aware that this also encompasses elements of performance.

The report includes the following chapters:

- Chapter 1: Introduction
- Chapter 2: Literature review
- Chapter 3: Australian RTO and RVTS workplace-based assessment audit
- Chapter 4: GP registrar, supervisor and ME use of consultation observation as an educational and assessment tool
- Chapter 5: A qualitative investigation of acceptability of WBAs in Australian General Practice Training
- Chapter 6: Evaluation of EPAs
- Chapter 7: WBAs as a predictor for RACGP exam performance
- Chapter 8: Designing a Workplace-based Assessment Framework for Australian General Practice Training
- Chapter 9: Workplace-based assessment Framework for Australian General Practice Training

This report provides details of the rigorous methods and findings which underpin the development of the final chapter and stand-alone document *Workplace-based Assessment Framework for Australian General Practice Training*.

References

Govaerts, MJB, Van de Wiel, MWJ, Schuwirth LWT, et al. (2013). Workplace-based assessment: raters' performance theories and constructs. *Advances in health sciences education,* 18: 375-96.

Govaerts, MJB, Schuwirth, LWT, Van der Vleuten, CPM, et al. (2011). Workplace-Based Assessment: Effects of Rater Expertise. *Advances in health sciences education*, 16(2):151-65.

Govaerts, MJB, Van der Vleuten, CPM, Schuwirth, LWT, et al. (2007). Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Advances in health sciences education*, 12(2):239-60.

Harden RM, Gleeson FA. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, 13(1):41-54.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance*. Academic Medicine*, 65 (9): S63-67.

Norcini J., Blank L.L., Arnold G. K., Kimball, H.R. (1995). The Mini-CEX (Clinical Evaluation Exercise); A Preliminary Investigation. *Annals of Internal Medicine,* 123 (10): 795-9.

Royal Australian College of General Practitioners (2015). *Competency profile of the Australian general practitioner at the point of Fellowship*. Available from: https://www.racgp.org.au/education/Registrars/fellowship-pathways/general-practice-education-framework (accessed February 2019).

Swanson, DB, Norcini JJ. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine*, 1(3):158-66.

Weller, JM, Misur, M, Nicolson, S, et al. (2014). Can I leave the theatre? A key to more reliable workplace-based assessment. *British Journal of Anaesthesia,* 112(6):1083-91.

Stream 1
Literature Review

Mr Shaun Prentice (GPEx)
Assoc Prof Jill Benson AM (GPEx)
Prof Lambert Schuwirth (Prideaux Centre, Flinders University)
Dr Emily Kirkpatrick (GPEx)
Dr Cat Kirby (Eastern Victoria GP Training)

GPEx In collaboration with

Flinders UNIVERSITY GP training

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

GPEx would also like to acknowledge:

- The Working Group for this stream including: Mr Shaun Prentice (GPEx), Assoc. Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr Cat Kirby (Eastern Victoria GP Training), Dr Taryn Elliott (GPEx) and Ms Michelle Pitot (GPEx).
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.

GPEx In collaboration with

Flinders UNIVERSITY GP training

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Introduction

Historically, clinical assessments in postgraduate medical education were largely ad-hoc and ward-based. This prompted a movement away from such assessment measures so as to improve reliability and objectivity, however gains in reliability were minimal and overwhelmed by perceived reductions in ecological validity. A compromise was reached with the introduction of the mini-CEX by John Norcini [1]. Since this point, workplace-based assessments (WBAs) have proliferated throughout postgraduate medical education. A major appeal of such assessments is the direct assessment of what a trainee does in their workplace, targeting the highest tier of Miller's pyramid [2] for assessing clinical competence (refer Figure 1). This avoids inferring competence on the basis of assessments targeting lower tiers of the pyramid, thereby improving the validity of WBAs relative to other assessments such as multiple-choice questions or objective structured clinical examinations (OSCEs) [3].

Figure 1. Miller's framework for assessing clinical competence [2]

Despite these conceptual advantages, WBAs have received much criticism from trainers and assessors alike [4,5]. Understanding these perspectives on the limitations of WBAs will be imperative to improving them. However, reviews in this field considering various perspectives are limited. Multiple systematic reviews have been conducted, but these have mainly been descriptive [6,7] or used meta-analytic techniques [8]. Although these have provided useful consolidations of the literature, they are unable to offer a deep analysis of the main phenomena of interest, particularly the role of interactions between users, tools, organisations and context. Similarly, reviews have often considered outcomes of WBAs and user perceptions in isolation [6,8] where modern conceptualisations of validity recognise the importance of the interplay between these factors [9-11]. To permit a deeper engagement with this literature examining such interactions, we conducted a hermeneutic literature review to examine WBAs. The aims of the review were:

1. To understand the psychometric and economic properties, effectiveness and perceived acceptability of WBAs tools and methods in postgraduate medical education
2. To understand the interaction between WBA tools and the registrar, supervisor and/or medical educator (ME) in terms of meaning-making and forming valid judgements
3. To understand the required assessment literacy of registrars, supervisors and MEs to optimise the utility of WBAs.

Methods

Theoretical background

A hermeneutic review is based on the philosophy of hermeneutics, which proposes readers come to a text with a 'horizon': their current knowledge of a field. A text also has a horizon which comprises the knowledge it contributes to that field. In reading a text, the reader's and text's horizons fuse to broaden the reader's horizons and facilitate deeper engagement with the texts. This generates a cycle whereby the reader's horizon influences how they approach a text, and the text influences the reader's horizon [12].

In the context of a literature review, a hermeneutic cycle can be used. This involves an iterative cycle of engaging with the literature to explore questions and answers, which in turn prompt seeking of and engaging with more literature. This approach lends itself to a qualitative analysis of the literature, with the hermeneutic cycle continuing until saturation is reached [12].

Procedures

To begin, we developed the following definition of WBAs:

"Workplace-based assessments are the tools and the processes used in the collection of data about the trainee's performance in the workplace, and the judgement of their competence by an assessor, for a range of summative and/or formative purposes[1]."

Based on the three aims of this review, three overarching scoping questions were developed:

1. What are the psychometric and economic properties, effectiveness and perceived acceptability of WBA tools and methods presented in post-graduate medical education literature?
2. What is known about the interaction between WBA tools and the registrar, supervisor and/or ME in terms of meaning-making and forming valid judgements?
3. What is known about the required assessment literacy of registrars, supervisors and MEs to optimise the utility of WBAs?

From these questions, more specific sub-questions were developed (refer Appendix A).
A research librarian developed a search strategy for the following databases: Ovid Medline; Ovid PsycInfo; Ovid Emcare; and ProQuest (including ERIC). The search strategy was designed to identify articles pertaining to WBAs in the postgraduate medical education literature. The country of articles was limited to Australia, New Zealand, the United Kingdom (UK), Canada, the Netherlands and Scandinavian countries so as to maximise the comparability of results. The search strategy is detailed in Appendix B. The search yielded 211 non-duplicate articles.

The titles and abstracts of all articles were screened by one author (SP), who categorised them loosely according to the three scoping questions. He then selected 10 articles [4,5,13-20]. Two coders (JB and SP) independently inductively coded these articles using a grounded theory approach and a consensus meeting was held to discuss coding structures. Based on this coding, specific questions were refined to recognise the emergence of new questions answered or raised by the

[1] Current thinking in medical education is shifting from a 'formative vs summative' dichotomy of assessments to one of 'low', 'medium' and 'high' stakes. However, the reviewed literature has used the terms 'formative and summative'. In the interests of being faithful to this literature, we have therefore decided to use the traditional 'formative vs summative' terminology.

literature. A revised coding structure was developed which fused both authors' coding structures. This coding structure was expressed in a codebook and visual thematic maps.

A further ten articles were selected by one author (SP) for the second round of coding [7,21-29]. These articles were disseminated to a larger group of coders (JB, CK, EK and SP) who were instructed to code using the coding framework from an NVivo template. This coding involved a blend of *a priori* coding using the structure, and inductive coding to revise or add to the coding structure based on the new articles. The four coding structures that arose from this round of coding were reviewed and fused into a new revised coding structure, again documented in a codebook and visual thematic maps. At this point, modifications were made to the scoping questions. Given the importance of WBAs was being explored in the literature reviewed, a new scoping question was added to specifically focus on this ('Question 0'). Furthermore, the concept of 'effectiveness' was moved from scoping question 1 to scoping question 2 in the form of 'outcomes', while the role of the training organisation was added to scoping question 3. The revised scoping questions were as follows:

0. Why are WBAs viewed as important forms of assessment in the postgraduate medical education literature?
1. What are the psychometric and economic properties, and perceived acceptability of WBA tools and methods presented in the postgraduate medical education literature?
2. What is known about the interaction between WBA tools; the registrar, supervisor and/or ME; and the context in terms of outcomes, meaning-making and forming valid judgements?
3. What is known about the required assessment literacy of registrars, supervisors and training organisations to optimise the utility of WBAs?

Additionally, the content coded under each question was explored to identify questions with little or no literature answering them. These gaps were compiled and used as the basis for the selection of the next round of ten articles to code [30-39]. These articles were again disseminated to four coders (JB, EK, SP and LS), along with the revised coding structure. Coders were again instructed to use a combination of *a priori* and inductive coding. The four coding structures from the final round were again reviewed and fused in a consensus meeting. All coders were satisfied with this coding structure and felt that saturation was reached.

Results

Note: for the purpose of the results section, <u>underlined</u> text refers to themes and *italicised* text refers to sub-themes.

Question 0: Why are WBAs viewed as important forms of assessment in the postgraduate medical education literature?

Various perspectives on why WBAs were perceived as important emerged from the literature. The ultimate strength of WBAs was their perceived capacity to ensure the doctors of tomorrow provide <u>high-quality care</u> for their patients [14,35]. WBAs were seen to achieve this objective through a number of strategies. Most fundamentally, WBAs are defined by the <u>assessment of a trainee's performance in an authentic setting</u>. This feature was overwhelmingly identified as a key strength of WBAs [4,7,14,16,17,19,21,22,24,26,28,29,31,32,34,38]. This aligns with Miller's [2] pyramid; by assessing the 'does' level, making inferences about a trainees' performance from lower levels is unnecessary.

Additionally, WBAs were appreciated because they provide an opportunity to assess non-clinical competencies which are felt to be equally important for providing high-quality care [17,26]. Stakeholders emphasised the value of WBAs in facilitating the provision of feedback on a trainee's performance [4,7,18,21,26,27,32,35]. In particular, trainees emphasised how WBAs *'force' supervisors to provide them with feedback*, something they felt was important given the absence of feedback in other domains of training [7,21,28].

Other perspectives on the importance of WBAs identified how they were uniquely capable of facilitating important attributes of the assessment process that are difficult to integrate into other, non-WBA, formats. Stakeholders explained how WBAs permit the assessor to become the instrument of assessment, thereby allowing the assessor's expertise to be utilised [15,17,34]. Stakeholders expressed also how this is difficult to implement in other assessment formats. Furthermore, stakeholders appreciated how WBAs permit a holistic assessment of a trainee's competencies rather than deconstructing it into specific aspects [5,21], and that a trainee's progression over time can be mapped [21]. Finally, because WBAs focus on a trainee's performance, stakeholders felt that WBAs encourage trainees to engage in self-learning [4,21,26,35].

Summary

WBAs are valued by stakeholders because they are perceived to ensure doctors are trained to provide their patients with high quality care. Stakeholders predominantly believe WBAs facilitate this by assessing a trainee in an authentic setting. WBAs were also seen to afford opportunities for assessing non-clinical competencies and providing feedback to trainees. Other strengths of WBAs included that the assessor becomes the assessment instrument; holistic assessments that map trainees' progression over time are permitted; and trainees are encouraged to engage in self-learning.

Question 1: What are the psychometric and economic properties, and perceived acceptability of WBA tools and methods presented in the postgraduate medical education literature?

Economic Properties

Very little information could be found regarding economic properties of WBAs in the literature. At the end of the second round of coding, 'economic properties' was identified as a gap, yet additional information was not identified in the third round. One included study estimated large time (and therefore financial) costs were associated with the implementation of a WBA system [23]. Nevertheless, several issues were raised regarding the broader concept of the feasibility of WBAs. The most commonly reported concern in this respect was regarding the *time requirements* of WBAs and difficulty *organising* these within the constraints of already busy *workloads* [4,5,7,18-22,28,31,32,35]. In particular, the *technological platform* under which WBAs were administered could - paradoxically - increase the time requirements of WBAs [21]. Stakeholders also raised concerns about the *high number of WBAs* required to produce reliable judgements [24,29,30,33,34]. Such concerns were compounded by the large time costs associated with conducting individual WBAs. No information could be found regarding any cost-benefit analyses or other economic analyses of WBAs.

Summary

There appears to be a dearth of literature examining the economic properties of WBAs; this could be an important avenue for future research. Regarding the broader concept of the feasibility of WBAs, the *time requirements* of WBAs were seen as prohibitive, particularly when attempting to organise them within already *busy workloads*.

Psychometric Properties

When coding psychometric properties of WBAs, we examined both the purpose of WBAs and the more traditional psychometric properties (e.g. validity, generalisability and reliability). Regarding purpose, most WBAs were identified as formative or summative, with a few references to WBAs forming part of a programmatic assessment portfolio. However, few articles described repercussions of performance on a WBA. Of those that did, usually failure to complete set WBAs resulted in remediation [20,28,32].

A common theme that arose regarding the purpose of WBAs concerned the tension between formative and summative purposes of WBAs [4,7,14,17-19,21,26-28,30-32]. Trainees and assessors alike view *formative and summative as mutually exclusive categories*, emphasising that using WBAs for summative purposes undermines any formative benefits that might arise. In particular, assessors felt

> "*...a conflict of interest between creating a learning experience for the trainee and the need to document competence...*" [19].

This conflict between their role as an *assessor and a teacher* frustrated them [4,17,32]. Some also observed that using WBAs for summative purposes *undermines the authenticity* of the WBA [4,17,18], because trainees 'game' the system for favourable outcomes by choosing assessors and cases [4,18,20]. Similarly, authenticity may be undermined by trainees' *stress of being assessed affecting their performance* [4,18,28].

The tension between formative and summative purposes is also seen as adversely affecting the utility of WBAs. Some stakeholders reported that *trainees engage less with the feedback* from summative WBAs, since the focus is on performance rather than learning [18]. Similarly, assessors report that the *purpose of a WBA affects their decision-making;* assessors are inclined to be more lenient in feedback and decisions for summative WBAs because of the great weight assigned to their decision [28,30]. This is especially the case if assessors do not feel confident in justifying their decisions [28,30]. To overcome these issues, some have suggested that the intended purpose (formative or summative) of each WBA be clearly communicated to stakeholders to ensure appropriate use [17,18].

Stakeholders' views on purpose related to the tension between tailored and standardised assessment. Some stakeholders favour standardised assessment, as this means frameworks can be used, ensuring comparability of assessments by reducing subjectivity. This aligns with the perspectives of 'objectivity' and 'accountability' identified by de Jonge [14]. Conversely, others believe assessment should be tailored to the needs of the trainee to maximise its utility, aligning with the perspectives of 'adaptivity' and 'agency' identified by de Jonge [14].

The literature reviewed that explored the more traditional psychometric properties of WBAs can be grouped into four categories.

The first details the measures taken to establish the validity, generalisability and reliability of WBAs. Most studies based their investigations on generalisability theory [13,23,24,29,33], with one using standard reliability theories [22]. Yet, some authors have argued that social constructivism should be the preferred epistemological framework to approach the generalisability of WBAs given the paramount importance of the context [14,15]. Similarly, whilst some papers used either Kane's or Messick's validity frameworks [5,23,29], some have argued that an interactional framework should be used to account for how different aspects of the WBA (e.g. trainee, assessor, tool, environment, case etc.) affect the outcome [14,17]. Multiple techniques to validate WBAs were described in the

reviewed literature, including correlations with other measures; factor analyses; inter-item comparisons; piloting and simulating; and mapping assessments to outcomes. Additionally, Archer [13] argues that the validity of multi-source feedback (MSF) (and hence other WBAs) cannot be assumed, but must be subjected to regular quality assurance measures. Generally, WBAs have been found to have low reliability, but research has found that when scales are redefined to use entrustment-based scales, reliability increases [29,33].

The second category of research concerns identifying factors which frustrate the psychometric properties of WBAs. The majority of the literature has focussed on such factors related to users. Assessor factors, such as the role of the assessor (e.g. doctor vs nurse) have been found to affect the reliability of results [34]. Similarly, more senior assessors tend to be more stringent [13] and set exam-like conditions, which can affect the validity of the encounter [19]. Much research has also found reliability is affected by assessors' psychology, including their biases, stereotypes, cognitive load and level of assessment and clinical expertise [13-17,28]. Related to this, the purpose of a WBA (i.e. formative or summative) affects assessors' decision-making, as they are wary of the significance and implications of their decisions [18]. Similarly, trainees have been found to 'game' WBAs to maximise their scores, also undermining the validity of the assessment [4,5,7,13,19,25]. More broadly, users' attitudes towards; available time for; and training in WBAs affects the way in which WBAs are used (e.g. rushing to complete them retrospectively), which in turn compromises the validity of WBAs [4,14,18-20]. Similarly, the validity of a given WBA can be affected by the relationship between the assessor and trainee. From one perspective, friendship can introduce leniency, undermining the validity of the assessment [5,13,18,19,28]. Conversely, some argue that this relationship permits longitudinal assessment, thereby improving accuracy of assessment [14].

Additional frustrating factors have been identified. One important factor is how the context can affect the psychometric properties of WBAs, with issues such as case difficulty and specialty affecting the scores produced [13,34]. A problem for psychometrics with respect to WBAs is that

> *"…cases are unpredictable and cannot be scheduled or repeated…"*[29].

To compensate for this, some argue for the importance of deliberately sampling a diversity of cases for collation [17,24]. Problems with tools have also been identified, for example poor sensitivity [5]. Additionally, the wording of scale items (e.g. norm-based vs entrustment) has been found to affect reliability [23,24,29,33], while scores for individual items cannot be compared [13,34]. Furthermore, systemic issues around WBAs have been described with respect to a lack of benchmarking data, making it difficult to make comparisons and high stakes decisions [22,26].

The third category concerns determining the optimal number of WBAs and how frequently they should be used. Some stakeholders argue that in order to establish high psychometric reliability in WBA results, many WBAs must be completed. One study identified that trainees believe between ten and twenty WBAs is sufficient to develop an appropriate overview of their competence [28]. Similarly, research examining WBAs suggests that approximately ten WBAs produces appropriate levels of generalisability [24,26,28]. However, this varies dramatically based on the wording of scales, with entrustment-based scales tending to require fewer assessments [29,33]. When considering reliability and generalisability of WBAs, it is also important to acknowledge the argument some make that the more WBAs that assessors and trainees are required to complete, the greater the likelihood users will develop assessment fatigue, which may undermine the validity of the WBA in the first place [24]. Additionally, some argue that there should not be a 'golden number' of WBAs required for all registrars, but that the number should be sensitive to trainees' performance. They argue this on the basis that fewer WBAs are required to establish reliable estimates of competence for trainees who are performing well [5,13].

The final domain of literature examining psychometric properties of WBAs concerned their association with improved skills or educational outcomes. Evidence exists to support the discriminatory capacity for a variety of WBAs regarding trainees' competence, including the mini-CEX [22,38], OCAT [23], MSF [21], TAB [38], CbD [38] and ESR [38]. Furthermore, one review determined that WBAs demonstrate changes at the Kirkpatrick Levels of 2 and 3 i.e.

> *"...a modification of skills and attitudes or behavioural or willingness of learners to apply new knowledge and skills..."* [7].

Users also perceive that the provision of feedback that occurs through WBAs produces positive behavioural changes [21,26], particularly for MSF [7]. Yet, much evidence has also contradicted these findings. Several studies have found WBAs to identify either no or very few trainees performing below competence [22,23,29]. Similarly, WBAs have been found to have low predictive validity for identifying trainees in difficulty [21,38]. This has led some to argue that WBAs have no benefits for use as a summative tool [22,23,29]. Finally, systematic reviews have concluded that WBAs do not "...affect a demonstrable change in practice..." p. 138, [21]. Yet, part of these conflicting findings may relate to a variety of limitations that have been identified in the postgraduate medical education WBA literature. Barrett and colleagues [30] have identified a number of such limitations, including an absence of theoretical frameworks guiding research; limited study designs (e.g. lack of pre-post analyses and triangulation); and not critically appraising the influence of potential researcher bias. Perhaps most significantly, they also identify that evidence of learning is measured through performance changes, a metric which does not capture affirmation of good practice. Furthermore, researchers have identified multiple gaps in the literature, including:

- Assessments of (especially behavioural) outcomes arising from WBAs [7,17,30];
- Examination of the effects of repeated use of WBAs [30];
- Investigations into the use of WBAs for trainees in remediation [30];
- Researching differences between tools [30]; and
- Defining what constitutes an 'underperforming trainee' [30].

Summary

Regarding the purpose of WBAs, the dominant issue that emerged concerned the tension between formative and summative purposes of such assessments. In particular, it was expressed that using WBAs as summative assessments *undermines their advantages and utility*. Whilst multiple frameworks have been used to assess the validity, generalisability and reliability of WBAs, the general consensus is that WBAs have low reliability unless entrustment-based scales are used. Within this context though, such low reliability is likely explained by the various factors which frustrate these psychometric properties, including users (e.g. assessors' roles or seniority, or user's attitudes); the purpose of a WBA; and the relationship between the assessor and trainee. Research generally concludes that ten WBAs are required for a sufficiently generalisable result, although factors such as assessor fatigue and trainee competence will affect this. There is mixed evidence regarding whether WBAs improve trainee performance, although part of this may be due to limitations with the literature (e.g. limited study designs).

Perceived Acceptability

Another critical aspect of the properties of WBAs is understanding how they are perceived by stakeholders. The literature suggests that rather sizeable proportions of trainees and assessors alike (approximately 40%) believe that WBAs are of low value and may even detract from the quality of training [18]. Interestingly, research suggests that older trainees [31] and international

medical graduates (IMGs); [5,31] are more likely to view WBAs favourably. Nevertheless, a variety of themes regarding stakeholders' views of the positive and negative attributes of WBAs emerged in the present literature review. Often such viewpoints were affected by stakeholders' position regarding the tailored vs standardisation debate.

Positive attributes

The most commonly cited positive attribute of WBAs related to the feedback provided. Trainees in particular appreciated how WBAs were an opportunity for them to *receive feedback* in the first place, something they felt was lacking across the remainder of their training [4,7,18,21,27,28]. Trainees also held positive views of WBAs when the feedback was in *narrative form* that provided guidance regarding *personal development*; *identified areas of improvement*; and provided a *holistic overview* of their progress [7,20,28,32,35]. Stakeholders also reported that WBAs permitting *feedback from multiple people* were useful for providing such a holistic overview [21]. Related to this, stakeholders perceived WBAs as a '*fair*' form of assessment [5,20,22]. Assessors specifically attributed this fairness to the more holistic nature of WBAs compared to other forms of assessment [22].

Stakeholders were favourable towards WBAs that used clear definitions and frameworks. Regarding this theme, assessors focussed on how these definitions and frameworks *reduced the cognitive load* and *subjectivity* in their judgements [16,17], while making it *easier to provide targeted feedback* [28]. Trainees' views aligned with assessors', focussing on how definitions and frameworks *reduce the subjectivity* of WBAs [4,5,16] and help *produce actionable feedback* [7].

Even more basically, a fundamental strength of WBAs according to stakeholders is that they involve the assessment of performance, especially direct observation-type assessments [22,24]. This in turn leads to *high face validity* [22]. According to stakeholders, this attribute was perceived to facilitate a number of valuable qualities of assessment. By assessing in an authentic context, stakeholders believe WBAs *stimulate learning* and permit the *translation of learning into action* [17,19,20,28]. Stakeholders who viewed the ultimate purpose of assessment as *ensuring high standards of care* emphasised how the direct assessment of performance makes WBAs an essential form of assessment [14,31,35]. Similarly, some stakeholders argued that WBAs allow the early identification of trainees in difficulty, thereby 1) reducing the risk of patients receiving low quality care, and 2) ensuring remediation resources are allocated in an efficient and timely fashion [13,22,32].

An additional benefit of WBAs from both assessor and trainee perspectives is their capacity to facilitate mapping trainees' progression over time. In this sense, WBAs are viewed as powerful tools for directing further efforts for development [20,25,31,32]. Yet, some stakeholders suggested this benefit was only hypothetical [21]. Related to this, WBAs were identified as providing a unique opportunity to capitalise on the potential utility of the relationship between the assessor and trainee for long-term monitoring of progress and goal-setting [14,17,28]. This can also stimulate further learning opportunities.

Negative attributes

The overarching issue with WBAs identified by stakeholders (largely trainees) in the literature concerned the misuse of WBAs. Practices such as a lack of direct observation; retrospective form completion; and insufficient feedback were perceived to undermine the value of WBAs [4,18,19,21,31]. Regarding the causes of this misuse, a plethora of themes were identified. A common theme was user disengagement with WBAs. From trainees' perspectives, *assessor disengagement* is consistently identified as a major limitation with use of WBAs, particularly regarding the quality of feedback provided [4,18,19,21,31]. This disengagement makes WBAs *cumbersome* to use, as trainees report having difficulty finding assessors willing to complete WBAs [5,25,31]. Conversely, very few assessors identify assessor disengagement as an issue [18]. Instead, assessors generally report

trainee disengagement with WBAs as undermining the utility of the assessment process, raising issues regarding trainees not taking responsibility for initiating WBAs or not recognising the learning opportunities WBAs afford them [18,21].

Stakeholders have identified that this disengagement is at least related to, if not caused by, a number of other issues. First, there is a sense of low assessment literacy amongst users. Trainees attribute a major cause of improper use of WBAs to assessors lacking the knowledge of how to use WBAs [4,18,21]. Consistent with this, some assessors have indicated they lack confidence in performing WBAs [28]. Yet, trainees have also reported receiving little or no training in understanding the purpose of WBAs and how to maximise the opportunities they afford [18,19]. Disengagement can also be traced to concerns regarding the definitions, frameworks and tools being used in WBAs. Assessors report frameworks can be cumbersome to use because of their extensive lists and use of vague and/or contentious definitions [16,17]. Trainees take issue with the importance assigned to trivial items; low reliability of tools; and the use of reductionist frameworks [5,31,32].

Assessors and trainees alike identify a major disadvantage of WBAs as their high costs, particularly those relating to time [18,20,21]. Related to this, stakeholders (especially trainees) find difficulty integrating WBAs into their existing workloads, often due to the large time demands of WBAs [7,18,21,22]. Trainees also note that WBAs can interfere with their work, for example by generating a staged environment in a consult or reducing ward efficiency [17,27].

Stakeholders expressed concern at the psychometric properties of WBAs. Specifically, some stakeholders were *sceptical of the evidence base* supporting WBAs, especially their level of reliability. The core concern seems to be a low capacity for identifying poorly performing trainees [5,19,28]. Additionally, stakeholders (especially trainees) believe WBAs have poor content validity insofar as they focus on assessing clerical skills whilst overlooking other important competencies such as teamwork [4,5,17-19,22].

Trainees also identified problems arising from WBAs being used for summative purposes. Trainees report how summative WBAs lose their formative value [18,19]. There are many reports of trainees *gaming* WBAs by choosing cases and/or assessors to attain a higher score [18,19]. Trainees also report summative WBAs as being stressful events [4,17,18,27,28,31]. Both of these factors undermine the fundamental advantage of WBAs by *compromising the authenticity* of the encounter being assessed [19,20,32]. Another issue related to using WBAs for summative purposes is that, even when WBAs are used formatively, trainees still suspect a summative element, thereby compromising the quality of even formative WBAs [7].

Trainees have raised issues regarding the unfairness of WBAs as well. A common problem is the variability that can arise from the *relationship* between the assessor and trainee, whether that be *leniency* or *stringency* [5,18,19,28,32]. Unfairness is also thought to stem from *problematic implementation* of WBAs, particularly where there is little or no training in the use of assessments and their technology platforms [5,18,21,31,33].

Trainees identified several other negative attributes of WBAs. Some trainees felt WBAs detract from the quality of training because of their low discriminative capacity [19] and by interfering with opportunities for self-learning [17,31]. Similarly, some argued that WBAs discourage excellence by focussing on minimum competence [19]. At the extreme, some trainees felt that they could be penalised under a WBA system if they began their training at a point of competence and could not demonstrate progress [5]. Finally, some trainees found feedback to be problematic, particularly where it is too context-dependent and not easily generalisable [4,28], or where it had been rushed and so of low quality [19].

Summary

Stakeholders have identified a plethora of positive and negative attributes of WBAs contributing to their acceptability amongst users. The most commonly identified benefit of WBAs is the feedback that trainees receive. WBAs are also seen to provide a holistic assessment and can be enhanced when using clear definitions and frameworks. Additionally, WBAs are appreciated for providing a means for an authentic assessment of a trainee's performance. Yet, WBAs are reportedly widely misused, with multiple negative attributes of WBAs and their use associated with this. Much of this stems from user disengagement, arising from low assessment literacy; contentious definitions, frameworks and tools; onerous time demands; and use of WBAs for summative purposes.

Question 2: What is known about the interaction between WBA tools; the registrar and assessor; and the context in terms of outcomes, meaning-making and forming valid judgements?

To approach this question, we considered the WBA from three points in its lifecycle: factors affecting how the encounter of a WBA occurs; factors affecting the decision-making of the assessor and the outputs of the WBA (i.e. feedback and scores); and factors affecting the utility of the outcomes from a WBA (i.e. learning). Additionally, we considered secondary outcomes arising from performing WBAs.

What factors affect how WBAs occur?

We first considered factors which contribute to the way in which the actual assessment (and the lead up to it) occur. We posed two specific questions and then more broadly considered this part of the WBA lifecycle.

- The first specific question concerned determining what influences which WBA is chosen in the first place. The literature suggests that these decisions are informed by two factors. The first factor concerns the feasibility of performing the WBA in the context, for example the feasibility of using video recording apparatus [27]. The second factor is stakeholders' (in particular assessors') preferences for certain WBAs [25,27].
- The second specific question considered factors determining who acts as the assessor for a WBA. The most common factor highlighted here was trainees selecting their assessor, oftentimes selecting assessors who will favourably assess them [18,19,32]. Related to this, trainees report that difficulties in the relationship between the assessor and trainee affects selection of assessors [19,21]. Selection of assessors included whether prospective assessors had received training in use of WBAs, and platform issues (e.g. not having passwords to access online platforms) [31].

Outside of these specific questions, we also examined broader factors which contributed to the conduct of WBAs. Most of these factors concerned users of WBAs. Several attributes of the assessor were identified in this regard. An important factor was assessor's attitudes towards the assessment. For example, Bindal, Wall, Goodyear [31] write that 27% of staff reported

> *"...refusing to do the assessments because they did not like doing them."*

Similarly, other studies suggest that an assessor's attitudes towards WBAs affects their engagement with the assessment and how thoroughly assessments are performed [14,27]. These attitudes relate to assessors' perspectives on the purpose of the assessment. Castanelli, Jowsey, Chen, Weller [32] write that the perception of a WBA as a tool for learning provides an impetus for overcoming barriers and engaging with the assessment. Conversely, if the purpose of WBAs is

unclear to assessors, they are more likely to complete them as administrative tick-box exercises [32]. Finally, an element of self-reflexivity was raised in the literature, with the suggestion that assessors who are aware of their own practices may serve as better models for trainees and be more observant of trainees' performance [37].

Similarly, multiple trainee factors were identified as contributing to the conduct of WBAs. The most common issue raised here is trainees feeling stressed by the experience. This commonly manifested through *nervousness* during the encounter and the generation of a staged environment which trainees felt undermined the authenticity of the assessment [18,27,28]. Levels of stress were identified as being at least partly determined by *trainees' comfort with being observed* [37]. A further trainee attribute that stakeholders commented on was the trainee's training level; assessors appear to find junior trainees more engaged in the process and receptive to feedback [28]. Finally, Patel, Agius, Wilkinson, Patel, Baker [38] noted that trainees' level of competence may also affect how they approach WBAs, with trainees of lower competence choosing assessors less likely to give them negative feedback.

Factors outside of users' control were also identified as exerting influence on the conduct of WBAs. One such factor is the definitions, frameworks and scales of a tool. These attributes were seen to provide a structure to guide assessors' observations and, in turn, decision-making [7,16,17,23,26,28,30,32,35,39]. In addition, training organisation requirements imposed regarding WBAs can also be powerful. Requirements regarding a *minimum number or frequency* of WBAs to be completed may stimulate inactive trainers and trainees to complete WBAs in the first place [27]. Whether a training organisation stipulates a WBA as having a formative or summative *purpose* influences how trainees approach and engage with the WBA [4,18]. The resources provided to complete the WBA (e.g. online platforms) and providing training in their use also affect the way WBAs occur (e.g. the capacity to capture multimedia data) [39].

Finally, interactions between users and external factors also determine how WBAs occur. Such interactions can affect the authenticity and representativeness of the encounter, such as the *assessor observing the trainee* and, through this, generating *stress* in the trainee [17-19,27,28,37], or the *trainee's choice of cases or supervisors* so as to minimise negative feedback [18,20,32]. Similarly, the relationship between the trainee and assessor has been identified as potentially facilitating or impeding WBAs [17,21]. For example, the working relationship may affect whether trainees approach supervisors to act as assessors. A further dimension to the issue of the relationship concerns *role clarity*, for example trainees being unclear whether they can ask assessors for assistance in a WBA [28]. Stakeholders have also reported the extent to which they plan WBAs (for example, allocating time to perform the WBA) influences the quality of the assessment experience [21,27,28,31,32,37]. Nevertheless, opportunistic WBAs are still seen as potentially providing unique learning opportunities [4]. Similarly, the time demands of WBAs were identified as potentially resulting in rushed or retrospective completion [18,20]. Finally, the context in which WBAs occur has been identified as important, with the content of unique encounters producing quite different scenarios for assessing [27].

Summary

It seems that what determines which WBA is chosen is largely driven by the feasibility of performing a WBA in a given context and users' preferences for certain WBAs. Related to this is the matter of who performs the assessment. This appears to be largely determined by trainee choice, which can be affected by the relationship between the assessor and trainee. In terms of the actual conduct that occurs within the context of a WBA, the dominant themes concerned assessor (e.g. attitudes towards the assessment) and trainee factors (e.g. stress; training level). Beyond this, factors such as tool attributes and training organisation requirements, as well as interactions

between these factors (e.g. relationship between trainee and assessor) can determine how a WBA occurs.

What factors affect the decision-making process and outputs?

This aspect of the WBA lifecycle is concerned with the cognitive processes occurring as assessors are receiving and processing information, and making decisions about trainees' performance (i.e. outputs). Naturally, most research in this area has focussed on assessor factors. The dominant factor here is the assessor's long-term psychological make-up. One aspect of this is the assessor's level of *clinical expertise*; as argued by Gingerich, Kogan, Yeates, Govaerts, Holmboe [15]

> *"...assessors must know and be able to accurately assess a trainee for the presence or absence of skills that define high quality care."*

Thus, the assessor's capacity for observation and decision-making will be considerably influenced by their own level of clinical expertise [15,17,37]. A potential manifestation of this concerns the *frame of reference* an assessor uses when making a decision. Some assessors may use their own level of clinical expertise as their frame of reference, meaning assessors with poor clinical expertise may deem trainees with equally poor clinical expertise as competent [15,16]. Conversely, some assessors may make comparisons with what they believe is the acceptable level for a trainee of a given level, or with their observations of trainees of a similar level [16,17]. Additionally, assessors' *stereotypes* can affect their recall of observed behaviour and judgements about trainees [15]. Related to the assessor's long-term psychological make-up is their confidence and competence in using the assessment. This discomfort can result in 'failure to fail': assessors not flagging trainees who should be flagged [29]. Levels of confidence and competence regarding WBAs are inherently associated with levels of assessment literacy. Training in the use of assessments can assist developing an understanding of expected competency levels for trainees at different stages of training; what is being assessed by WBAs; and how to justify unsatisfactory scores [17,28,33]. A further factor is the role of the assessor, which has several components. The assessor's *profession* (e.g. nurse vs clinician) can affect the feedback provided [38], while the *seniority* of clinician assessors has consistently been associated with stringency of decisions and quality of feedback [13,19]. Furthermore, there is the more pedagogical question the assessor faces as whether they identify as an *assessor or a teacher*. Relating to the aforementioned tension between formative and summative purposes, assessors report a tension between assessing a trainee for an assessment decision and advancing the trainee's learning [32]. Finally, the assessor's perceptions of the tools can have an effect on decision-making and outputs. Specifically, assessors' views on whether they believe *tools are sufficiently comprehensive* to make assessment decisions can influence their use of these tools [23,28]. Similarly, the outputs and decision-making are influenced by *assessor's engagement* with WBAs [19,20].

Additional factors have been identified in relation to the tool. The most common items raised in this respect concern the *definitions, frameworks and scales* used by the tool. These structures have been identified as influencing how assessors think about a consult; informs their assessment decisions; and helps them to construct feedback [4,5,16,23,26,28,29,34]. For example, tools which identified trainees as 'needing supervision' rather than 'performing at a lower level' reportedly make it easier for supervisors to raise concerns with trainee performance [32,33]. A notable advantage of using such definitions, frameworks and scales that stakeholders raised was the increased standardisation of analysis [4,5,16,28]. Other attributes of the tool, including the *length of the tool* and the *number of sources of information* a tool collates (especially for MSF) affect the outputs from an assessment [30]. Finally, the way in which WBAs are implemented (e.g. technology platforms) can affect the modality and timeliness of recording notes and preparing feedback [18].

Several contextual factors were also raised. An issue some raised was the amount of time that assessors have to develop and deliver feedback as affecting the quality of feedback [20]. Further, some assessors identified that, even if they wanted to provide feedback about skills development, they lack the time to help trainees in this skills development [36]. The other contextual factor raised is that the environment in which feedback is provided alters how feedback is delivered to trainees [20]. For example, in a busy environment feedback may be impaired by interruptions or an inability for trainees and assessors to speak candidly [20]. Additionally, the encounter itself will have a major influence on the assessor's judgement and decision-making [20,34]. Although this is an inherent strength of WBAs, stakeholders view this as problematic when *trainees purposely selected cases to maximise favourable assessment outcomes* [20].

The final area of factors affecting decision-making and outputs are interactions between the stakeholders, context and the tool. The most commonly raised factor is the interaction between the assessor and trainee. First, whether a *relationship* between assessor and trainee exists in the first place is thought to have an effect on feedback, particularly when trainees choose assessors for favourable outcomes [19,20]. The literature also indicates that *trainee seniority* affects outputs, with assessors becoming more critical in their assessments of more senior trainees [7]. Yet, the most important subthemes to emerge here concerned the *trainee's reaction to feedback* and the *implications of feedback*. Quite commonly, assessors reported adjusting their assessment decisions or avoiding feedback so as to prevent negative trainee reactions to feedback or to avoid trainees being flagged [7,14-16,28-30,38]. Such alterations were reported to be most common in situations where face-to-face feedback was provided [28]. Related to this, assessors reported that the results of previous assessments for a given trainee could affect their comfort with providing negative feedback [30]. Specifically, if a trainee had previously received negative feedback, assessors reported that this would make it easier for them to provide negative feedback. User's views on the authenticity of assessments also affect how users report using these tools, whether this be as intended or hastily and retrospectively completing forms [21]. Finally, the degree to which assessor's cognitive structures align with the framework of a tool influences the feedback provided and decisions made by assessors [15]. For example, some find reductionist checklists interfere with assessor's expert judgements, whereas use of entrustment-based frameworks may assist assessors [15,34].

Summary

When considering how a decision is made and feedback/scores delivered to a trainee, the assessor's long-term psychological make-up (e.g. *level of clinical expertise, stereotypes*); assessment literacy; perceived role (e.g. *assessor vs teacher*); and perceptions of the tool appear to have a major impact. Similarly, attributes of the tool (e.g. *definitions and frameworks used*; *length*) can also influence the assessor's decision-making. Additionally, contextual factors, such as time-constraints and the environment in which feedback is provided can pressure assessors' decisions and reporting, whilst the encounter that is the subject of the WBA itself will naturally play a major role in outputs. Finally, the interactions between these factors and others can be considerable influencers, for example the nature of the interaction between the assessor and trainee can affect assessor's decisions, how feedback is framed and what is communicated.

What affects the utility of the WBA outcomes?

The third question was concerned with factors influencing the learning and engagement with feedback (i.e. the outcome of a WBA). Most of these discussions centred on feedback. There is much description of what attributes make for useful feedback. A critical attribute of feedback reported consistently in the literature is that *feedback must be narrative* [4,5,7,13-17,20,30]. Stakeholders contrast narrative feedback with numerical feedback, reporting that narrative feedback is easier to interpret and more meaningful [7,14-17]. Another vital attribute of feedback is it being *targeted* to

specific attributes of the directly-observed assessed encounter [4,7,14,17,26,27,30,37,38]. This often is accomplished through providing *observation-based feedback* rather than making inferences, for example focussing on a trainee's posture rather than inferring their level of interest in a consult [7,17,26]. Additionally, feedback should *focus on the processes* involved rather than the specific context of an encounter, as this facilitates translation of learnings to new situations [26]. Stakeholders also identified that, to help trainees to make the best use of their feedback, feedback should be *actionable* and include *goals* [4,17,20,26,27,30,36]. Furthermore, these goals should be *tailored* to trainees' learning goals and needs [14,17,26,27]. Stakeholders also emphasised the importance of providing feedback in a *timely* manner, so trainees perceive this feedback as relevant and feel empowered to action it [17,26,27]. Trainees, especially senior trainees, said they appreciated feedback when it is *holistic* so they can develop an overall picture of their progress [5,20]. To be useful, stakeholders believe it must be *honest*, since dishonest feedback prevents trainees from identifying areas of improvement, which undermines a central tenet of the purpose of WBAs [14,26,37]. The final important attribute of feedback concerned its framing. Trainees value feedback when it is benchmarked *relative* to assessor's expectations for a trainee of their training level [28]. Similarly, stakeholders indicated that feedback should be framed using *standards or norms*, particularly those based on frameworks, as these provide clear criteria for trainees to work towards [16,37].

Additional considerations regarding feedback have been raised in the literature. Stakeholders recommend to avoid focusing on *personal attributes* in feedback, as this often leads to negative reactions from trainees and stifles learning [27]. Feedback should also avoid *inconsistencies*, for example contradictions between details in tick-boxes and free-text responses [38]. There is debate regarding whether feedback should be primarily *positive or negative*, or contain elements of both. Some have argued that positive feedback is important to reinforce positive attributes and stimulate further learning [17,37]. They contrast this with findings that negative feedback is typically met with critical appraisal of the validity and source of the feedback rather than learning from the feedback [27]. Given this evidence, others have argued that feedback should not be strongly negative, but framed in a constructive way and tempered by also including positive feedback [7]. Research also suggests that the *context* in which feedback is provided affects how trainees use this feedback. For example, feedback provided in front of patients is not preferred by some [37]. Similarly, there is debate about the optimal *delivery format* of feedback, with some preferring written feedback to minimise the effect of trainees' reactions to the feedback [27], while others arguing that both verbal and written feedback are important [4]. Evidence also suggests that the *source* of feedback will influence how it is used. When receiving feedback from fellow clinicians, trainees tend to assess the credibility of the feedback in terms of the perceived expertise of the assessor [17]. Conversely, research suggests that feedback from patients is generally accepted and used by trainees [7]. Finally, the *topic* the feedback addresses will also influence its use, although there is debate as to whether professionalism or clinical feedback is more often used [7].

In addition to the plethora of factors concerning the feedback itself, trainee factors also play a major role in the use of feedback. The most common trainee factor identified in this respect is the *self-reflective skills* of the trainee [4,14,17,26,27], which may also be developed through the WBA process [15,21]. This is related to how *receptive* the trainee is to receiving feedback [14,17,26,27], as well as their *motivation* towards learning being either performance- or learning-oriented [14,27]. A further factor concerns how much *training* trainees have received in how to use the assessment they receive [18,31]. This training may also impact trainees' *interpretation* of feedback as indicating a success or failure [26,27], and their *confidence in actioning feedback* [27].

Apart from the feedback itself and trainee factors, assessor factors and interactions between the assessor and other factors are also identified as affecting the utility of feedback from WBAs. The primary assessor factor is the *assessor's level of engagement* with the assessment, which can

influence how and whether trainees engage with feedback [5,18,20]. This effect likely arises from the influence of assessor engagement on how WBAs are conducted and the development of feedback. The relationship between the assessor and trainee again plays an important role. In the context of outcomes, this relationship has the potential to foster greater engagement with feedback. This is especially so when the relationship is long-term, as this facilitates *mapping a trainee's progression* and makes trainees feel the *assessor is invested in their learning* [5,7,14,17,19,20,26-28,37]. This relationship can also be useful in planning WBAs in accordance with learning goals to maximise their utility [17,26,27]. Finally, planning may also help to overcome issues around time constraints in provision of feedback, as limited time hinders trainees' capacity to reflect on feedback [19,20].

Summary

The foremost determinant of how stakeholders utilise WBAs lies in the qualities of the feedback. For example, users consistently report the importance of feedback being *narrative*, *observation-based*, *holistic*, *goal-focussed* and *timely*. While feedback should avoid *inconsistencies* and *personal attributes*, there is debate as to whether it should be primarily *positive or negative;* delivered in certain *contexts*; and around the *optimal modality* for delivery. Beyond the feedback itself, the trainee is also paramount, encompassing factors such as *self-reflective skills*; *receptivity*; *motivation*; and *assessment training*. Similarly, the assessor's level of engagement with the assessment and the relationship between the assessor and trainee appear important in how trainees utilise the feedback they receive.

What are the secondary outcomes of WBAs?

In addition to the primary outcomes of WBAs in relation to feedback, the literature identifies numerous secondary outcomes outside the assessment context that can arise from WBAs. These primarily involve the various effects WBAs can have on the relationship between the assessor and the trainee. Some stakeholders argue that WBAs have a *positive effect* on this relationship by providing an opportunity for assessors to become better acquainted with their trainees' performance [28,32]. Conversely, some have found WBAs can have a *negative* effect on this relationship. These manifested through the relationship becoming strained by trainees pursuing assessors for documentation, or trainees viewing supervisors as holding a more summative assessor role than a formative supervisory role [18,28]. WBAs can also impact on the broader practice/hospital workforce. Some identified that MSF was a useful tool for motivating other staff members [7]. However, most believed WBA had negative effects by decreasing efficiency, and interfering with trainees' consults and autonomy [17,22,27,31]. Some trainees, especially senior trainees, felt that a positive side-effect of WBAs is that they prompt supervisors to increase their supervision [28]. A further benefit reported by some assessors was benefits to their own practices, learning new strategies from trainees or improving their teaching [16,17,32].

Summary

Notably, a number of secondary outcomes appear to arise from conducting WBAs. While some argue that WBAs provide an opportunity for the relationship between the assessor and trainee to develop, others suggest that WBAs may tarnish this relationship by factors such as assessors adopting a more summative than formative role. Similarly, whether WBAs motivate the broader practice or hospital workforce or decreases efficiency is debated. Some positive effects that users have noted are that trainees believe WBAs encourage assessors to increase their supervision, while assessors report occasionally learning new practices from their trainees.

Question 3: What is known about the required assessment literacy of registrars, supervisors and training organisations to optimise the utility of WBAs?

The majority of literature coded under this question concerned training of assessors. This likely reflects findings that suggest a large minority (between 25 and 40%) of assessors have not

received any training in the use of WBAs [4,18]. Training is hypothesised as being important for three primary reasons. First, training is thought to improve the quality of outputs and outcomes of WBAs [5,7,15,20,26,28,29]. Specifically, trainees believe that assessors need more training to improve the *quality of feedback*, especially regarding the vocabulary of feedback [7,20,21,26,28]. Second, training is thought to reduce threats to the validity of WBAs by improving standardisation and reducing inconsistencies [4,13-15,26]. Third, assessor engagement may be improved by explaining both *why* WBAs are important and *how* to conduct them [4,17-19,24,26].

To accomplish these aims, various types of training have been devised. These primarily target assessors' *knowledge in how to complete WBAs* [16-18,23], but others also provide *meta-cognitive* training in conducting observations or improving clinical skills [15]. Research examining the outcomes of such training has noted various benefits. The most common benefit of training that assessors report is feeling they are providing higher-quality feedback to trainees, largely due to refining their *feedback vocabulary* [14,16,17,26]. Assessors also report that training gives them greater insight into how they make judgements based on their observation by using formalised frameworks [16,26]. This also influences how they observe trainees by being more targeted in their observations, thereby reducing the *cognitive load* of WBAs [16,17]. Finally, assessors also report that training improves their confidence in using WBAs, largely from their *richer vocabulary* for providing more targeted and constructive feedback [16,17,26].

Nevertheless, a number of issues in relation to training assessors have also been identified. First, training may not address assessors' preferences for using certain WBAs if the WBA does not align with the assessor's fundamental beliefs about assessments [14]. This is most notable in instances where assessors prefer holistic assessments over checklists. Similarly, assessors may not agree with definitions used in tools, for example definitions of what constitutes 'competence' [16]. More broadly, assessors have indicated reluctance to change their practices regarding WBAs despite training if there is a lack of a culture change at their institution, as this may reflect negatively on them if they are perceived by trainees as being relatively stringent [14]. Finally, there is generally a low rate of voluntary uptake of assessment training programs amongst assessors [18].

Although scant, there is some literature discussing the training of trainees in the use of WBAs. One UK-based study reported that a large proportion of sampled trainees had been trained in using WBAs [31]. The primary anticipated benefit of training trainees in WBAs is to help them understand the purpose of WBAs with the hope of improving trainee engagement [18,19,24,26]. Training is also thought to help trainees understand how to use WBAs, particularly *how to interpret their feedback* [14,26,28]. Although very little literature has examined outcomes of trainee WBA training, there is evidence suggesting that training increases engagement with WBAs [31].

In addition to training assessors and trainees in the use of WBAs, the literature also demonstrates that factors at the levels of the training organisation and broader medical system also affect the utility of WBAs. At the training organisation level, two dominant themes emerged. The first concerns the culture of the training organisation. For WBAs to be effective, the training organisation must promote a *culture of assessment **for** learning* rather than assessment ***of*** learning [17,37]. Where the former will promote learning-oriented motivations, the latter will promote performance-oriented motivations which may undermine the validity and utility of WBAs. Additionally, training organisations must ensure that there is *stakeholder engagement* in the development and maintenance of WBAs. Many assessors reported that being included in the designing and implementation of WBAs improved their attitudes towards, and engagement with, WBAs [16,17]. Stakeholders also reported the importance of having *uniformity* across the training organisation for WBAs, particularly with respect to the use of scales and definitions. This was viewed as promoting standardisation, consistency and fairness in the assessment process for

trainees [14,16]. Within this uniformity though, some argued that the training organisation also needed to promote a degree of *flexibility* so that assessors and trainees could tailor WBAs to the trainee's needs [14,27]. To maximise the utility of WBAs, several training organisational administrative requirements were thought to be critical. These included mandating a minimum number of assessments from assessors of a certain seniority, and providing pairings between assessors and trainees to facilitate the establishment of longitudinal relationships [4,13,14,17,19,27].

At the broader level of the medical system, stakeholders identified that two fundamental issues plagued WBAs. First, stakeholders identified that the broader medical system does not promote a culture of feedback [21,27,37]. This lack of feedback outside of the training context, combined with a blaming culture, does not normalise feedback, potentially detracting from trainees' seeking and acceptance of feedback within WBAs [37]. Additionally, some stakeholders perceive a disconnection between the ideals of high-quality care and the reality of the medical system. This can serve to undermine the effectiveness of feedback trainees receive [16].

Summary

While the literature suggests that assessor training in use of WBAs is low it is thought to facilitate improvements to the quality of WBAs (e.g. *higher quality feedback*; *reduced threats to validity*; *higher assessor engagement*). Accordingly, multiple types of training have been designed, each targeting different facets of assessor assessment literacy. While this training has been associated with some self-reported improvements (e.g. *higher-quality feedback*, *greater confidence*), the effects of such training appear to be inhibited by factors such as assessor preferences and organisational culture. Less literature has considered the concept of trainee assessment training, although this suggests it may increase trainee engagement with WBAs. Perhaps even more important than these factors though, is the role of training organisation and medical system factors, such as culture (e.g. *a pro-feedback culture that encourages learning*; and *engaging stakeholders*). Change in factors at this broader level is likely to be far more effective than assessment literacy interventions alone.

Discussion

Key findings

The present review provides a thematic overview of important issues relating to WBAs in the context of postgraduate medical education. Overall, stakeholders view the purpose of postgraduate medical training as ensuring that future doctors provide high-quality care for their patients. The key mechanism by which WBAs facilitate this is by providing feedback based on the assessment of a trainee's authentic performance in their workplace.

We then explored the economic and psychometric properties, and perceived acceptability of WBAs. From an economic perspective, the key factor involved in determining the feasibility of WBAs is their time requirements. Psychometric research identifies that reasonable reliabilities can be established for WBAs, particularly when scales use entrustment-based definitions. Additionally, a number of practical factors, in particular whether a WBA is used for summative or formative purposes, has been identified as considerably affecting the validity of a WBA. There is conflicting evidence regarding whether the use of WBAs is associated with changes in behaviour, although there are a number of limitations in the research exploring this association. The most favourably viewed attributes of WBAs by stakeholders are that they involve the provision of feedback and the assessment of performance, and allow trainees' progression to be mapped over time. WBAs which use clear definitions and frameworks are favoured over those with more arbitrary standards. Many negative perceptions of WBAs arise from the misuse of WBAs. This misuse may be traced back to

user disengagement; low assessment literacy; incongruences in definitions and frameworks; difficulty implementing WBAs into already busy workloads; and WBAs being used for summative purposes.

When examining factors determining the outcomes; meaning-making and forming of valid judgements in WBAs from all three points of the WBA lifecycle, common themes emerged. These entail user attributes (such as preferences for WBAs; user engagement; and training in use of WBAs) and the definitions, frameworks and scales used by tools. However, the majority of critical factors arise from interactions between these factors, such as the authenticity of an encounter being affected by user and contextual factors. Similarly, the relationship between the assessor and trainee can greatly affect the conduct, output and outcomes of a WBA. Interestingly, several secondary outcomes of WBAs were also noted, including effects for the assessor-trainee relationship; broader practice or hospital workforce; and assessor self-improvement.

The final domain we explored was the required assessment literacy for stakeholders to optimise the utility of WBAs. Most discussions concerned assessment training for assessors, with a variety of potential benefits of such training highlighted. While assessors seem to report some of these benefits are realised through training, such as an improvement in the quality of feedback, training of assessors is insufficient to optimise the utility of WBAs. Training in WBAs for trainees is thought to also be important, although this was discussed less in the examined literature. Nevertheless some evidence does suggest trainee assessment literacy training improves engagement. However, broader factors at the level of the training organisation and the medical system also influence the utility of WBAs.

Across a number of questions, several key themes were consistently pertinent. These dominant themes included the relationship between the assessor and trainee; the tension between formative and summative uses of WBAs; and low user engagement. That these themes emerged repeatedly in response to different questions suggests these are critical factors that must be emphasised when designing, implementing and evaluating WBAs.

An even more dominant theme that emerged in response to all questions was the important role of the interactions between the users, context and tool in determining the effectiveness and acceptability of WBAs. This is a fundamental deviation from traditional validity and reliability theories. Both these groups of theories tend to see validity almost uniquely as characteristics of the instrument and include the user only from the perspective of consequential validity, i.e. the impact the assessment has on the learning of stakeholders [40]. What has become clear from this review is that the multilateral interaction between the users, the instrument and their environment plays a dominant role in the validity and reliability of WBA.

Gaps

In the course of conducting this literature review, several gaps were identified. The first gap concerned economic analyses. In reviewing the literature for such information, we were required to broaden our definition of ‘economic’ factors from financial analyses to resources. Based on observed time demands of mini-CEXs, one reviewed study estimated financial costs of WBAs [22], however this did not encapsulate all costs in relation to implementation and additional maintenance, nor did it consider returns on investment. Information regarding the financial costs and returns of WBA implementation and maintenance would provide another useful dimension by which decision-makers could assess the value of WBAs.

A further gap in the literature was the repercussions of performance on WBAs. Although much of the included literature discussed the outcomes of WBAs, there was no discussion of how WBAs

were used in a summative assessment context. Examining the remediation procedures for trainees identified as struggling using WBAs would be another useful perspective. Given WBAs are an assessment, and a key function of an assessment is to highlight the strengths and weaknesses of a trainee, understanding the extent to which WBAs guide remediation processes is important to understand. For example, suppose external clinical teaching visits (ECTVs) are used in a training organisation to identify trainees needing remediation. If it were found that ECTVs do not provide sufficient information to identify the particular difficulties the trainee was facing to guide the remediation program, the ECTV would need modifying and/or assessors would need further training.

With respect to Scoping Question 3 about assessment literacy, the majority of the included literature considered assessment literacy and training for assessors. Yet, as has been demonstrated throughout this review, no single element of the collection of users, tools and contexts can be isolated with respect to WBAs. That literature has considered the role of trainees, training organisations and even the broader medical system is encouraging, yet there was very little of this identified in the present review. Hence, there also seems to be a gap in the literature considering training regarding WBAs for stakeholders and organisations other than assessors themselves.

Limitations

In using the approach of a hermeneutic review, the rigour of qualitative methodologies has been harnessed to deeply engage with this literature. However there are a number of limitations to this review. First, it is not a systematic review and therefore not exhaustive. Given the timeframe for this review, a systematic review was not possible. However, by using qualitative methodologies, the concept of saturation was also invoked. We therefore make no claims that this review is exhaustive or comprehensive, but that it provides a thematic overview of the dominant themes raised in this body of literature. Additionally, the scope of this review was confined to Australia and countries sharing a comparable medical training system to Australia. This decision was made so as to ensure the findings were relevant to the Australian context. The findings of this review, while potentially relevant, are not necessarily applicable to contexts outside of such training systems.

Conclusion

In reviewing the literature examining WBAs in postgraduate medical education, it is evident that the acceptability and effectiveness of WBAs are not only determined by the instrument itself, but by a complex set of interactions between the trainees, assessors, context and tool. Taking into consideration these interactions, we have developed a series of recommendations to improve the utility of WBAs:

1. Training organisations must clearly articulate whether a WBA has a formative (assessment *for* learning) or summative (assessment *of* learning) purpose to assessors and trainees to maximise its utility.
2. Different assessors should conduct formative WBAs, than those who conduct summative WBAs. Formative WBAs should preferably be assessed by the same group of assessors.
3. When used for summative purposes, WBAs should be collated programmatically.
4. WBAs should use entrustment-based scales, definitions and frameworks.
5. The competencies assessed by each WBA should be listed, and a description of how these competencies are assessed, provided to assessors.

6. The culture of the training organisation must be collaborative and be supportive of assessment for learning to ensure at least a minimally valid assessment process. More punitive and competitive cultures are likely to lead to gaming and cheating and thus invalid assessment.
7. Consistency and minimum standards must be enforced throughout a training organisation, but flexibility permitted within these standards to enable tailoring of assessments to trainees' individual needs.
8. Measures must be taken to ensure trainees' performance in WBAs is as authentic as possible. This may be achieved via:
    a. Training organisations allocating time for supervisors/assessors and trainees to complete WBAs to maximise their utility. This includes providing time for planning/organising WBAs.
    b. Trainees not being permitted to choose assessors or cases for WBAs whenever possible.
9. WBAs should be conducted by multiple assessors and in a variety of contexts.
10. Assessors, trainees and training organisations must be trained in the use of WBAs, including the tools; their purpose; platforms; and meta-cognitive awareness.
11. Benchmarking data should be available for WBA tools.
12. Trainees must receive high-quality feedback from WBAs about their performance. High-quality feedback is characterised by narrative feedback which is timely, holistic, actionable, honest, standard/norm-based, tailored and targeted/observation-based. Furthermore, it identifies strengths and weaknesses. It is not context-dependent, nor based on personal attributes.

References

1. Norcini JJ, Blank LL, Arnold GK, Kimball HRJAoim. The mini-CEX (clinical evaluation exercise): a preliminary investigation. 1995;123(10):795-799.
2. Miller GEJAm. The assessment of clinical skills/competence/performance. 1990;65(9):S63-67.
3. Norcini JJ. Work based assessment. BMJ : British Medical Journal. 2003;326(7392):753.
4. Shalhoub J, Marshall DC, Ippolito K. Perspectives on procedure-based assessments: a thematic analysis of semistructured interviews with 10 UK surgical trainees. BMJ Open. 2017;7(3).
5. Viney R, Rich A, Needleman S, Griffin A, Woolf K. The validity of the Annual Review of Competence Progression: a qualitative interview study of the perceptions of junior doctors and their trainers. Journal of the Royal Society of Medicine. 2017;110(3):110-117.
6. Labbé M, Young M, Nguyen LHP. Toolbox of assessment tools of technical skills in otolaryngology–head and neck surgery: A systematic review. The Laryngoscope. 2018;128(7):1571-1575.
7. Saedon H, Salleh S, Balakrishnan A, Imray CHE, Saedon M. The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. BMC Medical Education. 2012;12:25.
8. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: A systematic review and meta-analysis. Academic Medicine. 2015;90(2):246-256.
9. Durning SJ, Artino Jr AR, Pangaro LN, van der Vleuten C, Schuwirth LJAM. Perspective: redefining context in the clinical encounter: implications for research and training in medical education. 2010;85(5):894-901.
10. Kane M. Validation. In: Brennan R, ed. Educational Measurement. Vol 17. Westport: ACE/Praeger; 2006:18-64.
11. Prochaska JO, Wright JA, Velicer WFJAP. Evaluating theories of health behavior change: A hierarchy of criteria applied to the transtheoretical model. 2008;57(4):561-588.
12. Boell SK, Cecez-Kecmanovic DJC. A hermeneutic approach for conducting literature reviews and literature searches. 2014;34:12.
13. Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in a national programme. Archives of Disease in Childhood. 2010;95(5):330-335.
14. de Jonge LPJWM, Timmerman AA, Govaerts MJB, et al. Stakeholder Perspectives on Workplace-Based Performance Assessment: Towards a Better Understanding of Assessor Behaviour. Advances in Health Sciences Education. 2017;22(5):1213-1243.
15. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: Assessor cognition from three research perspectives. Medical Education. 2014;48(11):1055-1068.
16. Kogan JR, Conforti LN, Bernabeo E, Iobst W, Holmboe E. How faculty members experience workplace-based assessment rater training: A qualitative study. Medical Education. 2015;49(7):692-708.
17. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: The do's, don'ts and don't knows of direct observation of clinical skills in medical education. Perspectives on Medical Education. 2017;6(5):286-305.
18. Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. Advances in Health Sciences Education. 2016;21(2):455-473.
19. McKavanagh P, Smyth A, Carragher A. Hospital consultants and workplace based assessments: how foundation doctors view these educational interactions? Postgraduate Medical Journal. 2012;88(1037):119-124.

20. Mehta F, Brown J, Shaw NJ. Do trainees value feedback in case-based discussion assessments? Medical Teacher. 2013;35(5):e1166-1172.
21. Barrett A, Galvin R, Scherpbier AJJA, Teunissen PW, O'Shaughnessy A, Horgan M. Is the learning value of workplace-based assessment being realised? A qualitative study of trainer and trainee perceptions and experiences. Postgraduate Medical Journal. 2017;93(1097):138.
22. Brazil V, Ratcliffe L, Zhang J, Davin L. Mini-CEX as a workplace-based assessment tool for interns in an emergency department-Does cost outweigh value? Medical Teacher. 2012;34(12):1017-1023.
23. Halman S, Rekman J, Wood T, Baird A, Gofton W, Dudek N. Avoid reinventing the wheel: implementation of the Ottawa Clinic Assessment Tool (OCAT) in Internal Medicine. BMC Medical Education. 2018;18.
24. Homer M, Setna Z, Jha V, Higham J, Roberts T, Boursicot K. Estimating and comparing the reliability of a suite of workplace-based assessments: An obstetrics and gynaecology setting. Medical Teacher. 2013;35(8):684-691.
25. McConnell M, Sherbino J, Chan TM. Mind the Gap: The Prospects of Missing Data. Journal of Graduate Medical Education. 2016;8(5):708-712.
26. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Medical Teacher. 2007;29(9):855-871.
27. Pelgrim EAM, Kramer AWM, Mokkink HGA, van der Vleuten CPM. The process of feedback in workplace-based assessment: organisation, delivery, continuity. Medical Education. 2012;46(6):604-612.
28. Weller JM, Jones A, Merry AF, Jolly B, Saunders D. Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: implications for implementation. British Journal of Anaesthesia. 2009;103(4):524-530.
29. Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. British Journal of Anaesthesia. 2014;112(6):1083-1091.
30. Barrett A, Galvin R, Steinert Y, et al. A BEME (Best Evidence in Medical Education) review of the use of workplace-based assessment in identifying and remediating underperformance among postgraduate medical trainees: BEME Guide No. 43. Medical teacher. 2016;38(12):1188.
31. Bindal T, Wall D, Goodyear HM. Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise? Medical Teacher. 2011;33(11):919.
32. Castanelli DJ, Jowsey T, Chen Y, Weller JM. Perceptions of purpose, value, and process of the mini-Clinical Evaluation Exercise in anesthesia training. Canadian Journal of Anaesthesia. 2016;63(12):1345-1356.
33. Crossley J, Johnson G, Booth J, Wade WJMe. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. 2011;45(6):560-569.
34. Crossley J, Jolly BJMe. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. 2012;46(1):28-37.
35. Fokkema JP, Scheele F, Westerman M, et al. Perceived effects of innovations in postgraduate medical education: AQ study focusing on workplace-based assessment. 2014;89(9):1259-1266.
36. Holmboe ES, Yepes M, Williams F, Huot SJJJogim. Feedback and the mini clinical evaluation exercise. 2004;19(5p2):558-561.
37. Jenkins LS, Klaus von P. Family medicine training in Africa: Views of clinical trainers and trainees. African Journal of Primary Health Care & Family Medicine. 2018;10(1).
38. Patel M, Agius S, Wilkinson J, Patel L, Baker P. Value of supervised learning events in predicting doctors in difficulty. Medical Education. 2016;50(7):746-756.

39. Sandars J, Dearnley C. Twelve tips for the use of mobile technologies for work based assessment. Medical Teacher. 2009;31(1):18-21.
40. Messick SJEr. The interplay of evidence and consequences in the validation of performance assessments. 1994;23(2):13-23.

Appendix A: List of questions

Scoping question 1

Psychometric properties

- How is validity defined by the paper? E.g. Cronbach's construct validity, Messick's framework, Kane's framework etc.
- What was the described purpose of the WBA? E.g. assessment *for* learning v *of* learning (i.e. formative v summative)
- What measures have been taken to ensure the validity of the described WBA? E.g. blue-printing; staff development/supervisor training; infrastructural measures (e.g. parallel consulting rooms); organisational structure (e.g. super numeracy); validity research
- What approach to generalisability was used? E.g. standard reliability measures, generalisability theory, item response theory/Rasch modelling, saturation, transparency/note-taking/audit trails, organisational reliability

Economic properties

- What were the investments associated with the WBA? E.g. development and running costs etc.
- What were the returns on investment of the WBA? E.g. prevention of study delay, attrition; improved quality of registrars, probability of passing RACGP examinations etc.

Effectiveness & Acceptability

- What, if any, differences are there in skills outcomes between types of WBA tools?
- What, if any, differences are there in education and skills outcomes between those who do and do not participate in WBAs?
- What was the acceptability of WBAs for registrars?
- What was the acceptability of WBAs for supervisors?
- What was the acceptability of WBAs for MEs?
- Specifically:
    - Views and perceptions on assessment in general
    - views and perceptions on the need for training in general
    - specific views with relation to how WBA ensures safe and independent practitioners
    - how WBA stimulates and drives learning
    - how WBA stimulates and drives a quality orientated organisation
    - how time-consuming WBA is and whether that is 'value for money'
    - what specific aspects about the implementation of WBA have made it overly cumbersome or, vice versa, easier

Scoping question 2

- What factors (e.g. personality) of registrars, supervisors and/or MEs are associated with WBA performance?
- What factors (e.g. personality) of registrars, supervisors and/or MEs are associated with perceptions of WBAs?
- Are certain WBA tools more preferred by certain users than others? If so, what factors impact these preferences?

Scoping question 3

No additional questions were initially generated for Question 3

Appendix B: Search strategy

Database(s): ***Ovid MEDLINE(R)*** and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R) 1946 to November 01, 2018
Search Strategy:

#	Searches	Results
1	("work* base*" adj3 (assess* or evaluat*)).tw,kf.	451
2	workplace based assessment.tw,kf.	212
3	1 or 2	451
4	education, medical, graduate/ or "internship and residency"/	63234
5	(postgraduate medical education or post-graduate medical education or registrar* or intern* or GPs or general praction* or physician*).tw,kf.	1170899
6	4 or 5	1215889
7	3 and 6	222
8	exp Australia/ or (Australia* or Queensland* or New south wales or victoria* or tasmania* or northern territory*).tw,kw.	187631
9	New Zealand/ or New Zealand*.tw,kw.	65873
10	exp Canada/ or (canad* or alberta* or british columbia* or manitoba* or new brunswick* or newfoundland* or labrador* or northwest territories or nova scotia* or nunavut* or ontario* or prince edward island* or quebec* or saskatchewan* or yukon territory*).tw,kw.	215039
11	exp Great Britain/ or (Britain* or British or United Kingdom* or England* or English or Scotland* or Scottish* or Wales or Welsh or Channel Islands or Northern Ireland* or Irish).tw,kw.	575090
12	"scandinavian and nordic countries"/ or denmark/ or greenland/ or finland/ or iceland/ or norway/ or svalbard/ or sweden/ or (Scandinavia* or nordic or denmark* or danish or greenland* or finland* or finnish or iceland* or norway* or norwegian* or sweden* or swedish or Netherland*).tw,kw.	307901
13	or/8-12	1277498
14	7 and 13	102

Database(s): Ovid Emcare 1995 to 2018 week 43
Search Strategy:

#	Searches	Results
1	("work* base*" adj3 (assess* or evaluat*)).tw,kw.	271
2	workplace based assessment.tw,kw.	122
3	1 or 2	271
4	education, medical, graduate/ or "internship and residency"/	66688
5	(postgraduate medical education or post-graduate medical education or registrar* or intern* or GPs or general praction* or physician*).tw,kw.	441880
6	4 or 5	494115
7	3 and 6	115
8	exp Australia/ or (Australia* or Queensland* or New south wales or victoria* or tasmania* or northern territory*).tw,kw.	74671
9	New Zealand/ or New Zealand*.tw,kw.	20883
10	exp Canada/ or (canad* or alberta* or british columbia* or manitoba* or new brunswick* or newfoundland* or labrador* or northwest territories or nova scotia* or nunavut* or ontario* or prince edward island* or quebec* or saskatchewan* or yukon territory*).tw,kw.	77786
11	exp Great Britain/ or (Britain* or British or United Kingdom* or England* or English or Scotland* or Scottish* or Wales or Welsh or Channel Islands or Northern Ireland* or Irish).tw,kw.	135853
12	"scandinavian and nordic countries"/ or denmark/ or greenland/ or finland/ or iceland/ or norway/ or svalbard/ or sweden/ or (Scandinavia* or nordic or denmark* or danish or greenland* or finland* or finnish or iceland* or norway* or norwegian* or sweden* or swedish or Netherland*).tw,kw.	89966
13	or/8-12	367585
14	7 and 13	28

Database(s): ***PsycINFO*** 1806 to October Week 5 2018
Search Strategy:

#	Searches	Results
1	("work* base*" adj3 (assess* or evaluat*)).ti,ab.	159
2	workplace based assessment.ti,ab.	54
3	1 or 2	159
4	(postgraduate medical education or post-graduate medical education or registrar* or intern* or GPs or general praction* or physician*).ti,ab.	321144
5	exp Australia/ or (Australia* or Queensland* or New south wales or victoria* or tasmania* or northern territory*).ti,ab.	46782
6	New Zealand/ or New Zealand*.ti,ab.	10669
7	exp Canada/ or (canad* or alberta* or british columbia* or manitoba* or new brunswick* or newfoundland* or labrador* or northwest territories or nova scotia* or nunavut* or ontario* or prince edward island* or quebec* or saskatchewan* or yukon territory*).ti,ab.	46991

8	exp Great Britain/ or (Britain* or British or United Kingdom* or England* or English or Scotland* or Scottish* or Wales or Welsh or Channel Islands or Northern Ireland* or Irish).ti,ab.	192196
9	"scandinavian and nordic countries"/ or denmark/ or greenland/ or finland/ or iceland/ or norway/ or svalbard/ or sweden/ or (Scandinavia* or nordic or denmark* or danish or greenland* or finland* or finnish or iceland* or norway* or norwegian* or sweden* or swedish or Netherland*).ti,ab.	60870
10	or/5-9	333681
11	3 and 4	31
12	3 and 4 and 10	8

ProQuest:

Set#	Searched for	Results
S2	(noft(("work* based" NEAR/2 (assess* OR evaluat*))) AND noft(GPs OR "general practition*" OR physician* OR registrar* OR "post-graduate medical education" OR "medical education")) AND stype.exact("Scholarly Journals") AND PEER(yes)	97
S1	noft(("work* based" NEAR/2 (assess* or evaluat*))) AND noft(GPs or "general practition*" or physician* or registrar* or "post-graduate medical education" or "medical education")	133

Stream 2
Australian General Practice Training organisation work-place based assessment audit

Prof Lambert Schuwirth (Prideaux Centre, Flinders University)
Assoc Prof Jill Benson AM (GPEx)
Dr Emily Kirkpatrick (GPEx)
Dr Jon Wignes (GPEx)
Mr Shaun Prentice (GPEx)

41

GPEx In collaboration with

GP training · Flinders University · general practice training tasmania · MURRAY CITY COUNTRY COAST GP TRAINING · NORTHERN TERRITORY General Practice Education · RVTS REMOTE VOCATIONAL TRAINING SCHEME LTD · WAGPET · GP SYNERGY · GPTQ

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

GPEx would also like to acknowledge:

- The Working Group for this stream including: Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Assoc Prof Jill Benson AM (GPEx), Dr Emily Kirkpatrick (GPEx), Dr Jon Wignes (ModMed Institute), Mr Shaun Prentice (GPEx), Dr Taryn Elliott (GPEx) and Ms Michelle Pitot (GPEx).
- Collaborating Australian general practice training organisations: Eastern Victoria GP Training (EV), General Practice Training Tasmania (GPTT), Murray City Country Coast GP Training (MCCC), Northern Territory General Practice Education (NTGPE), Western Australian General Practice Education and Training (WAGPET), Remote Vocational Training Scheme (RVTS), GP Synergy and General Practice Training Queensland (GPTQ) for their hard work in providing data for this report.
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.

GPEx In collaboration with

GP training · Flinders University · general practice training tasmania · MURRAY CITY COUNTRY COAST GP TRAINING · NORTHERN TERRITORY General Practice Education · RVTS REMOTE VOCATIONAL TRAINING SCHEME LTD · WAGPET · GP SYNERGY · GPTQ

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Introduction

Regionalised Training

Australian general practice training is currently regionalised to states and/or regions within states, except for the Remote Vocational Training Scheme (RVTS) which covers rural and remote areas across Australia. This regionalised approach takes into account the significant unique regional variations in Australia including, but not limited to, differing geography, communities and health systems.

To cater for this vast regional variation Regional Training Organisations (RTOs) have been funded to deliver the Australian General Practice Training (AGPT) Program within specified geographical regions. There are currently ten RTOs (including the RVTS) that deliver training across the whole of Australia. For the purpose of this report when RTO is mentioned it will include RVTS.

To ensure a consistent quality in general practice training, each RTO is accredited by both the Royal Australian College of General Practitioners (RACGP) and the Australian College of Rural and Remote Medicine (ACRRM) to deliver training according to the Standards set by the College/s. Training must also be delivered according to the AGPT program policies. However, within this framework the training models used by each RTO vary considerably in order to cater for particular features of their region (e.g. urban/rural/remote), their allocated budget and the available staffing. This has resulted in a variety of different models used across Australian RTOs to deliver workplace-based assessment (WBA).

There is a vast array of WBA models being used across Australian RTOs with similarities and differences, and currently no shared understanding about what are the most appropriate WBAs for general practice registrar training. As we progress towards a transition in 2020 to College-managed delivery of the Australian General Practice Training Program, it is important to understand the current models of WBA used in Australian general practice, in order to plan for the future.

Overarching Goals

In order to achieve a deeper understanding of the current models used across Australia, the current project undertook an audit of RTO WBAs to achieve two overarching goals:

- To audit the WBA tools currently in use in RTOs.
- To map these against the core skills of general practice and to develop an overall map of WBAs.

Defining the focus for the audit

For the purpose of this project our definition of workplace-based assessment is:
Workplace-based assessments are the tools and the processes used in the collection of data about the trainee's performance in the workplace, and the judgement of their competence by an assessor, for a range of summative and/or formative purposes.

Grounding the audit content within WBA literature

In order to inform this project stream, the literature described in the previous chapter has provided a lens for prioritising data capture, analysis and presentation of the results of the audit. The literature defines a number of key qualities of workplace-based assessment (WBA) assessment which were taken into account when designing the audit content. These include:

- The validity and reliability of WBA are dependent on the users.

- The validity and reliability of WBA are dependent on the context within which they are being used.
- WBAs are often used for many different purposes.
- Support, monitoring, compliance and regulation are essential.
- A number of documented barriers exist to successful WBA administration.

The literature suggests that if all these prerequisites are not sufficiently met, workplace-based assessment is extremely unlikely to produce valid and reliable outcomes. These key qualities are described in more detail below.

Workplace based assessment is user-dependent
There is a fundamental difference between workplace-based assessments (WBAs) and the more traditional standardised assessments, such as written or structured oral assessments. In the more traditional standardised assessment measurement characteristics, validity and reliability, are built-in aspects of the assessment instrument. For example, it is defensible to claim that a multiple-choice test has a certain validity and reliability and that these aspects are relatively independent of the user. In WBAs on the other hand, the measurement properties of the assessments are highly dependent on the users. Although, the rubric and the form are still important, the main purpose of WBAs is to support the judgement made by the assessor. In most WBA situations the assessor, therefore, must observe as well as form a judgement at the same time.

Therefore, WBAs can only be used by users (assessors, learners, decision-makers) who have sufficient understanding of the assessment process and its goals, in addition to the requisite medical content expertise, thus there is a need for a high standard of user training.

Therefore, WBA users were explored as a key part of the WBA audit to determine who is responsible for using WBAs within the RTO contexts.

Workplace based assessment is context-dependent
It is not only the assessor who plays an important role in WBAs, organisational culture, infrastructure and facilitation are also essential. Where organisational culture, infrastructure and facilitation are suboptimal, WBA typically does not work well. Reliable, appropriate and user-friendly infrastructure and clear and logical facilitation are crucial in ensuring optimal feasibility and thus optimal 'return on investment' for the time and resources spent on conducting WBAs. Hence, the success of WBAs is far more dependent on the specific context compared to more traditional assessments.

The organisational culture and infrastructure have to facilitate an honest and open dialogue between the learner and their supervisor, assessor or mentor in order to be able to authentically evaluate strengths and weaknesses as well as progress.

Therefore, contextual issues, such as RTO organisational factors and the many different medical educator (ME) roles, were explored to better understand the differing contexts in which the WBAs are being used across Australian RTOs, and whether this impacts on the way they are used.

Workplace based assessment can be used for different purposes
Traditional assessment is mostly aimed at capturing the level of competence of the learner at a certain point in time, with the intent of determining whether sufficient learning has taken place,

and whether the learner is ready to move into the next phase of their training or career. WBA, on the other hand, is often used for many different purposes. The most obvious of these are: formative feedback, summative decision-making, assessment *of* learning and assessment *for* learning.

Therefore, the different purposes for which WBAs are being used has been included in this audit.

Support, monitoring and regulation is essential for workplace based assessment
The literature recommends that WBAs are delivered in an easily accessible, practical, outcome-based, standardised and integrated manner for both the assessor and trainee, where compliance with the WBA is monitored and regulated.

Therefore, in the audit we explored the different modes of delivery of WBAs and WBA feedback, and the systems that RTOs have in place to monitor and regulate compliance with WBA completion and actions after feedback.

A number of documented barriers exist to successful WBA administration
The literature documents a range of barriers to successful administration of WBAs.

Therefore, this audit explored the existing barriers within RTO models of WBA implementation.

Key audit questions
Based on the key themes described above, the Steering Group agreed on the following questions to inform both data collection and analysis:

- What WBAs are used and by whom?
- For what purposes are WBAs used?
- How do RTOs support, monitor and regulate the use of WBAs?
- How do WBAs map against the RACGP's domains of general practice?
- What barriers exist to WBA success in the AGPT program?
- How do individual, communication and contextual factors influence the use and outcomes of WBAs?
- What are RTO perceptions in regard to the link between WBA performance and remediation and exam success?

Method

Participants
All RTOs in Australia were invited to participate in the audit by the GPEx Lead Research ME (n=10). Nine RTOs agreed to participate. This resulted in data collection across nine RTOs covering every state within Australia.

Tools and process for data collection
The audit content was designed with reference to the literature and through consultation with the Project Steering Group, which included representatives from each of the nine RTOs involved in this project. Two tools were designed to gather data for this audit:

- WBA Audit tool

- ME role questionnaire

Workplace based assessment audit tool
This tool was designed using an interactive series of linked sheets within an Excel template. Each Excel template contained questions pertaining to:

- Global questions establishing the RTO's organisational context (eg. number of registrars and supervisors, ME role)
- Information about each type of WBA used by the RTO

The individual questions aligned to the overall questions identified above.

Questions required both quantitative and qualitative responses to ensure rich data could be captured.

This Excel template was sent to each RTO in December 2018. RTOs determined the most appropriate person to fill in this template on behalf of the organisation. All nine participating RTOs completed the Excel template and returned this to GPEx for review and analysis.

Returned audits were reviewed by the Working Group and RTOs were followed up to provide additional information if clarifications were required.

Some of the RTOs did not consider learning plans to be WBAs and did not include this information in their original data. All RTOs thought that the current method of requiring and monitoring learning plans was not beneficial for registrars' learning, which was why they did not all include them. In order to further explore what they actually did, we asked RTOs to go back and complete that section of the Excel template if they had not previously done so.

Medical educator role questionnaire
Based on the variety of responses received through the initial WBA audit tool, an additional ME role questionnaire was also designed and administered to determine the similarities and differences between the ME role across RTOs as not all RTOs had fully explained this in the Excel spreadsheets.

This questionnaire was designed by the project Working Group based on an analysis of the data received from the WBA audit and the experience within the Working Group. RTOs were asked to rate from 1-3 the roles of the MEs as administrator, as assessor, as teacher, as training advisor, in accreditation and development, in programmatic assessment, in recruitment to the RTO, in a specialised portfolio or in regards to additional training:

1=Does not apply to MEs

2=Sometimes applies to MEs or applies to some MEs

3=Applies to all MEs

The questionnaire was sent to each of the nine RTOs in March 2019. RTOs determined the most appropriate person to fill in this template on behalf of the organisation.

All nine RTOs completed the additional questionnaire and returned the data to GPEx for inclusion in the data analysis.

Workplace based assessment audit database design
All data captured through the WBA audit and ME role questionnaire was reviewed by the Working Group and coded and themed in order to determine the specific data fields to be incorporated into the relational database management system, Microsoft Access.

By using a database to store the information we were able to access and explore the data more flexibly. A database structure allows for different ways to sort, combine or group data. For example, the question how many different WBAs do the RTOs use and the question how many RTOs use a particular WBA requires a different grouping of the data.

A network diagram was constructed and shared with the Steering Group to show the data fields that would be captured by the relational database management system and how these data fields and data tables would relate with each other and interact in order to answer the agreed questions. Feedback from the Steering Group was considered in the final design of the relational database management system data-set.

At the lowest level of aggregation of the raw data there are 43 individual WBA 'events'. Each event means the use of a particular WBA by a particular RTO. It does therefore not represent each individual administration of that WBA as most are administered at multiple points in time[2] . For the initial 12 tables this data has had to be aggregated slightly because some responses to the audit questions could not be separated out at the level of an individual event. At that level of aggregation 39 events are identified. In this report tables should be interpreted to be based on the total of 39 events, unless specified otherwise.

Data analysis and reporting
The questions defined by the Steering Group were used to guide the data analysis phase.

The senior and junior researchers from the Working Group met for an initial data analysis day whereby all audits were reviewed to identity gaps within the audits, to remove additional non-requested data and to highlight key words prior to input into the database.

The Working Group met for two data interrogation days whereby data within the relational database management system was analysed to:

- Inductively explore the data for similarities, differences, and patterns within the data.
- Answer the agreed questions from the Steering Group.

The key themes identified from the network diagram were taken into account during this process. The final report was presented to the Steering Group for discussion and sign off.

Some questions in this report explore relationships between characteristics of the RTO and issues concerning the use of WBA. The relationships were studied with a Kruskal Wallis test. This statistical test is a non-parametric test. Non-parametric tests are used when the volume of data is small – typically a number of observations smaller than 30 – or when it cannot be assumed that the data is normally distributed. [3]

[2] Cf under "Frequency of workplace-based assessments" in this report.

[3] Statistical tests, such as the Kruskal Wallis, are typically used to gauge the likelihood that the effect found in the sample is a representation of the effect that would be found if the whole population were studied. If for example a study is done to compare the effect of a drug versus a placebo in a group of patients, the researchers want to be able to infer that any difference found in the group of study participants would also be found if all possible patients would be studied. In making these inferences two types of error can happen; there might be an effect in the sample leading the

In any situation where we have had to perform multiple comparisons, we have applied a Bonferroni correction.

Results

Variation in Organisational Contexts

Regional Training Organisation size and training ratios

Nine RTOs were involved in this audit. RTOs were from all states within Australia and covered both urban and rural training contexts. Table 1 below describes the variation in size of RTOs.

Table 1. Variation in size and training ratios between Regional Training Organisations

	Average	Minimum reported	Maximum reported
Total number of Registrars	425.37	96	1319
Total number of Supervisors	507.44	90	1117
Total ***FTE*** [4]of Medical Educators	12.55	5.10	35.08
Total number of training coordinators (number)	17.72	5	96
Ratio of Registrars: Supervisors	0.79	0.48	1.18
Ratio of Medical Educators : Registrars	0.04	0.02	0.05
Ratio of training coordinators : Registrars	0.06	0.01	0.27

Regional Training Organisation rural training volume

Whilst all RTOs delivered training across both urban and rural contexts, the total number of rural registrars varied across RTOs. Rural and remote training placements will generate higher costs and require greater flexibility, as MEs, external clinical teachers, training coordinators and on-site training will usually be based in a city or large regional centre. In some rural and remote areas it is more difficult to ensure good supervision and a registrar will need more support from the training organisation in the form of increased fly-in ME or ECT visits, and/or off-site supervision. The logistics of rural and remote training placements will also need more training coordinator time and this is reflected in the increased number of training coordinators in those RTOs with higher numbers of rural registrars.

researchers to conclude that there will be an effect in the whole population, but there is actually none. This is a type-I error (comparable to a false positive lab test). The converse, researchers falsely concluding that there is no effect, is a type-II error. For these errors an arbitrary cut-off of acceptable likelihood is use, the famous 5% or $p \leq 0.05$.

In our case we need to do multiple comparisons, for instance comparing whether there is an association between the type of WBA and the issues reported with WBA. If we accept a 5% likelihood of a type-I error in each of the comparisons the likelihood of finding 'significance' purely by chance increases with the number of comparisons. With 2 comparisons it becomes 9.75%, with three it is 14.26%, etc. This is similar to healthcare. When a patient undergoes a series of lab tests – especially when they are not hypothesis driven – the likelihood of finding a false-positive result increases with the number of lab tests. A simple way to counteract this is to make the threshold for significance lower than 5% using a Bonferroni correction. This correction simply means dividing the 5% by the number of comparisons. With 10 comparisons, for example, the p-value for significance would be set to .005 (0.5%).

[4] The FTE is by approximation because one RTO provided a count of MEs rather than the total of FTEs.

What WBAs are used by whom?

Types of workplace-based assessments

In total the 38 WBAs were used by nine RTOs. Unsurprisingly, a huge diversity of WBA methods and approaches are employed across Australia.

There are, on the other hand, also similarities between individual WBAs. In order to make the data more comparable, we have classified these individual WBAs into sub-groups. These are presented in Table 2.

Table 2. Classifications of the various types of WBAs

Group	Subgroup	Description
DOV	DOV- External Visitor	Direct Observation Visit by external observer- ME or External Clinical Teacher (ECT)
	DOV-supervisor	Direct Observation Visit by supervisor
	Video	Consultation Observations recorded on video device
Learning plan		Learning plan
MSF	MSF- colleagues	Multi source feedback from colleagues
	MSF- patients	Multi source feedback from patients
PETAL		Patient Encounter Tracking and Learning tool
Term Assessment		Term-assessments, including mid and end-term

Across all RTOs, WBAs could be categorised into five types as described in Table 2 above. However, not all types of WBA were used across all RTOs, and the combination and frequency of WBA use differed across RTOs.

Table 3 shows the number of RTOs using each of the categories of WBA.

Table 3. Types of WBAs used across RTOs

WBA subgroup	*Number of RTOs using this WBA (initial RTO reporting)*	*Number of RTOs using this WBA (revised RTO reporting)*
Term Assessment	7	7
DOV- external	9	9
DOV-supervisor	3	3
PETAL	4	4
Learning plan	4	9
MSF-colleagues	6	6
MSF-patients	3	3
Video	3	3

In the initial data collection, it can be seen from Table 3 above that only four of the nine RTOs reported that they used learning plans as a WBA. Since planned learning (for instance in the form of a learning plan) is a RACGP requirement, and the feedback from focus groups and interviews in Stream 3b of this project did not correlate with this finding, the Working Group approached all RTOs who did not report using learning plans as a WBA to clarify if learning plans were used. As a result of this all RTOs reported that learning plans were actually used, but they had not included them in the audit because they were not seen as a WBA.

This is not entirely surprising as the learning plans can be seen as being peripheral to workplace-based assessment, or alternatively as a document or activity in which the outcomes of various WBAs are collected and collated. As such, we felt that it was important to include the learning plans in this report as they play an important role in the integration of the assessment processes. As will be clear from this stream of the report, engagement with planned learning is of vital importance, as the collection and collation of assessment information is only useful if it leads to an analysis or reflection and the formulation and enactment of learning activities.

Table 3 above shows that learning plans and DOVs, which are both required according to the RACGP standards, are used across all RTOs. Term assessments are also used by most RTOs. Video review and patient feedback are only used by a third of RTOs.

Diversity of workplace-based assessments

The number of different WBAs used varied between RTOs. This variation is presented in Table 4.

Table 4. List of WBAs used at each RTO.

RTO	*Number of different WBAs*	*Types of WBA*
1	5	DOV-external; MSF-colleagues; PETAL; term-assessments; learning plans
2	6	DOV-external; DOV-supervisor; MSF-patients; Video; term-assessments; learning plans
3	5	DOV-external; MSF-patients; PETAL; term-assessments; Learning plans
4	4	DOV-external; MSF-colleagues; MSF-patients; learning plans
5	6	DOV-external; DOV-supervisor; MSF-colleagues; PETAL Video; term-assessments; learning plans
6	6	DOV-external; DOV-supervisor; MSF-colleagues; Video; Term-assessments; learning plans
7	5	DOV-external; PETAL; term-assessments; Learning plans
8	3	DOV-external; MSF-colleagues; learning plans
9	4	DOV-external; Video; term-assessments; learning plans

Table 4 shows that all of the participating RTOs use DOVs (Direct Observation Visits) by an external observer and three also use a DOV by the supervisor. Learning plans are used by all RTOs. The next most frequently used WBA categories are term assessments (n = 7) and MSF-colleagues (n = 5) and/or patients (n = 3). Three RTOs use video observations and four use some form of PETAL.

Frequency of workplace-based assessments

How frequently during training do the different classes of WBA get used (ie DOVs -ranges between 1-3 times per phase).

All RTOs use the DOV by an external observer at least once per semester in both GPT1 and GPT2 and only once in GPT3. When a DOV by supervisor is done it is only once for GPT1 and in one RTO this is also done in GPT2.

There is variety in the frequency in which MSF by colleagues are being performed: ranging from only once during the whole training period (n = 2) to once every six months. (n=3) and in one event

it is only when there is an identified need. When MSF by patients are being used they are all used only once during training.

PETAL is performed once per semester (n= 4).

Term assessments are undertaken 2-3 times per term/semester for all seven events.

Although there are some slight differences between the RTOs and the WBA tools, there are some commonalities as well. Firstly, all MSF for patients happen either only in GPT1 (n=2) or at the start of GPT2 (n=1). All DOVs take place during at least GPT 1-3 in all RTOs and in one RTO also in GPT4. DOV by the supervisor takes place in each semester of GPT1 and GPT2. PETALs and term-assessments take place throughout GPT 1 – 3 and in one RTO also in GPT4.

However, although WBA instruments continue to be used in GPT3, the frequency of their use decreases to one for the whole GPT3 phase.

Who uses the workplace-based assessments?

Table 5 shows who is commonly involved in assessing the trainee for each type of WBA, although there is wide variation across RTOs in regard to how the WBAs are organised and administered. For example, the audit shows that WBAs that require simple data collection (such as the MSF and PETAL) are often managed by the registrar, the practice manager, or another external party. Where more face-to-face interaction is required and more complex synthesis of information, the ME plays a prominent role in either conducting the WBA (such as a DOV) or as a connection between the supervisor, registrar and the training organisation (such as the term assessments).

Table 5. Assessors commonly using each WBA.

WBA	*Assessor*
DOV- External Visitor	ME or ECT
DOV-supervisor	Supervisor
DOV - Video	Supervisor or ME
Learning plan	Registrar and supervisor, ME generally signs off (n=6)
MSF- Colleagues	ME, practice manager, external company, supervisor
MSF- Patients	External organisation
PETAL	RTO, supervisor, ME
Term Assessment	Supervisor and training coordinator

Medical educator and training coordinator roles

When we explore the different roles of those who use WBAs, there are two roles that are less obvious and are different across all RTOs. One is that of the training coordinators and the other is of the ME. The following section provides a summary of the common features of these roles which has emerged from data collected across all RTOs.

Training coordinators

All RTOs described the role of a training coordinator. This role has been given different titles at different RTOs; commonly known as ‘Education Officers’ or ‘programme training advisors’ or just ‘admin’. The training coordinator is a non-medical administrative role. The specific responsibilities of training coordinator vary across RTOs, but typically they have the responsibility for

coordinating internal and external training communication, co-ordination of registrar placements and training records, coordination of training activity (including WBAs) and support to MEs.

This role often works closely with the MEs to provide administrative support, reducing the ME administrative load and providing a cost efficient alternative to training delivery (for example co-ordinating collation of recognition of prior learning, ensuring appropriate requirements have been completed and documented to ensure registrar completion of training).

They are an important part of the WBA team as they perform such duties as: coordinate DOVs with the ME or ECT, the registrar and the practice; send out the paperwork for the MSF or PETAL; and assist with ensuring that video equipment is sent out for video-reviews.

In many RTOs training coordinators also track compliance with and completion of WBAs, and follow up with reminders. In some RTOs this task is assigned to an ME, who can spend valuable time checking, emailing and phoning to ensure that WBAs are completed. In those RTOs where this task is performed by the training coordinator, they will use agreed processes to follow up completion of WBAs. In the case that a registrar has been followed up, and is still unable to comply with RTO expectations, training coordinator have the ability to provide additional support or refer this registrar to an ME, committee or regional head of education.

The training coordinator often has responsibility for a group of registrars throughout their training, and/or a number of practices that they always support. Therefore training coordinators often have regular communication with practice managers, registrars and supervisors. Because these key stakeholders have an ongoing relationship with the training coordinator, they may ring a training coordinator as the first port-of-call for a variety of queries.

Because of the level of interaction training coordinators have with practice managers, registrars and supervisors, they are often involved in the RTO flagging process. The types of flags that may arise from a training coordinator relate to: difficulties in the practice, in the relationship between registrar and supervisor, personal health issues, or lack of professionalism. Hence a skilled training coordinator can be an integral part of the RTO team as a broker, as their non-medical and non-hierarchical relationships are often less threatening to registrars and practice managers. RTOs described the importance of having a clear reporting and escalation pathway for the various issues that might arise. For instance, the flagging system should informally and formally include training coordinator input, and continued non-completion of WBAs should be escalated to an ME or committee. Training coordinator reports can be part of the programmatic assessment committee for sign-off of training requirements and for monitoring remediation.

Medical educators

The ME has a myriad of roles in the RTO team. The number of MEs, hierarchy, and diversity of specific roles varies enormously across the RTOs. Some of this variation is dependent on geography; numbers of registrars; the cross-over of ME and training coordinator roles; the teaching and WBAs expected of supervisors; whether MEs train, accredit or liaise with supervisors and practices; how many small groups are run; and whether MEs are responsible for the remediation of their own registrars or there is a specific remediation ME.

MEs often have a longitudinal relationship with their assigned registrars throughout their training and may be involved with them in small groups, direct observation visits, clinical teaching visits, training advice, WBA sign-off, flagging discussions and/or remediation. Their role is more likely to be that of mentor, support and coach than teacher, as they are one step removed from the day-to-day supervision of the registrar. At direct observation visits they are in part ‘representing’ the

training organisation and can informally assess the relationship of the registrar with the practice staff and the supervisor, as well as the practice context and its educational value. They can act as a broker between the practice, supervisor, registrar and RTO if necessary and can triangulate opinions about the registrar with the supervisor.

Because of their longitudinal relationship with the registrar and because they have medical education training, MEs are well-placed to have an overall programmatic view of the registrar's progress. In addition, the training coordinators will usually escalate compliance issues with the registrar's assigned ME, and supervisors may ring the ME if they wish to verbally flag the registrar without putting anything on paper. In many RTOs, there are regular meetings of all MEs to discuss all registrars, and in others, only those registrars with any monitoring or action flags are discussed. In some RTOs flagged registrars are discussed by a different committee with input from the MEs.

The ME role questionnaire was analysed to determine similarities and differences in the ME role across RTOs. Table 6 shows each of the categories and tasks within the questionnaire with the total number of RTOs who responded that their MEs are involved in this role at least some of the time or that at least some of their MEs are involved in this role. All RTOs indicated that their MEs were involved in the majority of roles. Roles where there was a large variation across RTOs are highlighted.

Table 6. Roles of the medical educators in the different training organisations.

Category	*Task*	*Total RTOs (n=9)*
Administration	ME responsible for monitoring compliance with assessment completion	9
	ME documents training requirements for registrar portfolio	2
Assessment	ME performs direct observation visits of the registrar	9
	ME reviews assessments	9
	ME signs off on assessments	9
	ME provides feedback on individual assessments	9
	ME flags registrars for review	9
Teaching	ME leads registrars in small group discussion	9
	ME facilitates case-based discussions	9
	ME provides support and training to supervisors	8
	ME provides support and training to practice staff	4
	ME provides support and training to RTO staff	6
	ME has role in learning resource development	9
Training Advice	ME provides career counselling	9
	ME provides training advice	9
	ME acts as a registrar mentor	9
Accreditation and Development	ME involved in flagging practices	8
	ME involved in accreditation and assessing of practices	5
	ME involved in accreditation and assessing of supervisors	6
	ME involved in feedback and development of practices and supervisors	7
	ME participates in registrar support and coaching	9

Category	*Task*	*Total RTOs (n=9)*
Programmatic Assessment	ME participates in registrar monitoring	9
	ME has a picture of the registrar's overall progress based on assessment performance	9
	ME involved in programmatic assessment committee	6
	ME provides feedback on overall progress	9
Recruitment to RTO	ME participates in selection and recruitment of registrars	9
ME Training	ME is expected to, and is supported in, having additional training to that provided by RTO	9
ME role in Specialised Portfolio	ME leads remediation	3
	ME leads a specialised area within the RTO (e.g. Aboriginal Health)	8
	ME leads research and academic registrars within the RTO	4
	ME leads exam preparation	6

Table 6 shows that ME involvement is similar across all RTOs, with MEs being involved in all identified roles within the Assessment category. This pattern was similar across most roles within the Programmatic Assessment category. The one exception was 'ME involved in programmatic assessment committee' where only six RTOs reported this was part of an ME role.
The greatest level of variation in ME roles related to WBA was in whether:

- MEs lead remediation, (n=2 RTOs)
- ME documents training requirements for registrar portfolio (n=2 RTOs)

For what purposes are Workplace-based Assessments used?

In order to understand the reasons why the various WBA are used, RTOs were asked to describe the purpose for which they use each specific WBA in their training and assessment program. These were themed into three purposes:

- ***Summative*** - meaning that the outcome of this WBA is that a direct decision will be made with possible consequences for either remediation or progression is determined by the compliance with, or satisfactory completion of the WBA.
- ***Formative*** - meaning that the outcome of this WBA is only for feedback or teaching purposes, without direct ramifications for progression or remediation.
- ***Part of programmatic assessment*** - meaning that the outcome of this WBA is initially for feedback and to drive learning, but it can also be used, in conjunction with the outcome of other WBAs, to support decisions about remediation or progression.

The use of the terms 'summative' and 'formative' has mostly been replaced with 'high-stakes' and 'low-stakes' in the current literature. This allows for a spectrum of outcomes ranging through 'medium-stakes', which would incorporate WBAs used in programmatice assessment.

The table below shows that in most events and in most training organisations, WBAs are used for formative reasons, either as a sole purpose (n=12) or in combination with programmatic assessment (n=12). Only in 10 events are the WBAs used for purely summative reasons and in four events summative and formative purposes are combined. There are striking differences between

the different training organisations with respect to the espoused purposes of the WBAs. These are summarised in Table 7.

Table 7. Summary of WBA purposes by RTO.

RTO	*Number of WBAs used as formative assessments*	*Number of WBAs used as summative assessments*	*Number of WBAs used as part of programmatic assessment*
1	2	3	0
2	4	3	1
3	2	5	0
4	2	0	0
5	5	2	4
6	5	0	5
7	2	0	1
8	1	2	0
9	3	0	1
TOTAL	***27***	***15***	***12***

Table 7 shows that RTOs primarily use WBAs for formative assessment purposes. Four RTOs have reported that they do not use WBAs summatively at all. However, three of these indicate that WBAs are used as part of programmatic assessment, which might suggest that the information is used for decision-making at a later stage.

Two out of nine RTOs use the WBAs as a significant part of programmatic assessment, whereas four RTOs do not seem to use any form of programmatic assessment.

Note that the total number of purposes is higher than the total number of WBA events; this is because some WBAs are used for multiple purposes.

Table 8. Types of WBAs and Purpose.

WBA	*Total Formative*	*Total Summative*	*Total Programmatic*
DOV-external	7	2	2
DOV-supervisor	2	1	0
Learning plans	6	4	1
MSF-colleagues	4	2	2
MSF-patients	2	0	0
PETAL	2	1	2
Term assessment	6	2	3
Video	3	1	2

Table 8 above shows wide variation across RTOs in the purpose for which WBAs are used. There is no clear relationship between the type of WBA and its purpose, although learning plans and MSF-colleagues are used for summative purposes more than any other type of assessment. In using the word ‘summative’ in this context, many RTOs mean that compliance with the WBA is mandatory and do not necessarily mean that the outcome of the WBA is used as a barrier. The distinction

between summative and formative is less clear when the assessment program is more programmatically organised. In that event one could argue that eventually the information that arises from the assessment can and will be used for summative purposes. Hence it could be argued that all of the term assessments, learning plans, MSF-colleagues and video assessments eventually end up being used for summative or high-stakes purposes.

How do training organisations, monitor and regulate the use of WBAs?

Modes of delivery of WBAs

WBAs are delivered in different ways across RTOs. Table 9 summarises the mode of delivery across the different categories of WBAs.

Table 9. Mode of WBA delivery.

WBA	*Mode of administration*			*Comments*
	Paper	*Online*	*Face-to-face*	
DOV-external	0	1	9	Followed up with online report
DOV-supervisor	0	2	3	Followed up with online report
Learning plans	2	5	2	
MSF-colleagues	2	3	3	
MSF-patients	3	0	0	
PETAL	3	1	1	
Term assessments	1	5	2	Feedback from ME online
Video	1	2	2	
TOTAL	***12***	***18***	***22***	

Table 9 shows that a high proportion of WBAs are delivered using online methods, particularly the term assessments and the learning plans. In 17 events multiple modes of delivery are used, either a completion of the WBA and face-to-face discussion afterwards, or a face-to-face discussion first with a written/on-line report undertaken later.

WBA Written Feedback Structure

A further important aspect of validity of WBAs is the way in which the information from observing the registrar is summarised in a conclusion or score. Therefore, we explored how the different training organisations rate or scale the different WBAs, especially whether the 'scores' are expressed in numbers or in narrative. Table 9 shows the ways in which feedback was recorded across the different WBA categories.

DOV

All administrations of a DOV (15 across all 9 RTOs), have a narrative component. Some use narratives exclusively, but three DOVs also yield a numerical score. 10 DOVs are scored against a rating scale, either 'unsatisfactory', 'borderline', 'satisfactory' and 'excellent', or 'well below', 'below', 'at' and 'above'.

MSF

The seven administrations of multisource feedback are generally (five out of seven) rated on a three point scale typically indicating whether the performance of the registrar is below expectation, at expectation or above expectation, which is either operationalised with terms such

as 'behind', 'at' or 'ahead', or 'performing poorer than expected', 'performing at the level expected' or 'performing above expectations'. Alternatively, coloured bands, green, amber or red are being used. In one event the outcome of the multisource feedback determines whether the registrar is being flagged for follow-up or not, and in one event the results of the MSF are not scored but form the basis for a discussion with the supervisor. Although not all RTOs used the word 'flagging' in their response to the questions, we have interpreted all the rating systems, whether they be 'behind', 'below expectation' or such as an indication that the registrar is or is likely to be flagged in the course of their training.

PETAL

The PETAL, *patient encounters for teaching and learning,* is usually scored numerically with an additional narrative in two events. In all events the outcomes are being used to compare the performance of the registrar with others in the same stage of training, and in two events the outcomes are used to determine whether the registrar should be flagged for follow-up or not and whether their progress is deemed satisfactory or unsatisfactory. In one there is a self-reflection by the registrar on the results.

Term assessment

In five of the nine term assessments a combination of numerical scores were used. In two events the outcome was dichotomous, 'flagged' or 'not flagged', but in all other events a rating scale was used, mostly 'below expected level', 'working towards expected level', 'at expected level' or 'above expected level'. In one event, EPAs with entrustment scales were used in which a judgement was made where the supervisor would have to be; 'frequent direct in the room review by supervisor', 'needs on-site supervisor available in the practice', 'is safe to practise with phone access to supervisor' and 'is safe to practise unsupervised'.

The use of feedback from the WBAs

The score or narrative is important to inform the registrar where they are in terms of the development of their competence. Feedback, on the other hand can give the registrar more detailed information about their individual strengths and weaknesses and how best to improve. All types of WBAs generate outcomes that can be used to provide feedback. The next part describes how that feedback is used.

Table 10 summarises for each category of WBA how feedback is delivered and who is given access to that feedback.

Table 10. WBA Feedback mode of delivery and access

WBA	*Feedback Format*			*Who has access to the feedback*			
	Face to face	*Written*	*Online*	*Registrar*	*Supervisor*	*ME*	*Training coordinator*
DOV-external	6	1	8	9	8	6	3
DOV-supervisor	3		3	3	3	1	1
Learning plans	8	2	5	8	6	3	2
MSF-colleagues	5 (1)[5]	1	3	6	6	5	3
MSF-patients	3		3	3	3	2	0
PETAL	3	2	1	4	3	3	0

[5] One is verbal but over the phone with the supervisor

WBA	Feedback Format			Who has access to the feedback			
	Face to face	Written	Online	Registrar	Supervisor	ME	Training coordinator
Term-assessments	5	1	5	7	4	5	3
Video	4	1	3	4	4	3	2

Table 10 shows that the majority of WBA feedback is given face-to-face as well as in an online format. Term assessment feedback, most commonly used as a formative assessment, is not always given face-to-face with three of the eight RTOs who use term assessments only providing written feedback, some retrospectively after a registrar has left a placement. This may have an impact on the registrar's ability to efficiently apply this feedback.

In most events the registrar and supervisor have access to the feedback. In a number of events the ME does not have access to all feedback. The training coordinator does not often have access to WBA feedback.

RTO procedures for follow-up and non-compliance

All RTOs have procedures in place to follow up when the registrar fails to complete WBAs. But there are exceptions when it is expected that the WBA is purely a matter between the registrar and their supervisor (e.g., with learning plans). In the vast majority of the events the first follow-up is through email followed by a phone call when there is insufficient response (n=20). In some events it is either email or a phone call as the initial follow up (n=8). In seven events only email is used and in one event only telephone contact is used. For two WBA events either direct contact (not further specified) or continuous prompting through the ME takes place.

Failure to complete the WBA led to different consequences for the registrar. These are summarised in Table 11.

Table 11. List of consequences of failure to complete a WBA

Consequence	WBAs	Number of occurrences
Flagged	DOV-external; DOV-supervisor;	2
Flagged; non-progress	Learning plans	1
Flagged; remediation	MSF-colleagues, video, term assessment	3
Flagged; remediation; non-progress	DOV-supervisor, term assessments	2
Non-progress	DOV-external, Learning plans, MSF-colleagues, MSF-patients, PETAL, term-assessments, Video	8
Remediation	DOV-external, MSF-colleagues, PETAL, term-assessments	4
Remediation; non progress	DOV-external, MSF-colleagues, term-assessments	4
None	DOV-external, learning plans, MSF-colleagues, PETAL, term assessments,	12

Different RTOs have chosen different consequences for non-completion of WBAs. Those WBAs that are linked to the highest level of consequence (i.e. flagging and remediation) are DOVs, term assessments and MSFs. This may reflect the high-stakes value placed on these particular WBAs. Striking is the high number of events where no consequences were attached to non-completion or

the response has been N/A. When comparing these responses with the purpose of the assessment there were five events in which the purpose of the WBA was summative yet consequences to non-completion were deemed non-applicable. In all those events there were follow-up procedures to ensure the registrar completes the WBA, but there were no formal consequences listed for when the registrar still does not complete the WBA. We interpret this as an indication that the follow-up procedure up until now has been successful, but also that the local rules and regulations do not foresee a situation where the registrar does not complete, even after follow-up.

Non-completion of WBAs by the supervisor had three possible outcomes: withholding of accreditation (n=7), withholding of payment (n=10) or no consequences (n=24) and one unknown. The numbers represent the WBA events.

How do Workplace-based Assessments map against the RACGP's domains of general practice?

Each RTO mapped each of their WBAs against each of the RACGP domains of general practice. Table 12 shows the mapping for each WBA. It can be seen from Table 12 that RTOs tended to agree that DOVs and term assessments mapped against each of the domains. However, there was variation in RTO perceptions of the domains to which PETAL, learning plans and MSF were mapped.

Table 12. Number of RTOs which mapped WBAs in each Domain

WBA	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5
DOV-external	9/9	9/9	9/9	9/9	9/9
DOV-supervisor	3/3	3/3	3/3	3/3	3/3
Learning plans	2/4	3/4	3/4	4/4	3/4
MSF-colleagues	6/6	4/6	4/6	5/5	6/6
MSF-patients	3/3	2/3	0/3	3/3	3/3
PETAL	2/4	4/4	3/4	3/4	3/4
Term-assessments	7/7	7/7	7/7	7/7	7/7
Video	3/3	3/3	3/3	3/3	3/3

Note: The denominator in this table is the number of RTOs who use this WBA. Excludes learning plans for five RTOs who were not asked about domains.

The overall assessment numbers for each RTO mapped to the domains is in the table attached in the appendix[6].

Domain one: communication and the patient

35/39 events mapped onto this domain. These pertained mainly to the WBAs: DOV-external and supervisor, MSF-colleagues and patients, term assessment and Video. In the four events where the respondents indicated that they did not map onto this domain it concerned: learning plans (n=2) and PETAL (n=2).

[6] In this table the results of a total of 4139 events are used.

Domain two: applied professional knowledge and skills
35/39 events mapped onto this domain. These pertained to the WBAs: DOV-external and supervisor, MSF-patients, PETAL, term assessment and videos. The WBAs that mapped less to this domain were learning plans (n=1), MSF-colleagues (n =2) and MSF-patients (n=1).

Domain three: population health and the context of general practice
32/39 events mapped onto this domain and it is therefore the domain least well mapped by the WBAs. Mapping to this domain happened through: DOV-supervisor, DOV-external, Term assessment and Video. The events that did not map were in the WBAs: Learning plans (n=1), MSF-colleagues (n=2), MSF-patients (n=3) and PETAL (n =1)

Domain four: professional and ethical role
37/39 events mapped onto this domain. These pertained to all WBAs except for MSF-colleagues (n=1) and PETAL (n=1).

Domain five: organisational and legal dimensions
37/39 events mapped onto this domain. These pertained to the WBAs: DOV-supervisor, MSF-colleagues, MSF-patients, term assessments and videos. In four events the WBA did not map onto this domain, these concerned: Learning plans (n=1) and PETAL (n=1).

The DOV, both those performed by the supervisor or by an external observer, are seen to be mapping onto all five domains. This is not surprising as in directly observed patient consultations all important aspects of general practice have to be integrated and combined to provide good health care. The same applies to video observations. As with term assessments, these are typically used to integrate all information about a registrar's strengths, weaknesses and progress, which makes it logical that they would encompass all domains.

It is clear from these responses that all WBAs in-usage across the RTOs can map onto all domains, but the respondents did not indicate this to be the event for all WBAs on all occasions.

In summary, it is fair to conclude that the WBA program as a whole maps comprehensively onto all five RACGP domains, but that there is no specific association between the type of WBA and the domain mapping. This is a finding which is in alignment with the literature, which suggests that the content of the assessment (what is being asked or required of the trainee or registrar) is more important for the validity of the assessment than the form in which the performance is captured (how it is being asked).

What barriers exist to Workplace-based Assessment success in Australian General Practice Training?

Workplace-based assessment typically is only useful if the people who are using it take it seriously. Generally, there are two approaches to incentivise stakeholders to do this. One is to include the WBA in the assessment regulations, for instance by making the WBA mandatory.

Because of the mandatory nature of WBAs, completion rates are generally high. Most responses indicate 100% completion rate or at least 95% (n=3) or between 90-100% (n=2). Two responses were 'uncertain' and one 'variable'.

We asked for information from representatives of all participating training organisations about what they saw as the major issues concerning or hampering WBA from three different perspectives: that of the registrar, that of the supervisor and that of the training organisation. We

did not interview supervisors or registrars, as undertaken as part of stream 3b, however training organisation representatives were asked what they believe are the major issues for the various stakeholders. The outcomes are discussed first as a short narrative.

Problems associated with the WBA from the perspectives of the registrar

DOV

Three main themes could be distinguished in the open-ended comments in response to the question about the major problems associated with the DOV from the registrar's perspective.

Stress of being observed:

Registrars find it stressful to perform a patient consultation knowing that they're being observed and judged or evaluated. As one respondent puts it:

> *Threatening to have someone (who may not be known to them) sitting in and may disrupt the 'flow' of the consultation.*
>
> *Some indicate that it is stressful at first but when they become more used to it they find it a more valuable experience.*

Challenging logistics of setting up videos and finding suitable patients:

There are several logistical aspects that can make the DOV challenging. Finding the right patient, obtaining consent from the patient, finding a suitable time to ensure that all relevant stakeholders are in the same room and finally, ensuring that the video equipment is set up and working.

> *Logistics of organising the video, obtaining consent, potential impact on the normal 'flow' of the consultation, meeting the requirements of security of the information*

Quality of the assessment being dependent on the patient and expertise of the assessor:

From the viewpoint of standardisation, respondents felt that the individual patient selection, the complexity or simplicity of the event, and the experience or expertise of the assessor, all played a role in what they perceive as being – at least partly – a summative assessment.

> *Quality of the assessment is very dependent upon the patients seen and the assessor's abilities.*

Learning plans

The registrars see the learning plans merely as a tick-box exercise, which is overly bureaucratic. Therefore, they feel that the time commitment needed to complete them is not in alignment with their assessment or with any educational value. The prescribed format of the learning plan, as a tool that is not user-friendly has also been raised.

MSF

The main concerns around the MSF were related to their feasibility. In some events it is difficult for the registrars to find sufficient respondents or they require more support from the practice support staff. In a minority of events concern was expressed that the feedback, especially where it was negative, would be too confronting.

PETAL

The main concern regarding PETAL is the time commitment. Some respondents saw this WBA as relatively new and they have had very little experience. Some felt any educational or assessment value was not justified because of the time needed to complete this WBA.

Term assessments

The most important point of concern is the quality of the feedback. Several respondents indicated that either feedback is not given, given too late or that potential negative feedback is avoided and registrars are routinely marked as 'at expected level'.

In general, it is clear that most concerns around the WBAs are related to what could be called a 'return on investment' problem. Registrars are expected to invest time and effort into making the WBAs happen and feel that they do not receive sufficient educational or assessment 'return' of this investment. This is partly due to their perception of 'subjectivity', lack of expertise in the assessment judgement, and, most importantly, lack of useful and constructive feedback.

Problems associated with the WBA from the perspectives of the supervisor

DOV

With the DOV, the supervisors' concerns are mainly about time and planning and their own insecurity or lack of training in giving negative feedback.

Learning plans

The overall concern here is the lack of engagement. Supervisors find it very difficult to engage with learning plans if the registrars do not see them as a valuable activity.

MSF

Supervisors generally report no problems with the MSF. Some indicate that they feel that they are not sufficiently involved in the process.

PETAL

Supervisors expressed no major concerns with this WBA, except perhaps the fact that the time spent on PETAL is not spent on revenue-generating patient consultations.

Term assessments

The overriding concern with term assessments is the feeling that they are an overly bureaucratic system, which is unduly time-consuming. Some expressed concern about their ability to provide helpful feedback.

Problems associated with the WBA from the perspectives of the RTO

Training organisations indicate a myriad of different problems associated with the implementation and running of WBAs. For the purpose of this report these have been categorised into 12 overarching themes. In total these 12 themes were mentioned 43 times by the training organisations pertaining to different WBAs. The results are presented in Table 14.

Table 13. Frequency of identified issues from the perspective of the RTO

	Theme	*Illustration as an example*	*Number of times mentioned*
1	Cost	*Cost, large logistic exercise*	*12*
2	Workforce	*Availability of appropriately trained assessors to ensure consistency*	*17*
3	Compliance	*RTO needs to encourage the supervisors to complete the feedback reports with their registrars present as an option to support progress feedback*	*10*
4	Resourcing	*Resource intensive (requires data entry from paper forms)*	*12*
5	Online platform reliability	*Online learning platform has been problematic, in terms of registrars and supervisors being able to access their feedback forms. Some had completed them, and the system then did not save their work so they lost it.*	*2*
6	Patient demographic variability	*Unable to predict the number and range/type of patients they will see.*	*3*
7	Variable quality and quantity of information	*Part time registrars can struggle to have the requisite number of questionnaires completed in the time frame for data collection.*	*9*
8	Buy in	*The engagement of both registrar and assessor in the assessment process*	*10*
9	Training assessors	*Upskilling all supervisors to provide effective feedback and dealing with registrar's confidence to perform whilst being videoed*	*9*
10	Legal	*Secure transport, storage and deletion of the video recording*	*1*
11	Logistics	*Cancelled visits are hard to reschedule, visits can be expensive due to the distance and travel involved.*	*6*
12	Timeliness of completion	*Ensuring main supervisor completes report and ensuring that report is based on accurate data*	*5*

There is a striking pattern in the table of issues: cost, workforce, compliance, resourcing and buy-in are the most frequently mentioned issues. There seems to be a connection that in any WBA the stakeholders need to see a return on investment. The assessment is time-consuming and costly – not in the least because they detract from seeing patients (cost, resourcing), and therefore the users have to see an added value of the process. If this is not the case, and the issues suggest that it is far from optimal, stakeholders disengage (compliance, buy-in) and will 'go through the motions'. An important factor in this is to ensure that the assessment outcome and the feedback are managed by people with sufficient assessment expertise (assessment literacy as it is called in the literature[7]) to ensure a perceived return on investment. The table suggests that this is a central concern underlying most of the other issues reported here.

How do individual, interpersonal and contextual factors influence the use and outcomes of Workplace-based Assessments?

The issues identified in Table 13 could be associated with numerous aspects relating to either the WBA, the users or the organisation, and the data would allow for myriads of associations to be explored. However, from a statistical point of view this would not be advisable. As explained in the

[7] Popham WJ. Assessment Literacy for Teachers: Faddish or Fundamental? *Theory Into Practice* 2009;48:4-11.

methods section of this report, every statistical test has the possibility of an incorrect conclusion, especially concluding that there is an association or correlation where in fact there is not (the typical Type I error). When the acceptable level for the probability of this error is set to 5% ($p \leq 0.05$) this relates to each test. When performing multiple random (not hypothesis driven) statistical tests the likelihood of finding 'significance' due to a type I error increases. For instance, with 2 tests it goes up to 9.75%, with 3 to 14.26%, with 4 to 18.55%, etc.

Despite these caveats, we performed some exploratory analyses into possible associations between the issues and characteristics of the WBA or organisation, which are presented in the following tables.

First, we analysed the relationship between the registrar-related issues and whether the WBA is used for summative purposes or not. For these analyses, using Kruskal Wallis tests, independent samples were used, testing the number of mentions of each issue against the four classes of purpose (formative only, summative only, formative and summative, formative and part of programmatic assessment). The issues we tested were:

Table 14. Lists of identified issues from the perspective of the registrar, supervisor and RTO

Registrar issues	*Supervisor issues*	*RTO issues*
Anxiety	Scheduling	Costs
Non intuitive online system	Costs	Workforce
Negative patient perception	Workflow	Compliance
Honesty given power distance	Non-intuitive online system	Resources
Process is bureaucratic	Process is bureaucratic	Platform reliability
Time lag to feedback	Vested interest in registrar success	Patient demographic variability
Time lag to self-reflection	Disagreement with ECT	Variability of quality and quantity of information
Lack of quality/quantity of feedback	Lack of assessor expertise	Buy-in
Failure to fail	Failure to fail	Training assessors
Observation bias	Usefulness to great registrars	Legal issues
Decreased workflow	No issues, good process	Logistical issues
Unclear WBA purpose	Logistical issues	Timeliness of completion
Patient demographic variability	Legal issues	
Insufficiently trained assessor		
Performance disagreement		
Legal issues		
Logistical issues		

None of the issues showed a significant relationship with the purpose of the assessment from the registrar's perspective (according to the training organisation's interviewee), except for the decreased workflow ($p=.011$). However, as this may be purely due to chance, a Bonferroni correction was applied (to compensate for the multiple test effect) and consequently the cut-off

for significance was set at .0029. In this event the effect of purpose on the number of times 'decreased workflow' is mentioned as an issue is not significant.

For the supervisor and the RTO-related issues none of the effects is significant except for logistical issues with a p-value of .004. But here again we applied the Bonferroni correction, which would require a cut-off of .0038, and so the effect is also not significant (though marginally).

When combined across all types of issues there is no relationship between the total number of issues and the purpose of the assessment, none of the relationships were significant (Kruskall Wallis). This is not surprising given that in the individual tables almost none of the effects were significant.

The second set of associations we examined is between the type of WBA and the reported issues.

Table 15. Are the registrar related issues related to the type of WBA?

Issue	*p-value*	*decision*
Anxiety	***.001***	***Reject H_0***
Non intuitive online system	.661	Retain H_0
Negative patient perception	.866	Retain H_0
Honesty given power distance	.684	Retain H_0
Process is bureaucratic	***.011***	***Reject H_0***
Time lag to feedback	.159	Retain H_0
Time lag to self-reflection	.537	Retain H_0
Lack of quality/quantity of feedback	.509	Retain H_0
Failure to fail	1.00	Retain H_0
Observation bias	.476	Retain H_0
Decreased workflow	.043	Reject H_0
Unclear WBA purpose	.241	Retain H_0
Patient demographic variability	***.033***	***Reject H_0***
Insufficiently trained assessor	.137	Retain H_0
Performance disagreement	.712	Retain H_0
Legal issues	.601	Retain H_0
Logistical issues	***.014***	***Reject H_0***

One issue was significantly related to the type of WBA, anxiety related to the WBA. The other three, bureaucracy of the process, patient demographic variability and logistical issues were significant at the $p \leq 0.05$ level but not after a Bonferroni correction. However, these issues were also mentioned in the narrative descriptions of the concerns. Direct observation, which happens in the DOV-type WBAs is anxiety-provoking and the others far less so, and hence there is a relationship. Other WBAs such as learning plans are not seen as useful, but as bureaucratic tick-box exercises, and therefore they are seen as more bureaucratic than DOV, which are more intuitive in their perceived learning effects.

The next table presents the finding concerning the relationship between the type of WBA and the issue from the supervisors' perspective.

Table 16. Are the supervisor related issues related to the type of WBA?

Issue	*p-value*	*decision*
Scheduling	.202	Retain H_0
Costs	.622	Retain H_0
Workflow	.818	Retain H_0
Non-intuitive online system	.316	Retain H_0
Process is bureaucratic	.245	Retain H_0
Vested interest in registrar success	.349	Retain H_0
Disagreement with ECT	.087	Retain H_0
Lack of assessor expertise	***.022***	***Reject H_0***
Failure to fail	.094	Retain H_0
Usefulness to great registrars	.349	Retain H_0
No issues, good process	.172	Retain H_0
Logistical issues	.012	Reject H_0
Legal issues	.219	Retain H_0

Although the lack of assessor expertise and logistical issues are significant in the table they would not reach significance at the Bonferroni corrected level ($p \leq 0.0038$).
Table 17 presents the same results but now from the perspective of the training organisations.

Table 17. Are the RTO related issues related to the type of WBA?

Issue	*p-value*	*decision*
Costs	.115	Retain H_0
Workforce	.219	Retain H_0
Compliance	***.031***	***Reject H_0***
Resources	.166	Retain H_0
Platform reliability	.274	Retain H_0
Patient demographic variability	.183	Retain H_0
Variability of quality and quantity of information	.373	Retain H_0
Buy-in	.648	Retain H_0
Training assessors	***.042***	***Reject H_0***
Legal issues	.219	Retain H_0
Logistical issues	.214	Retain H_0
Timeliness of completion	.697	Retain H_0

The issues of compliance and training of assessors are significantly related to the type of WBA but none remained significant at the Bonferroni corrected level (.0042).

Compliance is mainly an issue for the MSF (patients and colleagues) and to some extent the learning plans. The issues around the quality of the training of the assessors was mention in combination with various WBA; DOV-external (n=2), learning plans (n= 2), videos (n=3) and with DOV-supervisor and learning plans.

We then combined all the issues per stakeholder perspective (registrars, supervisors, training organisation) and tested whether the number of mentions was significantly associated with the type of WBA.

Table 18. Are the combined issues related to the purpose of the WBA?

Question tested	*p-value*	*decision*
Association between all registrar-related issues and purpose of the WBA	.722	Retain H_0
Association between all supervisor-related issues and purpose of the WBA	.154	Retain H_0
Association between all RTO-related issues and purpose of the WBA	.331	Retain H_0
Association between all issues and purpose of the WBA	.398	Retain H_0

The final question would be whether there is a difference between the RTOs with more rural than urban registrars and vice versa. In a first round we used a Kruskall Wallis test to evaluate whether there is a significant association between urban/rural registrar characteristics of the RTO and the total number of issues and the totals per stakeholder groups. For this comparison we categorised the training organisation as either having more rural than urban registrars or as having more urban than rural registrars. The results showed that none of these relationships had statistical significance.

What are RTO perceptions regarding the link between WBA performance and remediation and exam success?

Whether there is a relationship between performance or completion of the WBA with exam results is largely unknown. One respondent noted that there "seems to be a correlation between failure to complete tasks and failing to pass exams", but most indicate that they don't know or were unable to find any relationship in their data.

Thus, in general, RTOs are of the opinion that WBAs are not predictive of college exam performance:

- In 22/43 the response indicated that there was no predictive value
- 14/43 responses reported 'unknown',
- In 4/43 events respondents assumed that there was some predictive value
- 3/43 believed the outcome of the WBA was predictive of success in the college exams (Video, 2 DOV external).

Discussion

Although the data is diverse, and it may be difficult to establish clear numerical patterns we feel that some important conclusions can be drawn from the audit.

What WBAs are used and by whom?

There are similarities in how WBAs and learning are conducted across the RTOs. Directly observed consultations in any form and learning plans are used by all, and term assessments by all but two of the audited RTOs. This aligns with the requirements in the RACGP standards and aligns with the literature. The acceptability, most appropriate process and usefulness of these assessments should be moderated against information from the focus groups.

However, there are differences in how the WBAs are conducted, and what the consequences of the WBAs are in terms of flagging, formative feedback or summative decision-making.

More than half of the RTOs did not initially identify a learning plan as a WBA. This finding may indicate that learning plans are not perceived as a WBA. We conclude from this that there is a difference in view on what constitutes WBA and what does not.

Video DOVs present a reasonably cost-effective tool for conducting a direct observation of a registrar's consultations, and yet are only used by fewer than half of RTOs. It would be logical to assume that logistical reasons (having to set up the equipment, etc.) would be a main reason for this but it was only mentioned as an issue from a registrar's perspective in half the cases and not at all from the supervisors' and RTOs' perspective.

WBAs that require simple data collection can be administered by less qualified personnel, potentially reducing cost (eg. PETAL and MSF). However, any WBA that is not completed without meaningful feedback to the registrar that they can use for their own learning, is not fulfilling this purpose. For this aspect of PETAL and MSF, MEs are critical in delivering or co-ordinating the required synthesis of information, self-reflection and feedback.

So, irrespective of who is assessing or coordinating the delivery of the WBA, interpretation of the outcomes always requires ME guidance and expertise. This relates back to the literature in which there is increasing indication that assessment literacy of the users pays a pivotal role in the validity of WBA.

Variation in Organisational Contexts

The data from the audit shows broad variation in organisational contexts within which workplace-based assessment is delivered. Findings that will be further moderated against results from other streams of research, which may inform the development of the WBA Framework are:

- One can see the diversity of WBA programs across the RTO jurisdictions as a weakness in standardisation. The data, however, suggest otherwise. The huge diversity has not led to a major systematic variability in successful 'graduates' of the programs. The diversity in WBA programs is associated with a broad variation in the size of RTOs, the rural training volume and the ratio of registrars to supervisors, MEs and training coordinators within the current general practice training environment. The regional training model may therefore actually be seen as effective or even essential to ensure that the AGPT program can cater for regional needs and generate future workforce that is distributed across all areas, including those of workforce need. To maintain this regional model, it is clear that a WBA Framework needs to be flexible to accommodate regional contextual and demographic differences. This, again, is aligned with the modern literature that increasingly argues that bespoke interactions are more valid in WBA than an overly structure and standardisation.
- MEs are consistently responsible for the delivery of workplace-based assessment and programmatic assessment across Australian RTOs. It is apparent that the ME role is essential to lead the delivery and co-ordination of WBA and programmatic assessment at RTOs. Results from this stream should be moderated against feedback from Stream 3b-the focus groups on WBAs.
- There is variation in whether MEs make decisions about registrar progression through a programmatic assessment committee. It is important to determine if better outcomes emerge from those RTOs with programmatic assessment committees. This will be explored in Stream 3d.

- MEs are not always involved in the documenting requirements within the registrar portfolio. This may be a role that can be more cost-effectively picked up by administration staff, allowing the ME further time to devote to the delivery of WBAs.
- MEs are currently expected to, and are supported to, attain additional training within medical education and education theory, however this varies amongst RTOs.

For what purposes are WBAs used?

- There is a huge variety in purposes for which each individual WBA at each moment is used. Some are used to make decisions in isolation – purely summative – and some entirely for feedback to the registrar without any consequences – purely formative. However the majority are used in conjunction with each other, for instances to determine a pattern of issues – like a flagging system. This should not be confused with programmatic assessment as that would require a regular review of all information in all assessment across methods to monitor and steer progress in all five RACGP domains. Although it has not been explicitly asked and though some RTOs report using the WBA in a programmatic fashion, the issues reported (logistics, buy-in, doubt about the assessment literacy of the persons involved in the WBA process, etc.) suggest that this is the most promising area for improvement of the WBA processes.

How do RTOs support, monitor and regulate the use of WBAs?

- All RTOs employ a WBA program across GPT1, 2 and 3 and sometimes also in GPT 4, and for the direct observations the frequency of assessments is similar across RTOs.
- All RTOs rely predominantly on face-to-face and online methods to provide the registrar with feedback. We have no data however on exactly how this feedback is being delivered, and we assume this depends on the specific interaction between the registrar, supervisor and ME.
- From the reported issues, it is clear that a solid and easy-to-use – online – infrastructure is essential and that any measure that alleviates the logistical and bureaucratic burden on registrars and supervisors is welcomed with open arms.
- The mode of delivery of feedback may seem efficient in most cases, the timing and the content of the feedback appear to be open for improvement in many cases.
- Finally, although in all cases there is an intent and procedures in place to follow up non-completions of WBA, these are not always associated with consequences for the registrar and/or supervisor.

How do WBAs map against the RACGP's domains of general practice?

- RTOs generally agreed that DOVs, term assessments and video review mapped against all of the domains.
- RTOs had a high level of variation in how they mapped MSF, learning plans and PETAL against RACGP domains
- An agreed mapping document needs to be created to apply across WBA delivered in the AGPT program.

What barriers exist to WBA success in the AGPT program?

- Concerns were expressed about the required level of assessment literacy of all stakeholders in the WBA process, especially the registrars and the supervisors. The MEs do play a role in bringing the stakeholders together and provide them with necessary knowledge and understanding on a more ad-hoc basis but formal training does not always seem to happen. As stated before, such assessment literacy is extremely important in making the WBA valid.
- Concerns were expressed about the bureaucratic nature of some processes and how they tend to lead to lack of buy-in and compliance. WBA assessment is reliant on the interaction between the learner (the registrar) and the knowledgeable other (the supervisor and ME) and without buy-in and compliance the processes run the risk of becoming ritual dances. These take time but do not lead to a more valid assessment or a more positive impact on development of competence. The issues flagged in Tables 15-17 may therefore be of primary concern.

How do individual, interpersonal and contextual factors influence the use and outcomes of WBAs?

- The individual, interpersonal and contextual factors can play a hampering role in the use and outcomes of the WBAs. When the process is seen as overly bureaucratic and not delivering sufficient return on investment (time, loss of workflow, etc.) there is the risk of retreatism or 'going through the motions'. Follow-up after non-completion can only play a minor role as there are few levers to enforce a genuine engagement. For this to happen the stakeholders will have to experience the processes as intrinsically meaningful, which according to our audit results is one of the challenges.

What are RTO perceptions regarding the link between WBA performance and remediation and exam success?

- RTOs generally do not perceive that there is a link between WBAs and exam success. However, those RTOs who did feel there may be a link, indicated that exam success or failure could be predicted by a Direct Observation Visit. This type of assessment seems to be highly valued by RTOs, who tended to have higher consequences for non-completion of DOVs, whilst also having clear mechanisms in place to ensure both face-to-face feedback and written feedback was available to all parties for this category of assessment.

Conclusion

In conclusion, the WBAs are varied in their implementation across the different training organisations, with varied purposes behind their uses. In all training organisations the content of the WBA has mapped across nearly all RACGP domains. This may be due to the fact that generally, the mapping onto a domain depends on what is being asked or required in a WBA rather than what form it takes. A DOV for example could be seen to focus more on domain 1 'communication and the patient' in one case and on domain 2 'applied knowledge and skills' in another. Logically, most cases will easily map onto multiple domains. The literature is not contrary to this and contends that what any assessment asks or requires of the candidate is more important than how the performance is captured (so, the content of the assessment is more important than the assessment format). What also becomes clear is that there is a considerable divide between the assessment as it has been intended and designed, and how it can be perceived.

The perception and utility of WBAs varies across the different RTOs, with each organisation adopting a fit-for-purpose approach. This is not a weakness of standardisation but rather a strength of the bespoke approaches that WBA requires. This being said, the strength of the WBA processes is dependent on the interaction between the stakeholders, and the MEs (and training coordinators) play a central role. We feel that MEs may be seen as the back-bone of registrar training. Their academic training, oversight, longitudinal relationship and brokering role, is not replicated in any other position. Because of their ability to provide rich input into programmatic and remediation discussions, they are integral to final decisions about whether registrars have completed training.

Appendix A: Overall assessment numbers for each regional training organisation mapped to the RACGP domains of general practice

WBA	*Training organisation*	*Domain 1:*	*Domain 2:*	*Domain 3:*	*Domain 4:*	*Domain 5:*
DOV-external	5	Yes	Yes	Yes	Yes	Yes
DOV-external	3	Yes	Yes	Yes	Yes	Yes
DOV-external	2	Yes	Yes	Yes	Yes	Yes
DOV-external	8	Yes	Yes	Yes	Yes	Yes
DOV-external	7	Yes	Yes	Yes	Yes	Yes
DOV-external	9	Yes	Yes	Yes	Yes	Yes
DOV-external	1	Yes	Yes	Yes	Yes	Yes
DOV-external	4	Yes	Yes	Yes	Yes	Yes
DOV-external	6	Yes	Yes	Yes	Yes	Yes
DOV-supervisor	5	Yes	Yes	Yes	Yes	Yes
DOV-supervisor	6	Yes	Yes	Yes	Yes	Yes
DOV-supervisor	2	Yes	Yes	Yes	Yes	Yes
Learning plans	5	Yes	Yes	Yes	Yes	Yes
Learning plans	3	No	No	No	Yes	No
Learning plans	1	Yes	Yes	Yes	Yes	Yes
Learning plans	9	No	Yes	Yes	Yes	Yes
MSF-colleagues	5	Yes	No	Yes	Yes	Yes
MSF-colleagues	8	Yes	Yes	Yes	Yes	Yes
MSF-colleagues	1	Yes	Yes	No	No	Yes
MSF-colleagues	8	Yes	Yes	Yes	Yes	Yes
MSF-colleagues	6	Yes	No	Yes	Yes	Yes
MSF-colleagues	4	Yes	Yes	No	Yes	Yes
MSF-patients	3	Yes	No	No	Yes	Yes
MSF-patients	2	Yes	Yes	No	Yes	Yes
MSF-patients	4	Yes	Yes	No	Yes	Yes
PETAL	1	Yes	Yes	Yes	Yes	Yes
PETAL	7	No	Yes	No	Yes	Yes
PETAL	3	Yes	Yes	Yes	Yes	Yes
PETAL	5	No	Yes	Yes	No	No
Term-assessments	2	Yes	Yes	Yes	Yes	Yes
Term-assessments	3	Yes	Yes	Yes	Yes	Yes
Term-assessments	6	Yes	Yes	Yes	Yes	Yes
Term-assessments	1	Yes	Yes	Yes	Yes	Yes
Term-assessments	9	Yes	Yes	Yes	Yes	Yes
Term-assessments	5	Yes	Yes	Yes	Yes	Yes
Term-assessments	7	Yes	Yes	Yes	Yes	Yes
Video	6	Yes	Yes	Yes	Yes	Yes
Video	2	Yes	Yes	Yes	Yes	Yes
Video	5	Yes	Yes	Yes	Yes	Yes
Total "Yes"		***35***	***35***	***32***	***37***	***37***

Stream 3a
GP Registrar, Supervisor and Medical Educator use of Consultation Observation as an Educational and Assessment Tool

Dr Cat Kirby (Eastern Victoria GP Training)
Dr Julie Willems (Eastern Victoria GP Training)

In collaboration with

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

The researchers would also like to acknowledge:

- The Working Group for this stream including: Dr Cat Kirby (Eastern Victoria GP Training), Dr Julie Willems (Eastern Victoria GP Training), Assoc. Prof Jill Benson AM (GPEx), Dr James Brown (Eastern Victoria GP Training), Dr Taryn Elliott (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University) and Ms Michelle Pitot (GPEx).
- Eastern Victoria GP Training for their support and contributions to the research and this final report.
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.
- With special thanks to all participants who contributed to this research.

Introduction

The advantages of teaching general practice (GP) registrars in the workplace have been outlined as having benefits for the GP registrars, supervisors, practices and patients alike. As Laurence et al. (2012) note that for supervisors, these benefits range from mentoring trainees into the profession through to helping them learn; for practices, the provision of an additional workforce with potential for succession planning; for patients, an appreciation of their role in training future general practitioners, longer consultations, and potentially new perspectives that trainees can provide on their conditions; and for the trainees themselves, the development of their trade and the potential to identify a practice to return to after completion of training. Assessment of learning and evaluation of competency in the workplace-based context are key to professional development.

There are a number of core competency-based workplace assessment tools available for workplace-based assessment of registrars in Australian GP training, and one such is consultation observation. Consultation observation (CO), also known as direct observation (DO), is used by medical educators and GP supervisors to assess registrars' clinical skills (Rencic et al., 2016) and competency (Khan et al., 2013). It also assesses the breadth of skills, such as 'soft' skills foe example effective communication (Essers et al., 2013).

In turn, this assessment provides the basis for feedback to the registrar, the tailoring of supervision to meet registrars' competence (Wearne & Brown, 2014), and grist upon which to frame further education. Beyond the assessment, feedback and education loop, the aim is to assist registrars to become self-reflective and self-evaluative, and using these multiple sources of information to inform their personal learning plans (Wearne & Brown, 2014).

Consultation observation is mandated by both Australian Colleges – The Royal Australian College of General Practitioners (RACGP) and Australian College of Rural and Remote Medicine (ACRRM). The RACGP mandates External Clinical Teaching Visits (ECTVs) which are half day sessions of observation conducted by a visitor to the registrar's practice. The visitor can be either a medical educator, a GP supervisor from a different practice, or in some cases, a GP from the registrar's own clinic. Further, there is considerable variation in approaches to consultation observations such as ECTVs across the various GP training organisations in Australia. This is recognised amongst training providers themselves, and has also been documented in past research (e.g., Usherwood, Kefalas, Knight, 2005).

ECTVs are framed primarily as a formative assessment tool in GP training that holds strong educational value for the registrar. For example, from the EV Training Practice handbook (2019, p.25):

The purpose of ECTVs are to:

- Help improve the registrar's skills, both as a GP and as a professional.
- Assist the registrar to develop a vision of what constitutes excellence in general practice consulting.
- Make an appraisal as to whether the registrar's knowledge and clinical skills are appropriate for their level of training.

ACRRM also mandates formative and summative Mini-Clinical Evaluation Exercises (Mini-CEX) during training, and these also involve consultation observation. Frameworks and procedures are provided to give structure to these consultation observation events by all Australian general practice training organisations (RTOs). Work has been done internationally to examine assessment

outcomes of Mini-CEXs as measures of competency and performance (e.g., Kogan et al., 2009). Work has also been done in Australia to examine assessment outcomes of ECTVs as measures of external exam performance and for the need for remediation. (Stewart et al., 2018; Magin et al., 2017).

The benefits of consultation observation as a teaching and assessment tool abound in the literature. One key aspect is that it occurs in an authentic setting (Norcini, 2003) and in real time. Another is that it can assess the non-clinical components, including the 'soft' skills of the registrars (Kogan et al., 2017). Positive influences on strengthening the relationship between the observed and the observee have also been noted (Weller et al., 2009).

However, concerns about the use of consultation observation for assessment and education have also been raised. These could be summarised as relating to cost limitations, patient safety concerns, nonstandardised settings, infrequent clinical events (Brydes et al., 2015), and variations between RTOs (Usherwood, Kefalas, Knight, 2005). Other areas of concern can include: assessor reliability and/or variability, judgement bias (Norcini, 2003), the impact of using consultation observation for assessment on its value for feedback and learning, barriers to being observed or receiving feedback, feedback that is subjective and not objective, unevenness in case mix (Norcini, 2003), observer's lack of educational training, potential of resistance by registrars and supervisors/MEs to engage in consultation observation, and negative influences on the relationship between the observer and the observee (Weller et al., 2009).

In addition, one cannot examine a tool without looking at the overlay of contributing factors including individual differences and contextual factors that co-exist in a given moment of time. For example, by considering the contextual factors in consultations of general practitioner registrars, evidence can be gathered as to how these might impact on the assessment in the authentic setting, including impact on such aspects as communication (Essers et al., 2013). Additionally, the impact of the key contributions of GP supervisors in the success of the process are unclear or unseen (Wearne at al., 2012). The determinants of the utility of consultation observation for assessment and education have therefore been identified as observer factors, observation tool factors, registrar factors and the context in which observation occurs.

Not having the evidence base to determine how beneficial consultation observation is as a tool is critical. As Brydges et al. (2015, p. 246) note, "establishing the necessary evidence base will permit educators and researchers to use these surrogates as the primary means of assessment during day-to-day practices". In order to progress consultation observation as a work-based assessment tool we need to identify not only what is already known in terms of the literature, but also what happens in reality. We therefore need to better understand how the triad of observer, observed and the observation tool interact within specific contexts, as well as the overlay of contextual influences.

By knowing these things, by building a body of evidence, we will be better informed in influencing how to design our consultation observation processes and the training and support we provide to our medical educators, supervisors and registrars in consultation observation. In this way, consultation observation can be used more effectively and reliably as an assessment tool and more meaningfully as an educational tool for feedback and reflection.

Aim and research questions

This study aimed to explore and identify important features and psychosocial dynamics of consultation observations as an assessment and educational tool. The study focussed on interactions between the observer, observed, observational tool and context of the observation,

and the ways in which these interactions may facilitate or impede consultation observation as a method of assessment and education.
Several research questions guided the study:

- How is consultation observation being framed, from an assessment and educational standpoint, in GP vocational training?
- What enables and impedes consultation observation as an assessment tool?
- What enables and impedes consultation observation as an educational tool?
- How can consultation observation be framed and supported so that it is both an informant of registrar progress and competency, and an educational tool?

Methods

Design

This qualitative study utilised in situ audio-recordings of debriefing discussions between registrars and their supervisor or ME following a consultation observation event, followed by semi-structured one-on-one interviews. The study focussed on formalised consultation observation events: External Clinical Teaching Visits (ECTVs; VIC) and Direct Observation Visits (DOVs; SA).

Research participants

Nine pairs of registrars and their observing supervisor or ME were recruited via two Australian RTOs (EV and GPEx). Sampling included participant pairs from both metropolitan (n=6) and rural (n=3) locations, and included registrars at various stages of training (GPT1 n=5; GPT2 n=3; GPT3 n=1). A summary of participant details is included in Table 1 below.

Table 1. Participant details.

Pair	RTO	Location	Registrar Training Level	ME or Supervisor Observing
1	EV	Rural	GPT1	Supervisor
2	EV	Metro	GPT2	ME
3	EV	Metro	GPT2	ME
4	EV	Rural	GPT1	Supervisor
5	GPEx	Metro	GPT2	ME (+supervisor)
6	GPEx	Metro	GPT1	ME
7	GPEx	Rural	GPT3	Supervisor
8	GPEx	Metro	GPT1	Supervisor
9	GPEx	Metro	GPT1	Supervisor

Data collection

Two types of data were collected from each pair. First, participant pairs audio-recorded the debriefing discussion(s) that occurred immediately following one or more of their observed consultations as part of their ECTV or DOV. As methods of debriefing and feedback vary across individuals, participants were given the flexibility to record one or more debriefing discussions as they would naturally occur. Secondly, one-on-one semi-structured interviews were conducted with each participating supervisor or ME and registrar in the two to six weeks following the ECTV/DOV. Whilst we aimed to undertake interviews within a two-week period, this time period was extended for some pairs due to participants being unavailable over the holiday period.

The interviews explored participants' perceptions and experiences of the consultation observation tool and process (i.e., ECTV or DOV). The interview questions (refer Appendix 1 and 2) were arranged around several focal points: the observer (supervisor/ME); the observee (registrar); the relationship between the observer and observee; contextual factors; and other issues of importance to the participant.

As part of the interview process, 20 interviews were scheduled. Of these, one was a repeat interview due to a failed recording on the audio device (R1). Interviews were digitally audio-recorded and ranged in duration from 26 to 66 minutes.

All audio-recorded data were transcribed verbatim and transcripts were cleaned and de-identified prior to analysis.

Ethics approval

This study was approved by the Monash University Human Research Ethics Committee (Project ID: 17491).

Research paradigm

We adopted an interpretivist approach to data collection and analysis; utilising both the recordings of in situ discussions and subsequent participant reflections to understand registrar, supervisor/ME experiences and perceptions of consultation observation. The interpretivist paradigm focusses on individual's experiences and meaning-making of given situations: "individuals with their own varied backgrounds, assumptions and experiences contribute to the ongoing construction of reality existing in their broader social context through social interaction" (Wahyuni, 2012, p.71). As researchers working within this paradigm, we have taken an 'insider perspective' drawing on our own experiences in general practice training, and teaching and learning, to inform the analysis.

Figured Worlds theory (Holland et al., 1998) was used as a theoretical lens for analysis of the debriefing discussion recordings. Figured Worlds highlights the interconnectedness and mutual reciprocity between individuals (e.g., registrar and supervisor/ME) and their social context (general practice clinic, RTO), and the meaning-making and formation of identity (e.g., the 'good doctor') that occurs in relation to interactions (Bennett et al., 2017). This socio-cultural theory has been used in broader education research to understand identity development of learners and teachers, and is now being applied within medical education (e.g., Bennett et al., 2017).

Data analysis

Two main approaches were taken to analysis of the interview and debriefing data.

First, a simplified conversation analysis (CA) approach was utilised to analyse the transcripts and audio-recordings of each participant pair's debriefing discussions. CA offers a method for examining the practices that participants use to construct an interaction. This approach was chosen as it is well suited to analysis of naturally occurring interactions (Maynard & Heritage, 2005). This analysis is about making the implicit explicit, examining what talk does as well as what it says. The analysis focused on interactional features and patterns that impacted on: the educational experience for the registrar; and, the value of the assessment of registrar performance.

Second, the interview data were analysed thematically using an inductive and deductive approach to identify key themes in the data. The thematic analysis followed the six phase recursive process

(Braun & Clarke, 2006; Clarke & Braun, 2013) that included (1) initial familiarisation with the data; (2) generation of initial broad codes; (3) searching for themes; (4) reviewing themes; (5) defining themes; and (6) naming themes. Themes were clustered in relation to broad categories of observer, observee, observational tool and context, and further analysed in relation to the guiding research questions.

Findings

The research findings are broken down into two sections. First, key findings from the debriefing discussions are presented, tying in data from the interviews regarding participant perceptions and experiences of the debriefs. Second, key themes from the interviews are presented and discussed. These findings are then brought together in the Discussion and discussed in relation to the guiding research questions and implications for the proposed WBA Framework.

Key findings from debriefing discussions

The recorded debriefing sessions offered insight into the common pattern of consultation observations and debriefs, and the variations occurring within that pattern. Analyses of the interactions between each observer and registrar also highlighted the focus, agendas and adopted positions of each person within the event. Figure 1 provides a summary of the ways in which consultation observation and debriefing events occurred, and some of the structural variations within this that were determined by each or both of the observer, and the registrar.

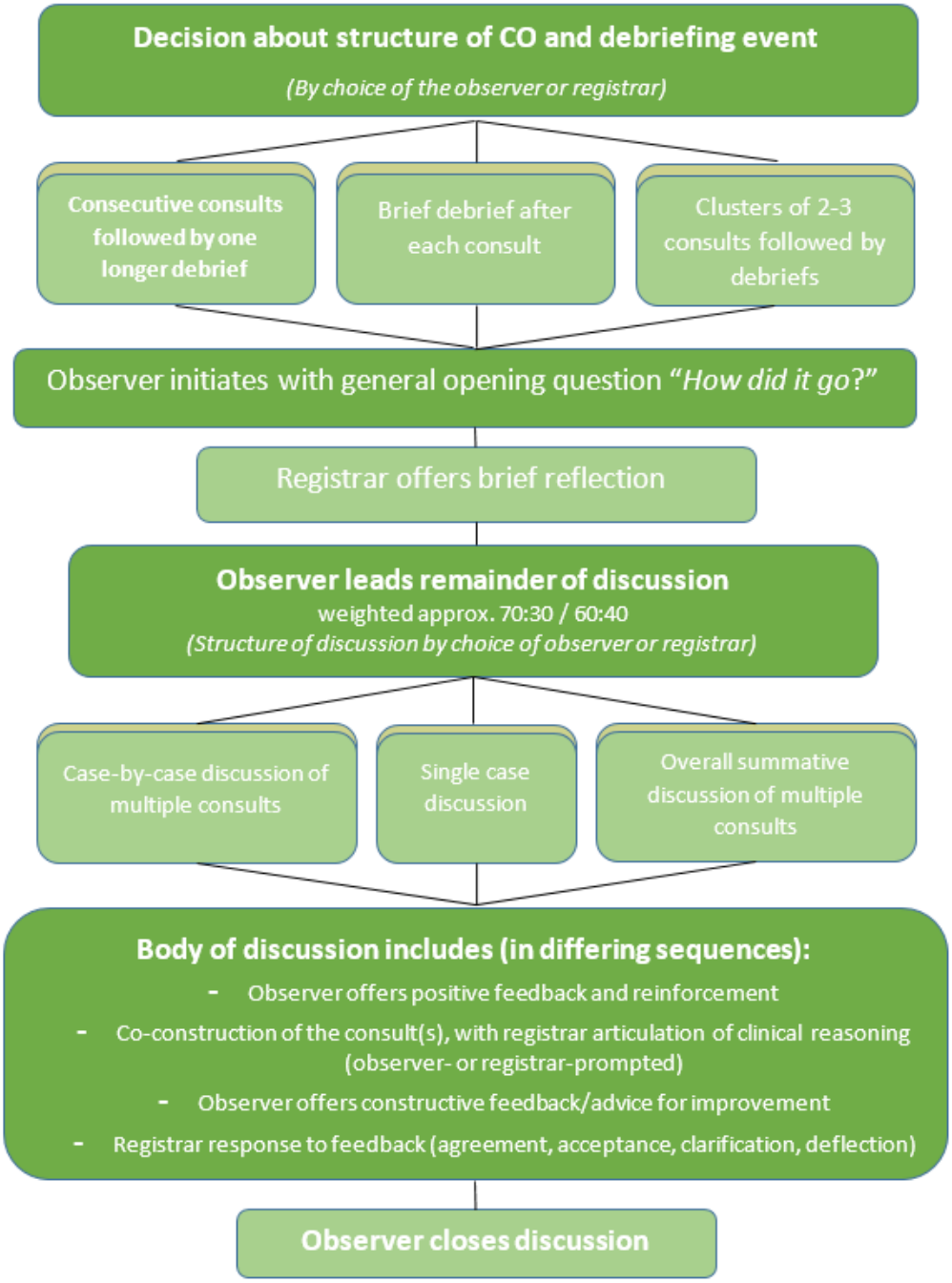


Figure 1. Consultation observation and debriefing phase structure.

Table 2 provides an overview of some of the notable structural, interpersonal, contextual and content features of the debriefing discussions. This shows some of the common features of consultation observation debriefs across individuals and RTOs, and also some of the observable variations that may impact on the educational value of the event. These are discussed below.

Table 2. Structural, interactional and contextual features of the debriefing discussions.

Structure, interactions and context	*Range*
Contextual factors	
Registrar stage of training	GPT1 to GPT3
Debrief with: supervisor or ME	6 supervisors, 3 MEs
Prior observations and/or other professional interactions between the pair (i.e., familiarity)	8 yes, 1 no
Structural factors	
No. of recorded debriefs	1 to 3
Duration of debriefs (minutes:seconds)	4:54 to 43:28
No. of observed consults prior to recorded debrief	1 to 7
Duration of ECTV/DOV event in total	Typically half day session, some variation (e.g., ad hoc single consults when time available)
Features of debrief	
Opening (question, statement, request)	Observer typically makes a brief statement on the consult(s) and offers a broad opening question to elicit registrar's initial reflections
Explicit vs implicit approach to structure of discussion	Sometimes implicit, sometimes discussed at beginning of debrief with the observer suggesting an approach or seeking registrar preference
Registrar self-evaluation and reflection	Present in each debrief
Observer feedback	Present in each debrief, both in terms of positive reinforcement and constructive feedback/advice
Registrar demonstration of clinical reasoning	Present in each debrief to varying degrees
Registrar advice/information seeking	Very little; mostly observer-led advice
Registrar responses to observer's feedback	Largely receptive and positive; some minor natural defensiveness
References to assessment tool (i.e., forms/reports to RTO)	Nil
Interpersonal dynamics	
Opener of discussion	Always observer
Ratio of talk (observer:registrar)	Typically 70:30 or 60:40 (approx.)
General control of the conversation	Typically observer-led
Interactional style between observer and registrar	Educator-learner; collegiate mentorship; maternal mentorship; familiar supervisory relationship (i.e., primary supervisor, or supervisor within clinic

The debriefing interactions were analysed to explore the way in which structural, interactional and contextual features may impact on the educational value of the event for the registrar, and the observer's ability to form a well-reasoned judgement about the registrar's competence based on their performance. The analyses also drew on the interview data to identify participant perceptions and preferences relating to these identified features, which are outlined briefly below.

Familiarity and past observations between registrar and supervisor/ME

Familiarity between observer and registrar appeared to have several benefits in terms of:

1. The registrar's level of ease in being observed, assessed and offered feedback.
2. The observer's capacity to form a judgment about the registrar's performance; particularly given the continuity of observing how the registrar has developed over time.
3. The perceived credibility of feedback offered by the observer (from the registrar perspective), as the observer offers reflections on the registrar's development over time.

Examples from three separate participant pairs showing their longitudinal relationship in action are provided below. The quotes highlight the benefit of familiarity and repeated assessments over time, in terms of the observer's capacity to offer specific feedback on the registrar's progress over time, and also the capacity to reflect on the registrar's enactment of feedback offered by the observer on previous occasions.

> *ME: "compared to my last visit, I think that you're really coming along well, particularly in terms of your confidence levels which is really nice to see because I think that one of the things that we talked about before was that I felt the message you were giving across was good but you perhaps weren't doing it in a very confident manner and I think that's really improved over the last few months and that you're actually able to express your management plans with a bit more conviction which I think inspires a bit more confidence from the patients as well…"*
>
> *ME: "…like I've said you've got better at deferring issues to later consults, and I reckon, because this is my second visit with you, I reckon you're much better at doing that now then you were the first time."*
>
> *ME: "… I think your approach with the headache diary was really good. Pulling that out, getting written information. Well done."*
>
> *Registrar: "Yes, using your advice from the last… "*

Sequencing of observed consults and debriefs

Debriefing after each observed consult offers valuable immediacy and specificity in terms of the feedback provided from observer to registrar, and the recollection of clinical reasoning offered by the registrar to facilitate the observer's assessment and feedback. Immediate debriefing after each consult also enabled observers to offer positive reinforcement for registrars, to reduce anxiety associated with the event. This approach also lent a sense of closure on each patient case before moving onto the next consult.

On the other hand, some participants engaged in one longer debrief at the end of all observed consults. In these instances, the initial stage of the debrief was often more generalised and feedback offered by the observer was summarised across observed consults, followed by case-based discussions. The observer sometimes appeared to refer back to notes to prompt their memory of each consult (as would be expected for this scenario). However, this suggests that some of the more specific detail of the observer's reflections and feedback may be lost in this approach.

As a 'midpoint', some participants preferred to cluster their observations and debriefs, i.e. two or three consults followed by a debrief, and then repeated again. This seemed to come down to the personal preference of the observer and/or the registrar, and still offered some immediacy and specificity in observer feedback and registrar self-reflections.

Contributions of debriefing to assessment

The debriefing discussions undoubtedly offer value in facilitating observer evaluation of registrar competence. The debrief offered observers the opportunity to dig deeper into the registrar's clinical reasoning; potentially facilitating more informed and valid assessment of the registrar. Registrars were often quick to articulate their clinical reasoning; to demonstrate their competency and mitigate any observations made by the supervisor/ME that they felt may not have been accurate.

Shared understanding and expectations of consultation observation

There appeared to be a common understanding of the ECTV/DOV as primarily an educational experience for the registrar. Each of the participating observers offered (1) constructive feedback with specific areas in which the registrar could improve or learn more, and (2) positive reinforcement for areas in which the registrar was doing well, as well as the registrar's developing identity as a GP. Registrars expected this of the observer, and were open and receptive to feedback.

Debriefing discussions were predominantly observer-led from beginning to end. Observers took the lead role in terms of steering the discussion (whilst also consulting with the registrar about their preferences), and also the amount of talk time. A shared assumption of the supervisor/ME's expertise and seniority was evidence in all but one pair. In the latter instance, the observer did not take a strong lead and positioned himself as a learner. The resulting discussion felt somewhat 'clunky', and the pair alternated in who was leading the discussion. An excerpt from their interaction is presented below:

> *Supervisor: "Clearly with that last consultation with the breastfeeding and this pregnancy one you appear to me to be very competent. I was learning things from you rather than…"*
>
> *Registrar: "Because you've got Person 2 at your clinic, don't you?"*
>
> *Supervisor: "Yeah."*
>
> *Registrar: "Yeah, so she's just amazing…"*

Positions adopted by observer and registrar

The participating observers were largely positioned (by themselves and the registrars) as the expert and mentor within the encounter. Each observer also positioned themselves as the champion; offering positive reinforcement and encouragement to the registrar as a developing GP. There was considerable generosity offered by supervisors and MEs in both regards; to create a useful educational experience for the registrar through specific constructive feedback and positive reinforcement. Observers positioned themselves in various ways to enact this; as the wise mentor, the maternal mentor, the experienced colleague, and as the learner themselves. Observers were also aligned as "diagnosticians of performance" (Molloy & Bearman, 2019) through their assessment of the registrar.

Registrars were positioned as both the developing GP (learner) and the competent GP. Both parties entered the encounter with the awareness that the registrar was there as a developing GP to learn from the valuable expertise of the observer. On the whole, the participating registrars enacted this role through enthusiastic receptiveness to feedback and advice. However, registrars were also positioned to demonstrate their capacity as a safe and competent GP to the observer and their RTO. These conflicting roles have the potential to interfere with the integrity of the event as an educational and assessment tool.

Key interview findings

Through the process of thematic analysis of the interview transcripts, several themes with accompanying sub-themes have emerged. The six main themes related to: the purpose of consultation observation; process, structure and observer roles; the benefits; the challenges; suggested work-arounds and ingenuity to overcome challenges; and the context within which consultation observation took place. These main themes are listed below with their subthemes.

An exemplar or two is provided to support each of the sub-themes. Of note, participating supervisors are identified in the results by the codes S1-S6, medical educators ME1-ME4, and registrars R1-9.

Please note that due to participant illness, there is still one registrar interview to complete at the time of the submission of this report, and therefore, inclusion of this respondent's contributions to the research cannot be included at this stage.

Purpose of consultation observation

When asked about the purpose of consultation observation, while most agreed to the various points of developing consulting and communication skills, and developing clinical skills, some teased apart the levels of importance in purpose. In particular, two participants – a supervisor and a registrar – identified a primary purpose that sits above both consulting and clinical skills, viz. the essential need for ensuring registrar safety.

Purpose 1 – registrar and patient safety

As part of the interview, S1 identified the primary importance of consultation observation as a means by which to check registrar safety as the primary purpose, over clinical or consulting skills. The GP supervisor elaborated: *"There's the question of, is the registrar safe? If there's something going on in that consulting room that wouldn't be identified if there wasn't someone sitting there, watching it. That's really the only opportunity we have to directly observe their consultations at the moment unless the supervisor does that themselves. And I think [that] once you get past the first barrier, if it's a disaster or if it's going alright, then it's a chance to really delve into either their clinical or consulting skills, depending on where they need some assistance."*

Personal safety, not just patient safety but also registrar safety, is a topic of importance in GP training conversations around Australia. It is not simply physical safety that is of concern, but also the registrar's psychological safety. As R2 reflected: *"I think it [consultation observation] can also be a source of psychological sort of support if you're not getting a great deal of mentorship from a clinic as has been the case for me in my previous year. Actually, the role of the medical educator observing and giving feedback takes on a huge role because it might be the only mentorship that I've had throughout the year. Which it really was."*

Purpose 2 – developing consulting skills including communication skills

The secondary purpose of consultation observation from the perspective of S1, was seen to be the registrar's development of consulting skills, including 'soft' skills like communication. Indeed, the critical importance of the development of communication, observation, and listening was emphasised by all respondents. As stated by ME1: *"I thought it was going to be largely clinical and picking up on things and giving kind of clinical teaching, but I actually find that the majority of what I do is to do with communication skills. And I think that that's something quite unique about that direct observation thing, that I think that people can look up clinical guidelines or the answer to a clinical question or get that from a supervisor kind of any time. But I think registrars aren't directly observed particularly much and I think that's quite a unique thing about the ECTV. And that's quite interesting because, quite often you know, you reflect back about how a patient looked or a cue that might have been missed or perhaps a way they reacted to something the registrar has said. And I think without that direct observation of that, it's something that people just don't notice."*

Viewing the observer and observee pair as a case of analysis in their own right, it is interesting that the registrar of ME1, the observee, reported to having a similar perspective on the purpose of consultation observation. When questioned, R2 replied: *"I think the main purpose is to have*

someone who's actually quite devoted to observing you and watching the development of your consultation skill and encompasses both the emotional and softer interaction skills...that's needed for doctor-patient interaction."

Purpose 3 – developing clinical skills

The other obvious purpose of consultation observation related to further development of clinical skills and knowledge. Reflecting this level of prioritisation, ME2 stated: *"And then, of course," the clinical content: are they clued up on what they need to think of, working diagnosis, important things they must not miss, do they ask the right questions?"* The 'and then, of course' comment reflects that while important, there are additional purposes and priorities. This is similarly reflected by S1: *"[I]f you get past the 'is everything safe?',* ***then*** *it's a chance to really delve into their either clinical or consulting skills …".* Thus the imperative is safety first.

A number of the registrars considered the purpose of consultation observation in assessing their clinical knowledge and skills. R5's comment is an example: "*…a lot of the core of observing these sorts of consultations, I guess on one aspect is…assessing clinical knowledge*."

In addition, and interestingly, it was also noted that this purpose of developing clinical skills was a two-way affair. The development of clinical skills was seen by some to be to the benefit of the observer, and not solely the registrar, as S2 and S3 identify. S2 described this as: *"Sometimes I find that, well especially in this case, R4 has done training that I haven't done. It's a practice that focuses a lot on antenatal care and obstetrics which is not my area at all. So often I end up learning some clinical things."* Similarly, for S3: *"[T]he visiting supervisors or medical educators enjoy it as well because there's a bit of a two-way learning in the whole thing, you do pickup things when you do the visits."* Therefore, consultation observation sessions can be of great personal benefit to the observer's own learning of up-to-date clinical knowledge.

Structure, process and observer roles

Some supervisors and MEs gave accounts of the structure of the observation of the consultation session, the process that they work through, how they envisaged their roles, and pathways to becoming an observer.

Structure

There were differences in the way that consultation observation was structured. Most were in a half-day block. A smaller number were shorter sessions, and in instances where the supervisor as observer worked at the same practice as the registrar, the session took place within teaching time. S5 commented of the half-day structure: *"We teed it up as a half day and one session and R8 at the moment is still having thirty-minute appointments, so it went for a bit over three hours. We did it in the morning."* Conversely, S4 spoke about a shorter session: "*So, it [the consultation observation session] was part of the hour or so I spend with R7 in formal teaching every week and what we did as part of that session, R7 actually had a patient booked in as part of that educational hour or so that we have and so I just did an assessment of one patient consultation that R7 had… and then we talked about the consultation, which was recorded. So the whole process was over half an hour..."*

There was also variance in when the debriefing sessions occurred in the structure of the consultation observation. For some, the debrief discussion took place between each patient as well as at the end of the session. As R3 explained: *"[S]o we had half an hour for each patient, 15 minutes for the consultation and then 15 minutes for discussion. And sometimes it did…, go a little bit over time with some of the consultations. But I thought the time afterwards to discuss them all was quite good."* Similarly, representing one of the observers who debriefed between patients and

at the end, ME4 stated that: *"I almost always debrief between patients, and most of them I would have debriefed between, I might have skipped one here and then for a while to help [the registrar] keep to time, but generally speaking I'd debrief between patients. It's more fresh in their mind and it also gives the opportunity to say 'I think you actually need to do x, y or z'. I think that's actually important you need to deal with that at their next review or, you know get that really, also get that feeling of, that direct feeling that you get after someone has left the room about whether you've dealt with that well or not because there are some times when you sort of think 'Oh that just didn't go well'. It's good to get that perspective right there and then."*

Alternatively, some opted to debrief only at the very end of the consultation session. As R8 explained: *"So it was a full morning session where I would pretty much practice as I would have had my supervisor not been in the room. So I would only turn around and speak to my supervisor if I was going to do that anyway by a phone call on any other day. There was no discussion of patients in between patients. There was no feedback about my performance until the very end when we had a formal debrief."*

Some observers were quite structured in the way that the consultation observation session was conducted, especially in relation to factoring debrief time during and after. A visiting supervisor, S6, discussed the mix of consultations and debrief sessions that she has found works best for both the observer and observee over years of experience: *"What parts work well, I think if it's a well-planned [consultation observation session], for instance if the patients are actually booked correctly; we've timed for debriefing, that works well. Obviously if there're not enough patients booked in it doesn't work, and if a registrar is over-booked then of course it doesn't work well either. So I think, as in any teaching exercise, planning is the key. So if I know that there's a [consultation observation session], or if I do an external clinical teacher visit, I get my practice manager to ring them up and say this is the way I want it, 'I only want two patients booked and then I want fifteen minutes, and then two patients booked, and then fifteen minutes, two patients booked, fifteen minutes …'. That's how it is, and I want fifteen minutes before any patients are booked, so I can talk to the registrar a little bit about their experience and how they're settling in. I think the most important thing is planning."*

As to the leading of the debriefing session, again there was variation. In leading and taking ownership of the debriefing discussion, some saw it as a two-way process, whereas others saw it as more observer-led. Reflective of a two-way process, S2 reflected that: "*I guess I was playing more of a leading role in that discussion than R4 was but it was pretty much a two-way discussion."* In contrast, other observers felt they directed the debriefing session, such as ME4's confession: *"I think I did, I think I can't help but talk. Yes."*

Process

Some observers, such as S4, had a detailed process they would engage in when observing a registrar: *"I guess as part of the consultation, I look at the very start of the consultation and so the way the registrar goes out to the waiting room and calls the patient in, I think that's an important start to a consultation. I look at the way the room is set up for the consultation as well and so the actual room set-up, how the computers are set, how the chairs are set up, the desk and so on. I also look at non-verbal cues that the registrar may be giving to the patient and also the patient may be giving to the registrar and whether the registrar is taking note of those non-verbal cues as well too. And then I look at various specific phases of the consultation. The history-taking, the examination, the discussion with the patient about the results of the history and examination, and the treatment plan that's organised for the patient, including investigations that may need to be done and developing a management plan with a patient and then any follow-up which may be required. I also*

then look at specifically the preventative health part of the consultation. The things that the patient may not raise but things that may need to be looked at."

Helping registrars move beyond a purely clinical focus

One of the benefits of the consultation observation process was seen to be helping registrars move beyond the clinical focus to the larger consideration of consulting and communication skills. To quote ME1: *"I think lots of registrars are and lots are receptive but I have a few who are just very, very focussed on the clinical stuff and want to know what they're doing wrong or right in terms of their clinical interactions with patients and I can tell aren't particularly interested in the other stuff I'd like to talk about. So I think that it doesn't work so well in that scenario. [INTERVIEWER: Mhmm. Is there anything...is there anything that you think can be done to facilitate, I mean because these soft skills are so important?] Yeah. Well, absolutely. So, in that scenario, I'll often say, well, you know, because I think a lot of that communication stuff is about opening things up a bit rather than going down one narrow path and so sometimes I'll try and role-play things or what if or what if you had tried it this way or how else you could have, you know, and where might you have got then. And, you know, I think you can sometimes get people with a little bit of a lightbulb, but not necessarily*."

Observer roles perceptions

Observers reported to having various perceptions of their role when undertaking consultation observation. These ranged from mentor through to educator, feedback provider, purely observer, assessor, or a mix of roles. S2 saw the role of observer as being one of a mentor: "*Well I think the role that would most closely fit how I approach it would be a mentor. That I'm going in there to help and provide some guidance. When I see something that I think they can improve, then I'll let them know."* Others saw the role of observer to be a teacher or educator. As S4 commented: *"I think the consultation we have with our patients is the most important thing we do in general practice and it's really important that we as doctors do it well. And I think it's a really critical tool that we have available to us. And if the supervisor, sorry, the person that's doing the assessment does it well, I think it's a really important teaching tool that can help the registrar as well."* Similarly, ME4 said that *"It's an opportunity for education which I'd probably do that more than anything else in the direct observation. And it's an opportunity for encouragement or just correcting, and I guess that's an education thing too, of the registrar. It helps them to get some idea of their standard as well."*

Others focussed on the assessment component, describing their role in that light, as did S3: *"I guess it's fairly broad. I mean the obvious initial thing is the assessment process that's required, which is probably the least important thing but maybe a bit of what it's driven. I think as a mechanism of assessing registrars and giving them some feedback on how to develop as quality GPs. That's not worded exactly right but reviewing their consultation style, clinical reasoning, relationship with patients, communication, and all the different aspects of a consultation that are important and being able to present to them what you're seeing, what you're thinking about that and giving them some advice or reflection on how they might improve in the future. All the registrars are in a stage, various stages of developing their own style and developing their clinical reasoning in general practice and the main thing is just to aid them along the way. There are some registrars that have specific issues and I think one of the useful things if they're flagged early on, then the visit can be directed at those specific issues. You know, things like communication or clinical reasoning, body language, relating to patients, whatever, if you know about those things you can specifically address those."*

Providing feedback following observation was another key role expressed. As S4, an observer in the same clinic as the registrar, noted: *"So again here in my own practice I really do see my role in that particular educational activity as one of observation and later feedback. I am happy to provide clinical advice if they need it as I would always do as the supervisor, but mostly it is an observing and feedback role for that particular activity."*

Finally, there was the perception of the role as being a mixture of things, of wearing multiple 'hats' at the same time, and dependent upon location of the observation, such as what S4 summarised: *"Certainly when I'm out assessing or looking at registrars and other practices, I am there as an assessor rather than a mentor. When I'm doing it with my own registrars, I'm really there as an educator and a mentor rather than an assessor, a formal assessor."* This also reflects the variance relating to the conducting of observations externally or within the registrar's practice.

Providing verbal feedback

The importance of providing verbal feedback in a timely manner was a key point of reflection for most of the research participants – observee and observer – alike. All the registrars in this research project were receptive to feedback, stating also that the feedback that they received matched their self-perceptions, though some registrars were more self-critical than the feedback they received.

R8 explained that the feedback also mirrored what was being placed into the formal written feedback to the RTO, and how the observer went to pains to explain this so that there would be no surprises: "*My supervisor went through pretty much the assessment he was putting into the, because he was physically doing the assessment on the online [RTO] system as we were seeing each patient. So he went through his feedback and what he had put in to that for each individual patient. So started with a summary of each patient, why they presented, a background of their issues, and then a kind of summary of my approach to the patient. Subsequent to that he would then give me feedback on certain areas that he had noticed, either that I had done particularly well or needed a bit of improvement. And it was really at those times that he invited my input as well. [The feedback] was quite mixed [in focus], but I would say professionalism and communication, if you consider that to be under the umbrella of [consultation] skills, then I think that was focussed on quite heavily in our feedback..I think in some ways I could anticipate what some of that feedback would be, because I had noticed that some of the things that he had pointed out was some of the things that I had noticed that I had done as well. So it was something that I was gonna pull myself up on as well. I was surprised that he had noticed some of the things that I had hoped that he wouldn't notice."*

ME4 noted the challenges of providing feedback to registrars who would not accept it. She was at pains to explain that this was in no way related to the registrar that she was observing, rather that the question prompted a memory when providing feedback did not go so well. ME4 said: *"This was from a few years back though and just lots of people had found the same issue with the registrar in question and really I just went through it in the same way as I would normally go about it and maybe I don't have very good strategies for dealing with people that are very closed off, I don't know. But, you know they can take on board what they take on board. I know other people that are a little bit more direct with people that have that kind of non-receptive attitude and I think that's a huge flag if you've got somebody that's not receptive to feedback at all that they're sort of a vulnerable person in terms of how they're performing in general practice I think."*

Providing written feedback

The move from paper to online feedback was noted by a number of observers. S6 noted of the transition from paper to online: *"before they were, [we] were doing it on paper. So, that was terrible*

and that was very, very backward and they gave me too-small squares to write stuff in. But no, I think the online format is actually much more useful, I can write and do the feedback all in a structured session. It's been much, much more streamlined since we've been able to do it that way."

S3 expanded, highlighting some of the trickiness about having to complete online that others mentioned: *"I mean the forms are pretty straight forward and easy to fill in that we have, I don't know how standardised they are across the country. But as a brief comments about, and the nice thing is it's not the one to five pick a number type of box, there's a written comment as to how they're going with the communication, clinical skills, education, all those sorts of different boxes to fill in, which are fairly easy and quick to fill in. The tricky thing with them at the moment is they're an online thing. So in the past they were handwritten and then faxed them off or post them off. Now they get done on the computer and there's just that trickiness of do you have a laptop in the room or an iPad in the room or do you write them down and then type them onto the computer. There's different approaches, nothing's perfect. But putting them on the computer is good because it's relatively quick and easy and legible and all that so that tends to work pretty well. Yeah."* It was interesting that this research participant mentioned the different approaches to completing the written feedback for those being observed and the RTO, but also queried whether the online tool was standardised across Australia.

As to the process of making notes for the purposes of assessment, variations were seen. One approach was 'assess as you go' which could be either by making notes on a notepad or typing onto a document or straight into the respective RTO's online feedback form. S4 explained: *"I usually do it as I go. I have a laptop with a fairly quiet keyboard and I don't want to miss, especially key phrases I like to be able to quote back to my …my registrar… verbatim any really interesting moments so I will do it as we go, and there is so much that happens during the consult I don't think I would trust myself to remember it all at the end, but that's my usual practice."* Others noted the potential for the process of typing to be a potential disruptor in the consultation observation process. S2 made a note to their registrar to advise them if the notetaking (in their case, on laptop) was disrupting the consultation. S2 said: "*I set, the rule is, sorry, not the rule, the guideline, or guidance before we start the consultation is, 'I will be writing on the computer. At any stage when you feel this would interrupt, please look at me, give me a sign, or just clearly say, 'Please stop'. If you feel I need to step out, and I don't pick up on that, please let me know'."*

Some observers used tools (whether in online form via laptop, or their notes) to hide/shield from the patient's enquiring gaze to affirm that the registrar was the primary point of contact, such as in the case of ME3: *"I don't know whether it's a good thing or not, but now that we're doing these [assessments] online, I sit there with my laptop, and I think that actually forms a little barrier anyway which is a good thing, and I explain to patients that I'm just typing some notes about R5 rather than about the consult and that's all good they think that's funny, so you know."*

Pathway to becoming a supervisor for consultation observation

Some of the participants in the observer category discussed their pathway into being either a supervisor or ME. S1 described: *"I've done quite a bit of, I sort of came into backwards in that I started as a medical educator back in 2009 but because of the way we run the ECTVs down here, I didn't start doing routine ECTVs until two or three years ago. However, in my role as a medical educator, I was often doing the troubleshooting ECTV so I started off doing the hard ones and then came back to doing the easy ones later on in the piece. I don't know that's the traditional path for doing this but that's all right."*

Benefits of consultation observation

A number of benefits to the opportunity for consultation observation were raised.

Occurs in real time

The immediacy of consultation observation was noted to be a strength. As S6 stated, the true benefit in being the direct and in real time: "*I think direct observation gives that immediate and direct formative feedback."* R4 spoke to the value of immediate feedback in consultation observation from the registrar's perspective: *"I think I have like a, probably a bit of a lack of confidence and so I think it's particularly useful. To sort of be able to have some definite feedback, particularly early in your training that you can sort of look at and reassure yourself because it is, going into general practice is obviously very different from any of the work that you do as a hospital resident or a HMO registrar of whatever, so, yeah, I found that quite reassuring. To know that I was doing an okay job. I mean, we did do some wave consulting and that sort of thing but this is better than that, so, I think, yeah, my own personal anxieties would make it particularly useful."*

Occurs in a real context

For other registrars as observees in the process, the value was that consultation observation took place in their natural working environment. S8 replied: *"I felt like I was being observed in my natural habitat, practicing how I normally would, as naturally as possible, trying as best as I can not to let the knowledge of someone else in the room assessing me influence how I would practice. So I think its's to get very direct feedback on my performance, which up until the observation I was getting more indirectly through my supervisor's observations on the types of questions I was asking when I was asking for advice."*

Relationship building and strengthening

The impact of positive relationship-building from the perspective of the observee and the observed can be one of the benefits. As R2 commented: *"I always find this particular medical educator…really encouraging and positive and really an advocate for the registrars who she's working with. I feel really lucky that I've been allocated to her because definitely the personality of the particular medical educator is going to have a big impact on your ability to seek mentorship and emotional support. But also not to lose too much confidence in a process that can sometimes feel like you're being observed and judged and watched. Yeah."* Looking at the response to the same interview question, the ME in the consultation pair reflected the positive relationship building, and the strength that because she had been her observer over an extended period, that it had grown over time. ME1 explained: *"I have certainly found R2 extremely keen to take on feedback and very receptive and very keen to learn as well. So that makes everything very easy because she is like a sponge and is very keen to be told, you know, to have some help in improving so that works really well with her…I think certainly when I first saw her having been aware that she was a relatively timid, I suppose, personality type and that therefore my feedback perhaps was a bit tailored to not scare her off but certainly over a period of time, that's become much less of an issue and so, yeah, I think she has developed confidence but also her robustness to feedback as well, which is good."*

The influence of a positive relationship between observer and observee has been already noted, but is worth repeating again. Looking at an observer-observee pair as its own unit of analysis was R7 and S4. R7 summarised the influence of a positive relationship that they had with their supervisor on the process: *"It makes it, makes it a bit more comfortable and S4's a very non-judgemental person, even if I completely, you know, stuffed up the whole consultation, had no idea what I was doing, he still wouldn't make me feel like an idiot. He's a very gentle person and, you know, I feel very confident having him sitting in on consultations and I can really just be myself and*

do my own thing. I don't feel like I've got, you know, a pair of eyes staring at me straight into the back of my head and scrutinising my every move. So yeah, I feel very relaxed when he's in the room with me and I think that just stems from us having a reasonably good relationship. We get on really well." In his interview, S4 commented of R7 when asked if there was ever any difficulty in providing feedback: "*Not with R7 because she's an excellent registrar. I guess when registrars perform well, it's a lot easier to provide feedback because it's usually positive*." S4 continued: *"I would hope if anything it [my feedback] has a positive impact on R7 and her future training and also on our relationship too as well too. Yeah."* Indeed, S4 expounded: *"She's been with us for twelve months, so she's due to change over. She's asked for a special dispensation to stay with us for another six months…I think she is hoping and we are hoping in the long term she might come back to our practice but I guess that's to be seen in the future."*

Similarly, R6 noted: *"[The observer is] very approachable, which I think is important because these sessions can be really intimidating and I can imagine if you had someone who, you know, just generally makes you feel uncomfortable, you probably wouldn't perform as well as you would otherwise…So, she was very relaxed and approachable and non-judgemental and very lovely and I didn't have any, you know, issues with saying what I thought or anything. I didn't even—I wasn't even really worried that she was there, which makes a big difference because, yeah, in exams and other times when you're under pressure, you definitely don't do as well as you probably would on a day where you're comfortable with someone, so, I think for her, I think she's very good at that job. I can imagine that there's a lot of different personalities doing it and that maybe would change things, but yes, definitely, her and I was fine."* Building of positive relationships can assist not only in learning the craft of being a GP, but can also build a future workforce.

Benefits of external visitors conducting the consultation observation session

Some observers highlighted the benefits of having someone external to conduct the consultation observation sessions. S3 commented that: *"It…gives the registrar a bit of an opportunity to ask questions about general practice, general approach from someone who's not their supervisor, which is an advantage to get a different perspective and also enable the registrar's to share any problems they might be having within the practice or personally that they don't feel like they can share with their supervisor."* Clearly there is the need for an independent contact beyond the practice for when things might not be going as well as hoped.

Some of the registrars also highlighted the benefits of having someone external to conduct the consultation observation sessions. In the extended quote from R2, a range of details were given including the affirmation of the psychological support that can be provided in their GP training at the hands of a skilled ME. R2 commented: *"I think the main purpose is to have someone who's actually quite devoted to observing you and watching the development of your consultation skill and encompasses both the emotional and softer interaction skills as well the medical knowledge that's needed for doctor-patient interaction. I think it can also be a source of psychological sort of support if you're not getting a great deal of mentorship from a clinic as has been the case for me in my previous year. Actually, the role of the medical educator observing and giving feedback takes on a huge role because it might be the only mentorship that I've had throughout the year. Which it really was."*

Benefits of the supervisor, ME and registrar team to shape the registrar's learning

One supervisor, ME and registrar trio participated in the research study, offering a unique insight into the roles of each other, a case of analysis in its own right. The identifiers for this trio are R5, S3 and ME3. In this case, the ME performed the consultation observation session, and the registrar's supervisor conducted the final debrief session. To explain this, R5 said: *"It was a bit of a combination because it was sort of depending on what my medical educator, because I think she had*

spoken to my supervisor after the observation was done and mentioned a couple of points. There was a sort of discussion surrounding what my medical educator told my supervisor but then there was also a bit of a discussion about how the observation went and how I felt about the whole process of the supervisor observation itself." ME3 added to the conversation: *"Well, I did a direct observation for this registrar and, as part of that I usually speak to the supervisor. I think I had informally spoken to him or he'd spoken to me at another occasion just flagging this particular registrar. And we had a discussion, I presume it was after I had finished the direct observation with him. Yes, so this had been a conversation with the supervisor." The ME continued: "Yeah, I mean, the supervisor gets a completely different and probably better, actually, view of what's going on with a registrar in terms of their consulting. Both from feedback from patients and things and also from access to their records and the questions that they're asking or not asking and patients coming in, you know, you just get a whole different slice of what's going on. And, so, and interestingly different supervisors in the one practice can have completely different views about a registrar as well. So, you know it's, it's almost one of those things of the more kind of opinions and views you get, I think, the better to be able to get any sort of really clear picture. The supervisor in question is actually a really conscientious supervisor and so he had some, he had some feedback that I didn't actually really observe when I was doing that direct observation. Although I had, when I had done it and GPT1, there were some issues that came up then and that I kind of went: 'Oh', about and, but they didn't when I came back for this particular visit."* S3 thought the addition of an ME to the team worked well. As S3 summarised: *"If it works well it's good. So when the medical educator does their visit, or when a supervisor does a visit to another practice, the ideal outcome is they have a meeting with the GP registrar's supervisor immediately afterwards to discuss their observations, any suggestions they have, what their analysis was, that sort of thing. That's the ideal situation, which happened with this. This sort of was picture perfect. So the visit was done, ME3 had the discussion with R5 and then we couldn't meet face-to-face immediately so we had a phone discussion in regard to him."* Clearly MEs and supervisors working together can assist each other in their particular roles in the GP registrar's training.

Registrars being seen as a 'real GP' giving a 'proper' consultation

Being seen as giving the patient a 'proper' consultation – being a 'real GP' – was a key element in the registrars' responses, as was learning and taking on feedback. As R2 expressed: "*I think my role is, well, first of all to make the patients feel like they're having a proper consultation. To take on any feedback a medical educator has provided. And to be open and honest about my experience so the training programme can be developed to reflect what's being communicated."*

However, a concern raised by one new registrar was that in having a visiting observer in the room (in this case a supervisor), it might lead to patient doubts about their abilities and competence. S8 expanded: *"… I did worry that the fact I had a supervisor in the room made my pateints feel less confident in my ability, because I wasn't sure perhaps what the patient had been told about why my supervisor was in the room. So the thought that constantly crossed my mind was 'Do my patients think that I'm being supervised because I'm not competent or that I need to be certified competent?'"*

Challenges

A number of challenges were identified in the consultation observation process.

On stepping in for case management

There was variance on the issue of 'stepping in'. As S5, a visiting supervisor summarised: *"I feel a bit uncomfortable providing management advice to patients who are not in my practice. I don't have a provider number for that practice; it just doesn't sit well with me. I will often do it a little bit if it is something straightforward but as much as possible, I try to encourage registrars on external visits to*

follow their normal procedure and speak to their own supervisor. So, yeah. That is the major difference because when I am supervising here with my own registrars then I am perfectly happy to be that first port of call if they have a question. I guess having said that, just because I am in the room with them I don't encourage them to defer to me, I really do want them to do what they would normally do (come up with their own diagnosis and management plan) and then only discuss it with me if they would otherwise have had any questions about it."

On this topic, one medical educator – ME1 – commented: *"My only role to step in and be obvious if I see something disastrous about to happen or something really should be done but they don't have the opportunity to change after the fact, then I'll step in to have 'Maybe, we should think about this.' But other than that, I don't really want to be there involved in the consultation…I think if there's something minor that needs to be changed whether it's change of a blood test or something like that, even that I wouldn't necessarily step in at the time. I would probably after consultation, I would say 'I know that you just did this but I'm wondering whether you might also want to do this and maybe it would be worth giving the patient a call back and sort that out.' I'd rather, I think it is very undermining if the registrar is trying to run their consultation and you step in 'No, no, no. This is how you should do it.' That patient is then going to lose a lot of faith in that registrar as the expert in the room."* Consideration to building up the registrar's confidence was therefore viewed as keenly as building up their competence.

One of the observees, R7, recounted their supervisor's willingness to step in: "*There was one very early in the year when I was just starting in general practice and had basically no experience with this particular present patient. And it was a situation where I would have stepped out and asked my supervisor anyway. So that was one of the provisos, I said 'S4 is here, he's watching the consultation but I'm just going to run this like I normally would.' And so, there was a particular question that came up that I couldn't answer, and I would normally have deferred to him anyway, so that was the only point where he's contributed to the consultation."* However, some supervisors, irrespective of whether visiting or in their own practice, still try to let the registrar know that patient management is their professional responsibility in the session.

Observer's familiarity with the patients

Another interesting facet identified in conducting the interview was in the issue of patients knowing the observer, and the dynamic that that has on the consultation observation. This is where the challenges of knowing – have familiarity with – the clinic's patient population. This was raised in several of the interviews from observers who were either rurally-based, living in communities, or observers who were part of the same clinic as the registrar, as the quotes from S5 and S6 both demonstrate. S5 commented: *"One of the patients actually had previously been scheduled to see me for a newborn check but R8 hadn't done one before so it was a little bit unusual for that particular consultation, in that the patient was told they would be seeing me and that I would be performing the newborn check. R8 wanted to see one before she did one herself. Apart from that it was just a standard morning of consulting. Thankfully I didn't actually know all of the patients particularly well. It was really only a couple of them that I needed to be clear of."*

S6 had a slightly more challenging experience: *"So, the last patient…it did derail because she [the patient] had an appointment to see me the next week to discuss some other problem further, so it did actually derail that whole consult a little bit, because she kept on deferring to me; she kept on talking to me instead of to R9. And even though I tried to deflect back to R9, because … I was looking at her, and saying, 'R9, this is your consult here, what do you think?' It was really difficult for her to get that control back, because of the dynamic. And that is the problem of sitting in in a practice where I am a principal and I've been here for nearly twenty years, and people come in to see the registrar because*

they can't get in to see me and then they look and I'm there! It's like, 'Oh, yes, well I want to talk to you, but I can't get into you for like three weeks!' So that was difficult. But of course, never happens in external clinical teacher visits, except the opposite happens—when I do an external clinical teacher visit, sometimes I see patients that I know through other social circles or professional circles, and then it's very awkward; then I have to sit out anyway."

Time challenges from the perspective of the observer

The time to conduct the observation was not raised as an issue per se. It was more the associated administrative requirements both prior to and after the event. As ME1 stated, the process of CO *"can be tricky at times. It's definitely, there are different challenges to it, I guess, than my clinical role in that it kind of spills out a little bit into my everyday life a bit more than my clinical work does, just in terms of report writing and telephone calls and that kind of thing. Yeah, so, yeah. But it's, you know, it's good to have a different challenge I guess."*

On disparity of observer feedback and observe self-perceptions

Whereas most observees in this research concurred with the feedback that they received, the challenges of incongruence between the feedback given by the observer, and the self-perceptions of the observee, were also noted by some reflecting on previous experiences. S4 commented: *"I guess when registrars perform well, it's a lot easier to provide feedback because it's usually positive. A couple of situations I've been in where you have to provide negative feedback, that is much more difficult. Especially, I can think of two situations where the registrar thought that they performed reasonably well but they really weren't performing and that is, that does become a difficult situation."*

Registrar's personal authority

One registrar spoke of a potential challenge with patients while being observed, is the lack of their own sense of authority in the situation. This was viewed through a gender and age lens. R4 explained: *"I think sometimes with the patient, when you have an older male GP supervising you and you're a young female registrar, they do sometimes lean towards the older male GP [observer]. But that's, you know, bias within the patient. That's a bit hard to do anything about, so. And I have introduced him as, you know, like a supervisory GP so that's reasonable for the patients too, I think."*

Registrar discomfort at being observed

Many of the registrars discussed the discomfort of being observed, though the importance of it being helpful to their training was simultaneously acknowledged. Some registrars spoke of the awkwardness of being observed. R5 noted: *"I guess having not having that much experience with direct observations in the past, I don't really have much to compare it to but I guess from my perspective, whilst it was a little awkward because I'd never really done this before, I probably had an advantage in that my medical educator was quite supportive and made me feel quite comfortable all throughout anyway and the whole process was reasonably straight forward and I can't really pin down any specific difficulties that I encountered during it."*

Personal factors, such as being shy or introverted, were also mentioned as having potential impact on the process. One registrar spoke about this in relation to others, even though it was not her own personal experience. R7 related: *"I can imagine, you know, if you had a more introverted, shy person who wasn't used to being watched or performing or doing anything like that it would perhaps, the anxiety of being watched would perhaps overshadow there, what they normally would be like. I don't think it would be perhaps an accurate reflection for them because if their performance anxiety became more of an issue for them, they, you know, people buckle under pressure. So, they*

may not conduct the consultation exactly as they would have if the supervisor wasn't sitting there watching so I certainly can see that there is some differences perhaps in how different personalities would respond to the process." Considerations to such individual differences need to be factored into the process, and considerations for additional support and reassurance where needed. Again this may come back to adequate registrar preparation for the event, and on the other side, robust observer training.

Variability between the practices being visited

There were some issues that were described between the different practices. ME1: *"In terms of what works well and what doesn't work well, hmm. I mean, generally, it seems to flow quite well. And sometimes that can be different at different clinics, particularly when you have newer clinics or clinics that aren't necessarily so used to the setup but certainly where Person 1 was working, they're very used to having registrars and, you know, in terms of the admin staff consenting the patients and letting them know and that kind of stuff, that all works very well there, and also the patients in general are very used to having registrars working there so the whole thing works quite smoothly. And again, because Person 1 and I have had a few sessions before, in terms of, you know, instructions and that kind of thing, we're quite, it wasn't particularly cumbersome so, yeah, it tends to work quite well. I'm trying to think of something that didn't work quite so well and I'm not sure I can think of anything."*

Dealing with negative feedback

Registrars anticipate being guided and shown areas where they can improve. However, the counter-effects of negative feedback was also brought up by some registrars when describing the experiences of some of their peers in contrast to their own experiences of consultation observation. For example, R2 retold the experiences of her peers, and the counter-effects that the negative feedback that they received, to express her relief of her own positive experiences, and how lucky she felt in her own situation: *"What I've heard from other registrars is that, at times, a sort of very critical medical educator has visited the registrar and they feel that they've been unfairly, sort of, provided with only negative criticism and it's actually really destroyed their confidence. I wouldn't change anything about my experience but I'm aware that it doesn't always work as well as this. [INTERVIEWER: So, [are you saying] that it's important that if negative feedback has to be given, it's also important to try and find some positive feedback to give to improve confidence?] I think that from what I heard in those cases, the negative criticism was probably not coming from a place of experience or knowledge. It was maybe more [laughs], it wasn't actually [pause] grounded in that experience or in evidence. It was maybe just a personal dislike of the registrar. I don't know…I wasn't present in those cases but I am aware that it's not always as streamlined and easy as it has been for me.* This warrants consideration, and perhaps additional training for the observer role is warranted.

The impact of cultural differences

The issues of cultural differences came up, both from the perspective of the registrar-patient interaction, and from the perspective of observer-observee. S4 commented on cultural issues in the registrar-patient interaction: *"I can think of one situation in particular where there was a junior registrar who was from another culture and he did an interview on a female patient and obviously there was significant cultural differences which meant that the [consultation] didn't go at all well."* From the perspective of cultural differences between observer and observee, S6 also suggested: **"***I think there could be cultural differences…I don't know. Again I'm sorry, I'm just generalising, but middle eastern males, being female and ethnically Chinese, I'm not sure whether that would be a problem, but I would have thought the main problem would come from … See as a student, or as the*

registrar, I would not want to be assessed by somebody who I didn't think had the qualifications to assess me, you know? So, I think that would be my main objection to being observed. So I need to know that that person is, not an experienced clinician, but they are either a medical educator or they have some sort of qualification or hold some sort of position that qualifies them to actually give me the feedback and, you know … in a way that I can receive it and act on it. I think that's the main thing that would be a barrier. But that could possibly be cultural barriers. It's possible."
How do I know how I am going with my teaching?

One supervisor wanted to extend consultation observation as a process so that supervisors and MEs received feedback on how they teach. This was interpreted as a cyclical process of reinforcement to assist them in refining their skills. S6 stated: *"Well … I think the problem also with the supervisor [role] is that we don't actually get feedback ourselves. So, I don't get feedback on my teaching; I have no idea. I have no idea whether my registrar thinks that they've improved or not. I can only feel that I've got more skills myself, so I feel more confident teaching. That's all I can say."*

Lack of instructions to registrars

For registrars, lack of instructions and training prior to undertaking consultation observation was raised as an issue. For example, R8 expounded on how the lack of receiving prior instructions and what the process would involve, other than a consultation observation session was going to be taking place: *"I certainly didn't get any additional information about what it would actually involve. To be honest, I was actually quite surprised that my supervisor sat there and did the assessment; putting in the patient details and his feedback into the online system as we were going. I had no idea that was going to happen. I knew he would be sitting in the room, but that part of it surprised me I didn't know that was going to happen."* This may have occurred due to an assumption that other stakeholders, such as the RTO, may have fully briefed the registrar on the process.

Lack of formal educational training of observers

The flip side of the coin as raised by one of the supervisors, was lack of training, including formal training, for the observer. As S6 explained after initially talking about the benefits of consultation observation as a tool: *"And that's the value of it, I think, because you can't see yourself, so you're really relying…on a trained supervisor, that this actually might mean, potentially with [consultation observation], I think lots of people do them that are actually not trained. So, they don't use teaching techniques that are formative, for instance. They may be a little bit more critical, or you know, everybody's got their own way of doing it but…So [consultation observation], I think the value of it lies in the quality of the supervisor performing the [observation]."*

Challenges with completing or accessing the online forms

For some – both observee and observer alike – the online forms and online feedback posed varying challenges. One visitor noted the challenges of providing feedback via the online form. S2 stated: "*The forms are a pain in the neck…But it's a real nuisance having, that you can't just have an offline form that would just be uploaded quickly. That would be my dream. That's just the sort of thing, like it takes so much time to write them up afterwards, although I already have them electronically. That's the one thing which is really a nuisance. But I do go through the online form with them initially, you know, fill out the registrar and the clinic's specific matters, together with them, actually."*

For one visiting supervisor, a system crash created major havoc, frustration and time delays. S2 related the tale: *"I really don't like [our online platform]. I started filling in her form and then halfway through I lost all the, I'd written all the cases in and then when I went to submit it something happened and anyway I lost it all so I had to start again. [INTERVIEWER: That sounds awful…a lot of*

work.] Well it was 10 or 15 minutes of typing. So there's that background feeling in me that [our online platform's] not, I don't like it. Anyway, so that's one thing…I don't know what, unless you type something out in Word and then cut and paste it but then that's most of the time is not going to be needed. Yep, not sure. [INTERVIEWER: So is anyone in the [RTO] aware…?] Yeah, I sent an email saying…You'll register my discontent. Which they noted."

For one registrar, accessing the online feedback to read the observer's feedback was just as problematic. R2 reflected: *"The only sort of criticism that I would like to make, most of this has all been very positive. I just think that the computer system is terrible. That I have to spend so long trying to access the ECTV written feedback and it's very random whether you're going to click on the right button at the right time. The computer program is called Pivotal and it's commonly sort of criticised for being so hard to navigate by both supervisors and registrars. Being able to access the written assessments, I think, is important because that's where the feedback lives and as useful as the whole process is, I think it would be nice if it could be backed up by a better way of accessing it."*

Overcoming challenges/ingenuity

Circumventing problems with online systems

While one observer commented on how pleased they were to move away from a paper-based assessment to an online system as the input boxes on the previous paper-based version were too small to write in, some participants – both observee and observer – discussed means to circumvent challenges found in using the respective online systems, such as creating their own template to mirror the online system which they could save along the way, and then copy and paste into the online portal when required. From the supervisor/ME perspective, this was to mirror the online assessment forms to overcome crashing. They could also work on this during the observation session. From the registrar's perspective, comments were made about building their own simple system that they could use to copy and paste into the online learning planner or, creating their own.

A beautiful example of this comes from R3 who explained: *"I mean I'm not exactly sure how beneficial the learning planner is. I mean I do use it and especially when I first started out I write out, you know these are the things I want to achieve and all of that but then what I've done now is I just have a OneNote kind of opened up on my laptop and then if anything comes up during my consultations I just write it on to that and then in the next night or two I can look at it. I think it's mainly because to go into the [official RTO] learning planner you have to login to the website and all of that, which just takes a bit more time. Especially when you're in a consultation and you're not actually logged into it and a topic comes up to mind. So I guess I do use a learning planner but it's just my own separate one that I have on my own WordPad as opposed to the actual formal one that's on the, I guess the useful part of it is before an ECTV visit, I do revisit it to kind of say, 'Oh, these are the things I initially setup to do and how am I progressing'. And it does give them a view of it as well."*

Similarly, S2 created his own Word version of the online form that he could complete during the visit, then copy and paste into the online form at the conclusion of, or following, the visit. S2 noted: *"I have to type it all in a word document on my computer. So, I do them almost instantaneously, cos I type them all up, and then I basically have my own little form, and then just cut and paste it into the online form."* As noted in the extended response by S2 in the previous section, he stated that his dream would be: "*that you can't just have an offline form that would just be uploaded quickly. That would be my dream."* This speaks to the possibility of RTOs having a pre-existing template that could be completed offline and uploaded.

In a similar vein, S4 made the suggestion of modifying the existing assessment rubric by extending the broader categories into the various components. They detailed: *"I think breaking down communication skills into things like language, use of body language as well as verbal language, the different ways in which you can establish rapport. There is no, thinking along clinical skills lines, there is nothing that directly addresses preventative health, just thinking off the top of my head now and I guess these are all sorts of things that you could try and sort of free-type. These are all—sorry I should explain the form itself has a lot of, you just have to tick satisfactory or unsatisfactory or I forget exactly what the scoring system is I think it is 3 or 4 options you can choose from in various domains, but there is no space underneath that to write regarding that particular domain. There is, further down at the end of the case, you can say things that the registrar did well and then areas for improvement and I guess that is where I would throw my comments in down there."*

Overcoming inconsistencies with RTO mandated training on conducting consultation observation

Solutions for the challenges of overcoming inconsistencies in approaches to structure and process when conducting consultation observation was addressed by S6. This supervisor suggested: *"I think training in how to do a [consultation observation] would be really good. In particular, talking about the planning, the structure, maybe mandating a particular way a [consultation observation] has to be done, because apparently, with [our RTO] anyway, culturally, historically, people just do it however they want to do it, so some people sit through a whole session and do it at the end, the feedback. There's no mandated structure, whereas I think having an organisation-wide protocol to say, '[consultation observation]s have to be this way, you have to do fifteen minutes, especially if it's an external clinical teacher visit, if it's an external clinical visit to somebody else's registrar, just to get the patient … the practice to schedule it in a particular way because you need to get to know the registrar first; what level they're at, how they're going, what their hours are, how much teaching they're getting, whether they're getting enough support, just that kind of initial orientation.' And then saying, 'Look, two patients, one break, two patients, one break, two patients, one break …' minimum of six patients for the [consultation observation] visit, the chat with the actual supervisor. So, mandating it that way, and perhaps funding training programs to talk about, 'Look, this is how you could do a [consultation observation] well.'*

Overcoming inconsistencies through formal training of observers as clinical educators

S6 discussed the importance and relevance of not just skill acquisition in conducting consultation observation, but also the importance of learning formal educational skills for the role of observer. This matches moves across the sectors for those in educational roles to acquire training in pedagogical theory and practices, such as those lecturing in the higher education sector. S6 explained: *"[M]y perspective is that I think supervisors of registrars ought to have a basic teaching degree anyway because we're teaching, and mainly that's what our job is. We are already clinicians, but we're not teachers until we develop some teaching skills. Otherwise, why have the teacher training, like why have trained teachers at schools for instance, you know? Just chuck any maths graduate there, and, but they can't teach. They can't teach; teaching is actually a skill. And particularly in the clinical situation, because it's so on-the-spot. It's on-the-spot, it's work-based, it's immediate, there's a lot of stuff around that teaching in the workplace that needs skill; that needs to be skilfully handled, like that feedback session in [consultation observation]. That's what I think. If not well handled, then the feedback's not taken on board, you don't get any learning, and you might get someone who is put off by it. So in our [RTO], what I've been banging on about is trying to get some of the supervisors to even do a graduate certificate in clinical education…[It] gives you the theory, the framework, the skills. It often talks about assimilation, it talks about again, how do you teach clinical skills? What are the current theories behind teaching a clinical skill? What are the barriers? So understanding all of those things, makes everybody a better teacher, I'm sure."*

The implications of context

The final theme relates to the issue of context – the arena in which the process of consultation observation in nestled. Much of the discussion centred around the physical space and layout. Other aspects discussed the practice culture, the supportiveness of the patient population, and how the RTO can be responsive.

Physical Space

The physical space in which consultation observation takes place was a large talking point, as noted.

Observee's familiarity with the consulting room layout

Familiarity of the consulting room's space was an issue for some registrars as they felt that unfamiliarity with the consulting room impacted their performance on the day. Reasons were given for a change of space, such as the practice wanting to allocate the registrar to a larger room for the CO session, in recognition that there would be an additional person (the observer) in the room. At other times, the practice rotated rooms and the registrar may not have consulted in that room prior to the observation visit.

R2 made two extended comments in relation to this: *"[S]ometimes the practice will put you, for example, in a different room to where you usually consult and that kind of thing can put you off because it can be, it sounds like it's not a big deal but you might be sitting in a different position. You're facing the patient in the opposite direction where you're not expecting someone to be observing you from the other side and the way your relation to the computer is affected and you don't know where all the paperwork is in that room. I think sometimes the clinic does that so that, they give a larger room to you on the day of the medical educator coming. One time when ME1 came, that actually unexpectedly put me off quite a bit."* They continued: "*For me, it was more the orientation of the room. Having to turn to my left to interact with the patient and then seeing an extra person on my left which was the medical educator. When there's already a small degree of emotional stress, like I'm being watched here and every move is being analysed, because it's true, that's part of the consultation, the body language is part of the consultation as well as what you're saying. So, when there's already one sort of source of stress in the consult, which is the presence of an extra person there to critique you, then the change in environment is an additional, sort of, stress to deal with and can result in, for example, you coming across as not very confident or hesitant and it's not actually reflecting the normal interaction."*

For R9, the larger room alternative was the practice's. As S6 explains to the question whether the consultation observation session was in a room that R9 has been used to practising in?
S6: "*No, we all rotate actually, so that was in a room that most of us rotate through, so it wasn't her room. But she's been in it before …"*

Sometimes practices and supervisors are cognisant of the impact that room changes may have on the registrar's performance. As S4 details: *"We try and, unfortunately, our registrars don't have their own room full-time but we try and limit their rooms to two or three rooms so that they get used to a couple of rooms. We also make sure in our, we've got twelve consulting rooms and we try and make sure that they're set up fairly similarly but every doctor has their own particular way that they want their room set up and so. But the way equipment is set up in the room is the same, so, when R7 goes from one room to another, she knows exactly where spare prescription paper is or where the Oroscope is or whatever. But R7 unfortunately doesn't have the opportunity to set the room up exactly how she wants it setup because she does move from one room to another, yeah."* Personal

agency in the consulting room set up and performance in consultation observation may be an interesting research angle to pursue in the future.

Room size

Not every practice is alike; some practices have small or larger rooms as their consultation spaces or a mixture of the two. This was reflected in the participant responses, and the overlay of the impact of perceptions of being in the way during the observation of the consultation. ME1 stated that: "*There was enough room in this room because sometimes that can be a bit of an issue. If you're kind of in the way wherever you are in the room, that can be a problem but this one was fine.*" Similarly, ME4 commented: "*Yeah from a, I guess, from a straight physical characteristics point of view, the small rooms are a problem but you can't do anything about that. That's the rooms. If there's a larger room where you can really get out of the way, That's quite uncomfortable where you're sitting there, sort of, arm in arm with the patient trying to not talk but the patient's looking across to you for is this right or, you know, what do you think? The more you can remove yourself from that, the easier it is to be a dispassionate party in the room."*

However, where some practices allocated a larger room for the purposes of a visit, there can be some downsides, as S6 noted: *"That room? That room we use [for consultation observation visits], well it has another role as a theatre, so actually looks a bit clinical and it looks like an out-patients' clinic room actually. But it's large enough, yeah; pretty good size."* The potential of larger rooms for consultation observation sessions may be beneficial in order to accommodate the registrar, patient, any support person(s), and the observer. However, the impact of registrar familiarity with the consulting room space, as noted above, is not to be underestimated.

Positioning of the chair configuration between observe, patient, and observed

An unexpected focus of attention in the interviews with the observers was about the positioning of themselves in the room in relation to the patient and the observee. Some preferred to sit where they could see the registrar. Others preferred to sit to see both registrar and patient. Others more the patient. However, these preferences are contingent upon the size and layout of the room and if these facets constrain them in their preferred placement or seating position. Some examples of these 'where do I sit if I have a choice?' consideration are as follows.

S3, for example, liked to see the registrar: "*So off in the corner, diagonally opposite where the registrar's sitting with the patient. Where the registrar's sitting, sort of diagonally opposite them, so the furthest away possible and away out of the patient's vision, that's where I tend to sit."*

Another visitor stated that: "*One factor is I think the visitor needs to be sitting in a spot where the patient is, you're not in the direct eye line of the patient so that they continue to focus only on the doctor, they're not looking over at the visitor. And in this situation it was, depending on which seat they sat in they might have been looking directly at me as well as the doctor or I would have been off to the side. That was the only factor that's, it kind of might be better to be sitting more behind their shoulder in a way. But no the room was adequate."* (S4)

Trying to be unobtrusive was a common discussion point from the perspective of the observers. As S6 commented: *"So I usually sit back, often with my knees crossed. I usually have a pen in my hand and some paper that I may be looking at writing on so I can be actually looking at something else rather than the patient. With just looking up from my pen and paper, so I am writing comments quite often as I sit there, so I just try and avoid the eye contact with the patient and just look up and see how things are going. But yes, I do position myself so that I try and keep out of the consultation as much as possible."*

S3 would have preferred to sit where the patient couldn't see him. However, he noted the room constraints over the opportunity to sit in a preferential space. S3 contemplated: *"You can't, none of the rooms are really structured such that you can sit where the patient can't see you. You can't sit behind them or right on the side of them so you're always sort of on the opposite wall but down in the corner is the best place."*

Similarly, the observees spoke about this details as well. R7, for example, said that: *"The supervisor chose where he was going to sit. And honestly in the room there was no other logical place for him to sit, because if he had sat anywhere else he would have been obstructing something like the examination table or the equipment or he would have been sitting directly next to the patient. Where he sat I think was the only possible place for him to sit."* In terms of the impact of contextual overlays to the experience of consultation observation, where the observer sits may impact the observation and this warrants further research interest.

Practice Culture and Patient Population

Practice Culture

On the whole, the benefits of a supportive teaching and learning environment from the perspective of the practice culture, was reiterated across all interviews. Reflective of this is R1's contributions: "I *think the practice culture is supporting learning and it's very educational. Even our supervisors themselves, I know they're very actively involved in teaching [with the RTO] and just being active. So I think the whole practice as a culture supports all of that."*

The alternative was expressed by R5: *"[I]f the organisation isn't there, like the practices put it down to say the registrar to get consent for patients and the practice isn't particularly proactive or helpful in organising appropriate patients as well then it can be quite tricky. I know there are registrars where the job has fallen on the registrar to gain consent for certain patients and then it ends up being quite awkward where they get two or three or more patients continually saying no, they don't want to be observed. And you know it is down to the patient in general and that can be quite tricky. So, as long as the practice is supportive and is able to help in organising all of that then it can be okay."* Practice support for the process therefore seems to be key.

Patient population

Similarly, it was felt that with practice knowledge, signage and familiarity with the practice, the practice's patient population seemed to be generally supportive across all interviews. However, it was felt that there might be some difference in the level of support between the socio-economic mix of the particular practice's patient population. As R8 explained: *"The only problem I have had is when we get new patients who kind of fly in and fly out whenever they can't get into their regular doctor, who are disgruntled about having to pay something, a GAP to begin with, who then has to come in to see a registrar who has to phone another GP supervisor to ask for advice. I think those situations, those patients, I can tell on their face that they're probably not that impressed. Particularly since they know they have to pay to be there. That's in contrast to clinics in kind of lower, more lower socio-economic suburbs that are kind of completely bulk billing. So the patients that go to those clinics have the expectation that they're only going to be seen for five minutes. So in which case, it's often just a simple complaint that a registrar might generally be able to deal with."*

Regional Training Organisation

Some research participants also considered how the relevant RTO might assist in the context of consultation observation. Just one example of how the RTO may be able to offer better support came from R7: *"I think overall, once you get your head around it, it makes sense. Initially I did have some trouble with it because our online learning platform through [our RTO] is very much based on*

the RACGP requirements and so a lot of the links are to the RACGP pages and just down the bottom it says 'ACRRM registrars don't have to complete this. Instead go here'. So initially I did some on the wrong paper and got a bit confused initially, but once I read that little fine print down the bottom, I worked it out and its fine now. [INTERVIEWER: So RTO instructions need to be a bit clearer?] I think so. Yeah. I think so."

Summary of findings in relation to the guiding research questions

Findings from the debriefs and interviews offer insight into how the observer, observee, the observational tool (ECTV/DOV) and context interact to form the consultation observation event. The resulting variations in these events, and participants' reflections and interpretations of a given event, provide us with information about (1) factors that can enhance consultation observation as an educational and assessment tool, and (2) the information, training and or support that is needed for supervisors, MEs and registrars.

In this section, the research findings are discussed in relation to each of the guiding research questions. The practical implications for future use of consultation observations as an assessment and educational tool in GP training are also discussed.

How is consultation observation being framed, from an assessment and educational standpoint, in GP vocational training? Second, how can consultation observation be framed and supported so that it is both an informant of registrar progress and competency, and an educational tool?

Consultation observation via clinical teaching visits (e.g., ECTV, DOV) were originally introduced into Australian GP training as a formative assessment tool to supplement in-practice teaching: "These assessments occur during training and are low stake assessments for learning. Formative assessments aim to generate powerful learning experiences from feedback and identify learning needs" (Ingham, Fry & Ward, 2016, p.919).

The perspectives of the RTOs participating in this study are aligned with this statement. Consultation observations are framed primarily as a learning opportunity for the registrar, and secondly as a means for RTOs to gain information about the registrar's performance and any red flags that require further intervention.

Participating supervisors and MEs recognised consultation observation as an assessment tool, and their requirement to provide a written evaluation for their RTOs. However, most regarded consultation primarily as an educational opportunity to provide feedback, advice and encouragement to the registrar. The debriefing sessions revealed a strong focus on the observer providing constructive feedback for improvement and positive reinforcement for the registrar's strengths.

Correspondingly, registrars demonstrated a high level of receptiveness to feedback from the observer, and reported that as a primary benefit of consultation observations for themselves. However, several of the registrars had significant anxiety and discomfort about the process of being observed and assessed during consultations. The underpinning awareness that an evaluation of their performance will be recorded and submitted to their RTO created understandable discomfort during the observations (and debrief) for many of the registrars; particularly those who identified as being shy or introverted, or less confident about themselves as a GP. This has the potential to impact the registrar's performance during the observed consultations and undermine their experience of what can be a valuable educational experience.

The implications of this for the framing of consultation observation in GP training are difficult. Arguably, registrars may benefit more from consultation observations if an assessment is not provided to their RTO. However, consultation observations are a central and timely source of information for RTO medical educators as to whether the registrar has appropriate skills and knowledge for their level of training, the areas that need further development, and the interventions that may be required.

The findings from this study suggest that consultation observations should continue to be framed as a formative assessment task designed foremost for registrar learning and development; and that strong emphasis is placed on this in the resources and training provided for registrars, MEs and supervisors.

What enables and impedes consultation observation as an educational tool?
This study identified several factors that appear to influence the educational value of consultation observation for registrars, as summarised below.

Enablers

- Longitudinal relationship (e.g., multiple observations over time) to facilitate familiarity, trust/credibility, and continuity of feedback over time as the registrar develops.
- Debriefs after every one or two consults, to enable immediate and specific feedback.
- Clear and shared expectations between observer and observee about the purpose and process of the observation visit, and the role of the observer in the room.
- Registrar input into the structure of the visit (e.g. opting for debriefs between each patient).
- A balance of specific constructive feedback for improvement, positive reinforcement of registrar strengths, and affirmation of their identity as a developing GP. Supervisor and ME training on feedback dialogue is warranted.
- Invitations for the registrar to seek advice and ask questions.
- A supportive learning culture within the training practice that enables appropriate preparation, and suitable space, for consultation observations.

Impediments

- Delaying all debriefing until the end of a session.
- Observation by a supervisor within their own clinic, due to familiarity with patients and an increased likelihood of patients trying to engage with the supervisor during observation.
- Lack of formal training for supervisors and MEs undertaking observations.

What enables and impedes consultation observation as an assessment tool?
The study also identified several factors that may impact the observer's ability to form a well-reasoned judgement about the registrar's competence, as summarised below.

Enablers

- Clear and specific guidelines for observers regarding:
  - The process of conducting a consultation observation visit, and the variations in approach that they can consider and utilise based on theirs and the registrar's preference.
  - Their role as an observer, and when it is reasonable to intervene within a consult.
- Training for less experienced supervisors and MEs, and those who wish to develop their skills.

- Repeat observations over time between registrar and observer, to assess the registrar's progression.
- Use of the usual consulting room in which the registrar sees patients, or a close approximation, so that registrars have their frequently used items at hand and are familiar with the layout and equipment within the space.
- Use of strategies to minimise the presence of the observer within the room (e.g., sitting out of line of sight of the patient and registrar, avoiding eye contact with patient, using a laptop or pen and paper as a symbolic barrier).
- Communication with the patient at the commencement of the consultation re: the observer's role in the room.
- Use of strategies to minimise registrar anxiety whilst being observed (re-iterating some of the points above):
    - Repeat observations over time to facilitate familiarity and trust between observer and registrar.
    - Familiar consulting space and specific positioning to minimise presence of observer.
- User-friendly methods of note-taking and assessment that can be populated during observations. As noted previously, some participants had created work-arounds to issues experienced previously in the completion of online forms, through creating their own template in Word or similar to mirror the online system, enabling them to complete the form during the observation and avoid problems with system crashes.
- A supportive practice culture that facilitates the necessary preparation and environment for consultation observations.

Impediments

- Observation by a supervisor within their own clinic, due to an increased likelihood of patients trying to engage with them during observation.
- Registrar discomfort at being observed, and the potential impact on performance.
- Overly complex or faulty systems for recording assessments.
- The time demands of preparing for and conducting the consultation observation and assessment for the RTO, which can sometimes result in shortcuts that risk compromising assessment validity.
- Lack of formal training for observers.
- Inappropriately scheduled patient consults (i.e., too few patients to enable a reasonable assessment, too many patients to enable sufficient time for debriefing).

Limitations of the study

Participants were recruited from two RTOs, offering insight into registrar, supervisor and ME experiences and perceptions of observation and assessment tools and processes for those two regions. Whilst some of the findings have generalisability to the broader community of GP training, there are likely to be specific variations in how other RTOs undertake consultation observations that have not been identified in this study. We suggest that further research is undertaken with additional RTOs and a larger cohort to identify whether particular elements of observational tools and processes have greater benefit in terms of registrar assessment and education.

In situ audio recordings of debriefing discussions were a notable strength of the study, offering a window into the way in which registrars and supervisors/MEs engage in consultation observations and subsequent debriefing and feedback exchange. However, this method also has recognised shortfalls in terms of (1) the potential impact of the presence of audio-recording on the discussions (e.g., greater self-editing); and (2) the potential for recordings being limited to the 'formal' component of discussions and not necessarily the additional and often valuable chat that

occurs before and after. The approach of this study, which accommodated the variations in methods of observation and debriefing, meant that participants did not always capture *all* of the discussions held between in relation to their ECTV/DOV. Whilst this more naturalistic approach minimised the imposition on participants, in future studies it may be beneficial to seek audio recordings of *all* debriefing discussions in relation to consultation events.

Conclusions and recommendations

This study offers insight into the commonalities and variations in how consultation observations are enacted as an educational and assessment tool, and how specific interactional elements of the observer, observee, tool and context may influence educational and assessment outcomes. However, the findings must be interpreted in light of the limitations of the study. Further research is needed across a broader cohort of registrars, supervisors and MEs from other RTOs, to further understand the differing approaches to consultation observation, the strengths and weaknesses of these various approaches, and the implications for future practice.

However, findings from the study provide some valuable direction for future use of consultation observation within the WBA framework arising from the broader ERG study. We recommend that the enabling and impeding factors outlined in the discussion are considered within the framework and potential resources and training arising out of the study for supervisors, MEs and registrars.

References

Bennett, D., Solomon, Y., Bergin, C., Horgan, M., & Dornan, T. (2017). Possibility and agency in Figured Worlds: becoming a 'good doctor'. *Medical Education*, *51*(3), 248-257.

Braun, V., & Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. http://dx.doi.org/10.1191/1478088706qp063oa

Brydges, R., Hatala, R., Zendejas, B., Erwin, P.J., Cook, D.A. (2015). Linking simulation-based educational assessments and patient-related outcomes: A systematic review and meta-analysis. *Academic Medicine, 90*(2), 246-256.

Clarke, V., & Braun, V. (2013). *Successful qualitative research: A practical guide for beginners.* London: Sage.

Essers, G., van Dulmen, S., van Es, J., van Weel, C., van der Vleuten, C., Kramer, A. (2013) Context factors in consultations of general practitioner trainees and their impact on communication assessment in the authentic setting. *Patient Educ Couns*, *93*(3), 567–72.

Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (1998). *Identity and agency in cultural worlds*. Harvard University Press.

Ingham, G., Fry, J., & Ward, B. (2016). Adding random case analysis to direct observation (ARCADO)-Updating the external clinical teaching visit to improve general practice registrar assessments. *Australian family physician*, *45*(12), 918.

Khan, M. A. A., Gorman, M., Gwozdziewicz, L., Sobani, Z. A., & Gibson, C. (2013). Direct Observation of Procedural Skills (DOPS) as an Assessment Tool for Surgical Trainees. *Journal of Pakistan Medical Students, 3*(3), 137-140.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*(12), 1316–1326.

Kogan, J.R., Hatala, R., Hauer, K.E., Holmboe, E. (2012). Guidelines: The do's, don'ts and don't knows of direct observation of clinical skills in medical education. *Perspectives on Medical Education, 6*(5):286-305.

Laurence, C., Docking, D., Haydon, D., Cheah, C. (2012). Trainees in the practice – Practical issues. *Australian Family Physician, 41*(1–2), 14–7.

Magin, P., Stewart, R., Turnock, A., Tapley, A., Holliday, E., & Cooling, N. (2017). Early predictors of need for remediation in the Australian general practice training program: a retrospective cohort study. *Advances in Health Sciences Education*, *22*(4), 915-929.

Molloy, E., & Bearman, M. (2019). Embracing the tension between vulnerability and credibility: 'intellectual candour' in health professions education. *Medical Education*, *53*(1), 32-41.

Norcini J.J. (2003). Work based assessment. *British Medical Journal, 326*(7392), 753.

O'Sullivan, B.; Russell, D; McGrail, M., Sampson, M. Warrington, A., Wallace, G., Bentley, M. & Couch, D. (2019). Factors related to rural general practitioners supervising general practice registrars in

Australia: 'A national cross-sectional study' [online]. *Australian Journal of General Practice*, *48*(1/2), 66-71.

Rencic, J., Durning, S. J., Holmboe, E., & Gruppen, L. D. (2016). Understanding the assessment of clinical reasoning. In *Assessing competence in professional performance across disciplines and professions* (pp. 209-235). Springer, Cham.

Stewart, R., Cooling, N., Emblen, G., Turnock, A., Tapley, A., Holliday, E., ... & Magin, P. (2018). Early predictors of summative assessment performance in general practice post-graduate training: A retrospective cohort study. *Medical Teacher*, *40*(11), 1166-1174.

Usherwood, T., Kefalas, B., Knight, A. (2005). External clinical teaching visits in the regionalised environment [online]. *Australian Family Physician*, *34* (12 Suppl), 10-2.

Wahyuni, D. (2012). The research design maze: Understanding paradigms, cases, methods and methodologies. *JAMAR*, *10*(1):69-80.

Wearne, S., & Brown, J. (2014). GP supervisors assessing GP registrars-theory and practice. Australian family physician, 43(12), 887.

Wearne, S., Dornan, T., Teunissen, P. W., & Skinner, T. (2012). General practitioners as supervisors in postgraduate clinical education: An integrative review. *Medical Education, 46,* 1161–1173.

Appendix 1 – Interview questions (supervisor/medical educator)

Introduction

Thank you for your willingness to participate in this research. The interview will go for between 45-60 minutes. I'll be asking you to reflect on your recent consultation observation event and subsequent feedback/debriefing discussion with your [Supervisor/ME or Registrar].

I encourage you to take your time with answers and be as honest as you can. We want to develop a clear understanding of how consultation observations are used and experienced by Supervisors and MEs who are performing the observation, and by the Registrars who are being observed. Ultimately we hope to develop a clearer understanding of how consultation observations can be best used as both an assessment and educational activity.

Your responses will remain completely confidential – you will not be identifiable in the reporting or presentation of any findings coming out of the study. The intention is to audio record this interview for transcription and analysis. Are you happy for me to go ahead with recording the interview?

If yes → Commence recording.

For the purposes of the recording, can you state whether you are a Supervisor/Medical Educator, or a Registrar?

[Response]

I'd like to ask some questions about your experiences and perspectives on consultation observation in general, and then more specifically about the consultation observation session and follow-up discussion you had recently with your Registrar [registrar name].

This interview will be broken into a series of four key sections. In the first, we would like to gain your insights on:

1. The assessor (the supervisor/ME

- To get a bit of background:
  + how much experience have you had with consultation observation?
- What do you see as the main purpose(s) of consultation observation?
- What do you see as your main role(s) in the consultation observation session?

Prompt: Do you have a particular "hat" on during the visit? E.g., assessor, educator, mentor?

- Could you tell me about the recent Consultation observation session that you had with [insert name registrar]?
- What in particular were you looking for while you were observing them?
  Prompt: were you keeping an eye out for specific behaviours or characteristics?
- How comfortable or confident were you with knowing what to look for during the visit?
- Did the visit give you enough opportunity to observe and form judgements about the Registrar?
  Prompt: why/why not? Could you tell me a bit more about this?
- Could you tell me about the follow-up feedback/debriefing discussion you had with the Registrar?

Prompts: How soon after the observation was it? Who led the discussion? What was discussed? [Try to ascertain how much the discussion focussed on assessment, and how much on providing educational support and feedback to the registrar.]

- Did you experience any difficulties in assessing and providing feedback to the Registrar?
  Prompts: Could you tell me more about this? How did you navigate/deal with these difficulties?

2. Relationship between the assessor (supervisor/ME) and the assessed (Registrar)

- How well did you know the Registrar prior to this round of consultation observation session?
  Prompt: have you assessed, taught or worked with the Registrar before?
- What are your reflections on the way you both interacted?
- Were there any notable interpersonal dynamics that might have impacted on:
  + Your performance,
  + the way that you communicated and engaged with each other,
  + and the follow-up discussion?

  Prompts: Could you tell me about this? How did you manage it?
  + The impact of the feedback from the visitor on you
- From the feedback discussion:
  + did your assessment of the Registrar match with their own self-assessment?
  + If there were disparities, how did you manage these?
  + How did you perceive that your feedback impacted on the Registrar?
  + In discussing the feedback, what were the things you felt you needed to say?
- If they have an ongoing relationship with the Registrar:
  + Did the consultation observation session have any notable impact on your relationship?
  + In what ways?

3. The assessor and the tool

- Thinking about your most recent ECTV, could you tell me which parts of the process and forms work well? Not so well?
  Further prompts:
  + Did the instructions assist or hinder consultation observation and feedback discussion? How?
  + Did the process and forms facilitate or hinder the interaction/discussion between you and the Registrar? How?
  + If they hindered either: how did you manage this?
- To what extent did the ECTV as a whole enable you to:
  + Form an accurate judgement of the Registrar's performance and competency?
  + Accurately report back your assessment to EV/GPEx/GPTT?
  + Openly provide feedback to the registrar? Prompt re: "private" vs "public" evaluation of the registrar.
  + Facilitate the registrar's learning

4. Contextual factors

In this section, we want to reflect on the effect that the environment plays in Consultation observation. We will ask questions about you as a person, the specific situation, and the broader context.

- Are there things about you as a person that impacted on the usefulness of the consultation observation?
- Were there things about the immediate environment that impacted on the usefulness of the consultation observation?
  + The physical space
  + The practice environment
  + The patient population
- Were there things about the broader training context that impacted on the consultation observation:
  + The training program perhaps?
- Do you think your current approach to ECTVs (consultation observation, feedback/debriefing, completion of forms) meets that purpose? Why/why not?
- Has your approach to the visits changed over time? How? What has influenced this?
- Do you take a consistent approach to ECTVs, or do you tailor it to the situation? For instance, the Registrar's stage of training or specific expressed needs, the clinic you're visiting, the registrar's supervisor.
- ***Do you think your approach to ECTVs is in line with:***
  + ***what other Supervisors and MEs do?***
  + ***EV/GPEx/GPTT's expectations about the visits?***
  + ***Why or why not?***

5. Final thoughts

- Are there any more comments that you would like to make on the process of consultation observation itself and as an assessment tool in your GP training?

Appendix 2: Interview questions (registrar)

Workplace-based assessment 3a – registrar interview schedule

Introduction

Thank you for your willingness to participate in this research. The interview will go for between 45-60 minutes. I'll be asking you to reflect on your recent consultation observation event and subsequent feedback/debriefing discussion with your [Supervisor/ME or Registrar].

I encourage you to take your time with answers and be as honest as you can. We want to develop a clear understanding of how consultation observations are used and experienced by Supervisors and MEs who are performing the observation, and by the Registrars who are being observed. Ultimately we hope to develop a clearer understanding of how consultation observations can be best used as both an assessment and educational activity.

Your responses will remain completely confidential – you will not be identifiable in the reporting or presentation of any findings coming out of the study. The intention is to audio record this interview for transcription and analysis. Are you happy for me to go ahead with recording the interview?

If yes → Commence recording.

For the purposes of the recording, can you state whether you are a Supervisor/Medical Educator, or a Registrar?

[Response]

I'd like to ask some questions about your experiences and perspectives on consultation observation in general, and then more specifically about the consultation observation session and follow-up discussion you had recently with your visitor [visitor name].

This interview will be broken into a series of four key sections. In the first, we would like to gain your insights on:

1. The assessed (the Registrar)

- To get a bit of background:
  - how much experience have you had with consultation observation?
- What do you see as the main purpose(s) of consultation observation?
- What do you see as your main role(s) in the consultation observation session?
- Could you tell me about the recent consultation observation session you had with your Supervisor/ME?
- Did the visit give you enough opportunity to have a discussion with the Supervisor/ME?
  Prompt: why/why not? Could you tell me a bit more about this?
- Could you tell me about the recent follow-up feedback/debriefing discussion you had with the Supervisor/ME?
  Prompts: How soon after the observation was it? Who led the discussion? What was discussed?
  [Try to ascertain how much the discussion focussed on assessment, and how much on providing educational support and feedback to the registrar.]
- Did you experience any difficulties in the feedback from the Supervisor/ME?

Prompts: Could you tell me more about this? How did you navigate/deal with these difficulties?

2. Relationship between the assessed (the Registrar) and the assessor (the Supervisor/ME)

- How well did you know the Supervisor/ME prior to this round of consultation observation session?
  Prompt: have you worked with the Supervisor/ME before?
- What are your reflections on the way you both interacted?
- Were there any notable interpersonal dynamics that might have impacted on:
  + Your performance,
  + the way that you communicated and engaged with each other,
  + and the follow-up discussion?
    Prompts: Could you tell me about this? How did you manage it?
  + The impact of the feedback from the visitor on you
- From the feedback discussion:
  + did the assessment you receive match with your own self-assessment?
  + If there were disparities, how did you manage these?
  + Did you give any feedback to the Supervisor/ME?
  + In discussing the feedback, what were the things you felt you needed to say?
- If they have an ongoing relationship with the Supervisor/ME:
  + Did the consultation observation session have any notable impact on your relationship?
  + In what ways?

3. The assessor and the tool

- Thinking about your most recent consultation observation session, could you tell me which parts of the process work well? Not so well?
  Further prompts:
  + Did the instructions assist or hinder the consultation observation and feedback discussion? How?
  + Did the process facilitate or hinder the interaction/discussion between you and the Supervisor/ME? How?
  + If they hindered either: how did you deal with this?
- To what extent do you believe that the consultation observation session as a whole enable the Supervisor/ME to:
  + Openly provide feedback?
  + Facilitate your learning?
  + Form an accurate judgement of your performance and competency?
  + Accurately report back your assessment to your organisation (EV/GPEx/GPTT)?

4. Contextual factors

In this section, we want to reflect on the effect that the environment plays in Consultation observation. We will ask questions about you as a person, the specific situation, and the broader context.

- Are there things about you as a person that impacted on the usefulness of the consultation observation?
- Were there things about the immediate environment that impacted on the usefulness of the consultation observation?
  + The physical space
  + The practice environment
  + The patient population

- Were there things about the broader training context that impacted on the consultation observation:
  + The training program perhaps?

5. Final thoughts

- Are there any more comments that you would like to make on the process of consultation observation itself and as an assessment tool in your GP training?

Stream 3b
A qualitative investigation into the acceptability of workplace-based assessments in Australian general practice training

Dr Emily Kirkpatrick (GPEx)
Assoc Professor Jill Benson AM (GPEx)

GPEx In collaboration with

GP training · Flinders University · general practice training tasmania · MURRAY CITY COUNTRY COAST GP TRAINING · NORTHERN TERRITORY General Practice Education

RVTS REMOTE VOCATIONAL TRAINING SCHEME LTD · WAGPET · GP SYNERGY · GPTQ

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

GPEx would also like to acknowledge:

- The Working Group for this stream including Dr Emily Kirkpatrick (GPEx), Assoc. Prof Jill Benson AM (GPEx), Ms Michelle Pitot (GPEx), Dr Taryn Elliott (GPEx), Dr Lina Zbaidi (Northern Territory General Practice Education)and Ms Kiara Beens (GPEx).
- Dr Lina Zbaidi (Northern Territory General Practice Education), who contributed data and the ethics backbone to this stream, and also Dr Nyoli Valentine (ModMed Institute) for assistance with focus group facilitation.
- The hard work of the Australian general practice training organisation collaborators who helped facilitate and arrange focus groups and interviews. We would like to thank Eastern Victoria GP Training (EV), General Practice Training Tasmania (GPTT), Murray City Country Coast GP Training (MCCC), Northern Territory General Practice Education (NTGPE), Western Australian General Practice Education and Training (WAGPET), Remote Vocational Training Scheme (RVTS), GP Synergy and General Practice Training Queensland (GPTQ) for their contribution.
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.
- With special thanks to all participants who contributed their opinions to this research.

GPEx In collaboration with

GP training · Flinders University · general practice training tasmania · Murray City Country Coast GP Training · Northern Territory General Practice Education

RVTS Remote Vocational Training Scheme Ltd · WAGPET · GP Synergy · GPTQ

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Introduction

In October 2018, the Royal Australian College of General Practitioners (RACGP) awarded GPEx a Special Education Research Grant with the aim of designing a workplace-based assessment (WBA) Framework for use within Australian General Practice Training (AGPT). As part of this grant, Stream 3b was established to undertake a qualitative investigation of the acceptability of WBAs in Australian general practice training.

The research for this stream was guided by the following research questions:

Supervisors and medical educators (MEs):

- What roles do supervisors and MEs have in WBAs of registrars?
- How do supervisors and MEs feel about their roles as assessors of registrars?
- What are the challenges and opportunities for supervisors and MEs in their roles as assessors of registrars?
- How do supervisors and MEs view programmatic assessment models and how they might impact on their roles?
- How do GP training organisations and training practice structures and characteristics impact the utility of WBAs?

Registrars:

- How do registrars perceive the roles of supervisors and MEs as work-based assessors?
- What challenges and opportunities do WBAs present for registrars?
- How do registrars perceive programmatic assessment models and how might moving towards these affect them?
- How do GP training organisations and training practice structures and characteristics impact the utility of WBAs?

In addition questions were asked about attitudes to the exams and to the place of WBAs in programmatic assessment.

Focus groups and interviews with registrars, supervisors and Medical Educators (MEs) were performed to identify the role of WBAs in general practice training and the perceptions and attitudes of those who are both using the WBA tools and administering the tools. A grounded theory approach was utilised to then explore the themes that emerged from the participants, and a subsequent literature review undertaken to gain greater depth of understanding of the utility and acceptability of WBAs. The discussion pulls together the results and literature review to synthesise the key concepts that emerged and how they interact in the WBA landscape.

Australian general practice training is delivered by regional training organisations and by the Remote Vocational Training Scheme (RVTS) who deliver training across Australia to registrars in urban, rural and remote locations. For the purpose of this report refernces to all general practice training organisations will be abbreviated as ‘RTO’.

Methodology

Participants

Between November 2018 and March 2019, 127 eligible individuals participated in focus groups and one-on-one interviews, these included medical educators (n=28), supervisors (n=41) and registrars (n=58), with participants representative of the nine GP training organisations partaking in the Education Research Grant. Participants were recruited through their general practice training provider, with no direct approaches made by the interviewers to participants.

Table 1. Research participant breakdown

Training Provider	*Registrars*	*Supervisors*	*MEs*	
1	11	10	4	
2	6	3	8	
3	10	7	0	
4	6	0	2	
5	4	7	1	
6	3	4	2	
7	8	4	3	
8	2	2	5	
9	8	4	3	
Total	58	41	28	***127***

Procedures

Focus groups were undertaken either online with video conferencing or face-to-face where possible, in line with ethnography. Focus group discussions were semi-structured, based on a set list of questions about WBAs, the role of supervisors and MEs, and perceptions and attitudes towards the RTO. The aim was to gain assessor and registrar views regarding workplace-based assessments in general practice training. Interviews continued to be undertaken until saturation was reached.

Ethics approval was obtained through Flinders University HREC, number 8214.

Analysis

Interviews were transcribed using a professional transcription service, with content analysis taking place whilst the interviews were undertaken, to ensure concepts could be further explored as the interviews progressed. Each transcript was coded individually in Nvivo Version 12 for Windows (QSR International, 2016). Themes were extracted from the transcripts taking into account the specificity and relevance of the comments made in relation to the original research questions. Coding was undertaken by two analysts, independently, thus increasing the reliability of coding with respect to the themes being uncovered. The analysis was undertaken in line with

Krippendorff, ensuring replicable and valid inferences from the transcripts (Krippendorff, 2004). Key quotes were selected from the interview transcripts to be included in the report.

Results

The results have been reported under the themes that emerged from the interviews and focus groups. The assessors, context, and interactions will be discussed first, followed by the tools and the utility. Finally stakeholder thoughts on current exams and programmatic assessment will be reported. Quotes from the coding of transcripts have been included to provide context to the narrative below. ME = Medical Educator, S = Supervisor and R = Registrar.

Medical Educator (ME) role

The strongest theme to emerge was the importance of the medical educator (ME) role. Exploration of the ME role was not initially one of the aims of this research project, however the crucial role that MEs play in workplace-based assessments and in overall registrar training became apparent. As such, it was decided that the ME role would be reported as part of this stream.

Not only does the ME have a teaching and educating role, but they are fundamental in offering pastoral support for registrars. MEs were seen as a 'port-of-call' for registrars when they needed help, clarification or training advice.

> *Registrar: "I see the overall medical educator as your safety net. It's that person who you can talk to if you think that you've got gaps so that you can start to address them. If you've got issues with training within your practice that's the person that you should be going to."*

In addition, the medical educator, who is usually a practicing general practitioner (GP) themselves, can be the 'go-to person' for the supervisor if they have any concerns about the registrar or additionally if they themselves would like assistance with their supervisory role.

> *ME: "They basically say that the number one thing that they really valued was the genuine support and the sense that we've got medical educators who are really trying hard to do things for them and that we genuinely believe in their capabilities of becoming good GPs."*

> *ME: "Some of it can be systems things, for example, the issue might be that the registrar's doing great, but they're not feeling well supported and supervised and sometimes that can be just having a conversation with the supervisor, when I get to meet with them, or it can be at a practice level, because it might be a new practice, where they haven't had registrars before."*

The ME is usually the key collaborator when it comes to programmatic assessment at the RTOs, monitoring assessment tasks and ensuring a registrar is tracking towards becoming a safe and competent practitioner.

> *ME: "As a training organisation senior medical educator…I'm looking for good quality day to day assessment information, so that we can put together an overall picture of that registrar, through programmatic assessment and get an idea of how that registrar is progressing through their training and what kind of additional support they might need. Where their placements might need to be in the future, where are their strengths, where are their weaknesses, so we can fill in the gaps."*

Concerns were mentioned about how MEs might feel if they took a more active role in assessing registrars in line with potential high-stakes assessments, considering that they fill such an important pastoral care role. The viewpoint was that the ME wore many hats and had many responsibilities to ensure the registrars is safe and competent, which appears advantageous when gathering data for programmatic assessment. The ME can hold the role of mentor, assessor, advisor, educator, remediator and pastoral carer on any one day.

> *ME: "I do have to be the bad cop sometimes already anyway so it is sort of a mixed relationship that they have with me. So I do a lot of our pastoral care and support for registrars but I also am the one that says you're not performing as expected."*

> *ME: "They have a dual role as well so they're pastoral care but they're also mentoring and career advice sort of that kind of point of what should I do next? I'm thinking of doing this and what are the rules around that. They also provide the summary of feedback so they do have to have some of those feedback conversations with the registrar."*

Supervisor role

> *Registrar: "The word supervision means a person who was there to make sure no one gets killed or knows that everyone is safe."*

The supervisor plays a vital role in ensuring that a registrar develops their primary care skillset and becomes a safe and competent practitioner, who values continuous learning and development. Several registrars and supervisors commented on the dynamics being more of a mentor-mentee relationship, helping to positively guide and educate registrars into becoming excellent GPs. They commented that those registrars who had the best supervisors, were often the most confident when it came to passing their exams and becoming a Fellow of the College.

> *ME: "Someone that can help shape you and they can really make or break your experience as a registrar, especially in your first terms, when you need them so much and you're going to them with so many questions. If you haven't got a good supervisor in your first term, it may turn you off the entire job. You might be looking at going back to the hospital pretty quickly."*

The supervisor is in the ideal position to undertake informal observation as part of their role, and thus incorporate these findings into their mid and end-term reports, incorporating entrustable professional activities around the supervision level the registrar needs.

> *Supervisor: "I think you've got a better understanding of the depth of knowledge and the clinical reasoning a little bit more than just being in there for one session or three hours."*

Discussions were raised in the interviews and focus groups around the role of the supervisor in comparison to the ME. One viewpoint was that the supervisor could provide further education and support to the registrars, as part of their progression. However, it became clear that the role of the ME and supervisor differ substantially. The supervisor provides day-to-day support, delivering a safe learning environment to allow the registrar to grow and develop competency and safety. It was emphasised that some supervisors don't necessarily want to take on additional responsibilities regarding assessment, with concerns around additional training, remuneration and time constraints.

> *Registrar: "There is definitely quite a bit of value in having somebody who should be a supervisor, more than somebody who just wants to be."*

The supervisor typically undertakes the term assessments at many RTOs, reporting back on how the registrar has progressed and their overall performance. However, there appears to be a lack of structured responsibilities around the supervisor role and what expectations exist at various RTOs. When discussing the supervisor's role in direct observation and initial safety observation visits within the first week, there was considerable disparity, with many reporting that they did not observe their registrar practising, but provided support when approached by their registrar. One of the challenges was that supervisors felt undervalued and under-remunerated at times by the RTO. Thus there is a need to fund supervisors to undertake a greater supervisory role, as much as is reasonable from the RTO budget.

> *Supervisor: "The system that we've got is that we're virtually doing this in an honorary capacity."*

Consideration should be given to the role of the supervisor and ensuring that supervisors are accountable for providing adequate supervision, but are also remunerated adequately for their time. In addition, there needs to be a regular review of the supervisor's supervising role to ensure that registrars are not being exposed to outdated and poor quality supervisors, resulting in a negative outcome for registrars.

> *Registrar: "Some have been really shocking and some just really outdated, even though they're well intentioned.*

Supervisors expressed considerable enjoyment in training registrars and believed they were investing in the next generation of GPs.

> *Registrar: "I see the supervisor as the day-to-day person that essentially makes sure that you're safe to practice and has an on the ground education role."*

Advantages reported by supervisors included the potential for work-force planning, ensuring their clinical knowledge was kept up-to-date, networking opportunities with other supervisors and that supervisor work provided a potential opportunity to then move into a medical education role.

Assessment Context

When developing the aims of this large Education Research Grant there was an overarching need to understand the context of assessment and understand the environment and culture of the delivery of WBAs, including the culture of the GP training organisation (RTO). As such, understanding the relationship of the assessor and the registrar and how these factors all interrelate are reported. The relationships were explored between the supervisor and registrar, ME and registrar, and the RTO, to establish whether there is an underlying tension between assessor and registrar and how this may translate in the context of assessment.

> *Registrar: "I think we need to make sure that we're protecting the expert nature of our Fellowship and I think potentially, I would worry if you're getting rid of more assessments, I would worry about the perception that that's going to have on whether or not GPs are experts, because they are specialists in this area."*

Relationship of supervisor and registrar

The relationship between the supervisor and registrar was explored across the interviews and focus groups.

> *ME: "There is a huge amount of trust that's required in that relationship."*

> *ME: "I think that there is no getting around that. It's the role of an apprentice and a teacher. You are responsible to the teacher, but you are also - the teacher is responsible for the apprentice as well. There is a huge amount of trust that's required in that relationship. Assessing as we go is what we do regardless of whether it's formalised or not."*

> *Registrar: "I think that that working relationship is pretty important with medical training in general, because undoubtedly we're all going to come across some sticky situations in terms of training. So you need a supervisor who can sit you down and go this is what you should have done to get out of this mess earlier or something like that, because generally I think there's a very collegial structure in general practice."*

The concept of intuition and 'gut feeling' when undertaking assessments was not explored specifically in this project. However, there were mentions in the focus groups of assessors making judgements based on background knowledge, with the collation of various viewpoints. The concept of how an assessor makes a decision could be explored in future research projects.

> *Registrar: "I actually think I'd prefer to be assessed by my direct supervisor because I feel like they would have a better idea of my practice every day….So they get a gut feeling about whether or not you're up to par and I think that is probably difficult to formalise an assessment that will play a big role."*

The role of a mentor versus that of an assessor was also raised.

> *Registrar: "Certainly, I needed my supervisor to be kind of there for me and knowing that he wasn't going to be super critical or anything like that. Like, I could just go, I'm getting really stressed out or something like that and I could go to him about those things. Whereas I think if I knew that he had more formal assessment stuff of me, I don't know, that might have made me feel more uncomfortable."*

The ability to be objective whilst undertaking an assessment was also discussed, and how this may have an impact on the apprentice-type relationship for the supervisor if they are required to also be the assessor. On the other hand, by understanding the registrar from different angles and spending time with them in environments such as the tea room, during hospital visits or in meetings, the supervisor is in an ideal position to gain an understanding of the registrar's overall abilities as a person.

> *Supervisor: "We're interacting a lot after hours as well and I guess I'm doing it in lots of different ways, not just assessing but monitoring to see how he's going, interacting as much as I can in lots of different ways, in social ways, on the floor, when we work and seeing patients."*

A positive relationship between a supervisor and a registrar has the potential to transform the registrar into a safer doctor who feels comfortable seeking support and guidance.

> *Supervisor: "I make a big effort to meet them beforehand. I think that's really important to build their confidence and if they're confident and they perform well and then then there's less problems."*

If registrars are not willing to engage actively in the relationship with their supervisor, it can negatively impact on their progress. There was much discussion about what constitutes a good supervisor, however it is important to also assess what constitutes a good registrar.

> *Supervisor: 'I'm not entirely sure that I would want to be in the role of the person that passes or fails someone because of that relationship you have with that registrar. You are actually effectively their employer. You're wanting to recruit good GPs to your practice. That doesn't mean, yeah, I just find the idea of me - I know if - I have known whether GP registrars, are going to pass or fail their exams and it's worked. I've said to my GP registrars I don't think you're ready to sit the exam and they haven't taken that on board and sat the exam and failed the exam."*

Tension when the supervisor is assessing the registrar

The power imbalance between registrars and supervisors was raised as a potential conflict across the various stakeholders being interviewed, however it was emphasised that this would be more of an issue if the supervisor was singularly responsible for deciding whether a registrar passed or failed their Fellowship. Issues that need to be considered will be that supervisors have various different personalities and unconscious biases, how engaged the supervisor is in active supervision and possible financial implications, such as if the supervisor was also the employer. Blurring the boundary was mentioned when it came to balancing the mentor and coaching role with the assessing role.

Supervisors reported that they often sugar-coated feedback and would struggle to give negative feedback to their registrar, particularly if they had formed a positive relationship with their registrar. Supervisors, having gained an overall view of the registrar, often felt that their registrar would improve over time and didn't want to 'flag' the registrar for being difficult. Some supervisors mentioned that by failing their registrar it may draw attention to their own skillset and imply a lack of ability to remediate the registrar.

> *Supervisor: "But I think there's that tension, there's always been this tension around assessor versus supervisor, versus also, included in the equation is employer. There's all these different tasks that a supervisor has to have and all different hats."*

The mid and end-term assessments, in conjunction with direct observation visits, provide a longitudinal picture of the registrar's progress, which form the basis of a programmatic picture, in order to collate multiple view-points.

> *ME/Supervisor: "I think that there is no getting around that. It's the role of an apprentice and a teacher. You are responsible to the teacher, but you are also - the teacher is responsible for the apprentice as well. There is a huge amount of trust that's required in that relationship. Assessing as we go is what we do regardless of whether it's formalised or not."*

It is evident from the transcripts that there is a necessity to separate the role of supervisor and assessor to ensure an objective assessment.

> *ME: "The person doing the summative assessment needs to be different from the person who's doing the formative assessment because of that relationship and the coaching and the mentoring and the pastoral care thing."*

Supervisors provide key pastoral support to registrars. If supervisors were undertaking high-stakes assessments on their registrar there is a potential that a registrar may not seek help and support from their supervisor, putting both themselves and potentially patients at risk.

Failure to fail, failure to flag and failure to give critical feedback

> *ME: "It is, the failure to fail and wanting to be everybody's friend."*

The concept of 'failure to fail' was noted as a key concern across the interviews and focus groups. There were concerns raised regarding the supervisor or ME holding previously formed opinions regarding the registrar's competency and safety. This may be opinions formulated from a personal interaction with the registrar or from previous experience with the registrar.

> *Supervisor: "I think it would be very challenging. I think it is challenging to fail people when you know them. There's also a sense of I have failed to remediate them over the whatever time I've had to teach them and for them to learn."*

General practitioners typically show personality traits of being caring and considerate, which may influence how they deliver feedback and how they assess a registrar. Comments from the interviews clearly demonstrated that some supervisors struggle to give negative feedback and if they had to decide on passing or failing a registrar, they would more likely give the registrar the benefit of the doubt and pass them.

> *Supervisor: "We're all too nice. We're - doctors and GPs are meant to be nice. We're trying to be nice to people all day. I think to try - then to take on an assessment role is difficult, and if it means that you're - this is not good enough, we tend to not want to say that."*

Written feedback given to registrars was discussed by participants as being something challenging, particularly if there were concerns around giving negative feedback and potentially upsetting the registrar. This may results in inaccurate written feedback and tendency to 'fail to fail'.

> *Supervisor: "If you've got a real concern, you go and talk to them. You are not going to put it on an assessment form."*

It is important that supervisors and MEs continue to have training in delivering feedback, to reduce the concerns around failure to fail. Registrars emphasised in the focus groups their desire for honest feedback and had a genuine desire to hear negative feedback, if it could result in an improvement in their performance. In addition, registrars found it unhelpful receiving benchmark comments, instead wanting specific feedback, rather than being told they were at the expected level.

> *Registrars: "They may be potentially marking you kinder and then you get to an actual exam and you feel like you're underprepared because you've been given this whole sense of security."*

> *ME: "I find that people are very kind on the workplace based assessments. I've just recently - and again this is anecdotal - but I've just recently had a registrar that we discovered right towards the end of her training, the supervisor phoned me and said she's only seeing two patients an hour and she's asking about nearly every single patient. She's passed her exams, she's done everything and after digging a little deeper we discovered that the person that had her in GPT1 thought exactly the same."*
>
> *Registrar: "I think if your supervisor has significant concerns about your readiness to be a fellow that should be taken seriously. Do I think that everyone's necessary going to be brave enough? No."*

Another point that was raised in the interviews was around the subsequent fallout that may occur if a supervisor failed a registrar or flagged a registrar as needing additional supports, particularly if that supervisor is also the employer of the registrar.

> *Registrar: "If your supervisor is failing you there would be a conflict afterwards to continue to employ you as well."*
>
> *ME: "I think the relationship is something that potentially can get in the way of a true assessment and introduces significant biases."*

Bias and potential conflicts around failure to fail have been raised as serious concerns, emphasising the importance of collecting multiple opinions in order to obtain a more realistic picture of the safety and competency of the registrar.

> *Supervisor: "As a supervisor you can't always be completely objective and you've got a relationship with the doctor that you're supervising. Whereas if you're examining someone you usually come in and you're quite objective so you can fail someone. I would find it very hard to fail my registrar. I couldn't be objective I don't think completely."*
>
> *ME: "I think that all RTOs are probably screaming, please tell us the truth. Tell us early, tell us often. There is always that reluctance."*

Assessor variability

Assessor variability is an important component when undertaking assessments, ensuring objectivity and validity. This was a key theme that arose from the interviews and focus groups. Contextual variability was also reported, highlighting the importance of the various roles that MEs, supervisors, practice managers and the environment all play, and how they interrelate.

> *Registrar: "I think because you're so isolated as a registrar sometimes in GP. It's just you and the patient. It's nice to have different people observe you and give their tips and their tricks and how they might do it. Just get those different perspectives."*

Consistency and objectivity were deemed important when it came to undertaking an assessment. Registrars, supervisors and medical educators reported both positives and negatives of having different assessors undertaking assessments. Overall there was a desire for both the same assessor and also different assessors, as both delivered different objectives. Having the same assessor allowed for continuity and reflection of progress.

> *Registrar: "It's very useful if the same assessor or the same visitor assessed you or observed you the second time around and they see that you... had improved on the things that you."*

> *ME/Supervisor: "So to think about in a practical sense one way of ascertaining some of those nuances and what not would be to potentially have the same either medical educator or supervisor visit a registrar that they've seen early in their training towards the end so that they can actually then make a comment."*

Whereas, a different assessor undertaking an assessment provided new insight and opportunities to identify different areas that were previously unmentioned.

> *Registrar: "I had a different person for every single one of them and I thought the different perspectives they brought and different style of feedback was always helpful."*

Variability was also raised as an important issue when it came to flagging registrars who were underperforming or needed additional support. Variability allows the registrars to also gain multiple perspectives.

> *ME: "We've certainly had instances where a registrar in one practice is given glowing reports and they move to another practice and then their supervisor contacts us with serious concerns about their abilities. So it's huge variability and some of that comes down to training and experience of the person completing them"*

> *Registrar: "So what my supervisor thinks is important is different from what my medical educator thinks is important. Getting two different opinions just gives you a bit more of an idea about where you should be pitching your practice."*

The role of WBAs

Participants were asked their opinions regarding workplace-based assessment and their thoughts about the utility and role of WBAs in the general practice training program. Overwhelmingly, all stakeholder groups supported the use of in-practice assessments as part of specialist general practice training, with discussions occurring as to how workplace-based assessment could be strengthened and further incorporated into the current training model.

> *Registrar: "So I think a workplace based assessment if done correctly, could factor in practically what it's like practising in your environment at the time."*

The concept of assessing consults that could not be assessed in an exam setting was raised. The real-life empathy of breaking bad news, unrealistic patient expectations, patients who required interpreters and time management for complex presentations were all raised as common encounters, but ones for which there may be no formal assessment. These can be the most challenging of consults, but also the most insightful when observing as to how a registrar manages the situation. Assessing registrars in their environment, seeing how they utilise the tools at their disposal, how they interact with other staff members, and how they greet new and existing patients, provides useful information as to how a registrar is progressing.

In addition, there was a strong sense that training should be tailored to the registrar, that there may be a need for additional assessments to be undertaken prior to deeming a registrar safe and competent for independent practice.

> *ME: "It needs to be tailored to the registrar."*

Registrar: "In my mind they're kind of looking at you from different perspectives trying to get an overall picture of your competency."

ME: "I think workplace-based assessments make sense really, because you're going to get a much better idea about what registrars are doing day-to-day. If you can have as many different sources of information feeding into that assessment as possible. So, I think a workplace-based assessment, programmatic assessment, is a really good thing."

Direct Observation

Direct observation is the key component of workplace-based assessment. The ability to assess the registrar in action, consulting with patients, is essential when it comes to training a safe and competent registrar.

ME: "With the direct observation you're talking about body language, you're talking about pacing, you're talking about organisation of the interview. You're talking about rapport, people skills."

It became apparent throughout the focus groups and interviews that RTOs were undertaking various forms of observation, with every RTO offering a different approach to observational visits. Types of direct observation included video review of consultations, external clinical teaching visits, assigned ME visits and supervisor observation. These subtypes will be explored later in more detail.

ME: "I am trying to concentrate more on the five domains, rather than just clinical skills...but it's more attitudes and ethics and yeah, communication skills, these sorts of things."

Supervisor: "There's a whole lot of things that it needs to assess, like referral patterns, the quality of referrals, the reflection on difficult patients or challenging cases, those sorts of things. I just think the exam is a tick, tick, tick type situation"

The unpredictability of general practice and the ability to observe the registrar in action, on their feet, provides an uncompromised view of their capability.

ME/Supervisor: "It was a really valuable observation, because it was not predictable and it was - there was uncertainty… it was a much better assessment, more valuable assessment than it would have been if they were all straight-forward consults".

Registrars saw the observational visit as a time for education, developing their learning needs and further refining their general practice skill-set.

Registrar: "It's actually very useful because I would rather make mistakes during those visits and learn from them after it during the feedback."

Registrar: "I think it's a very good way of assessing, but it's probably a very good way of assessing bad - like you know dangerous practice or catching dangerous situations than perhaps assessing the competency or how good you are in a way."

While there were many positives to direct observation, there were concerns raised regarding whether direct observation provided adequate insight into the full scope of general practice and how this might be incorporated as part of a barrier examination. The ability to 'game' the

consultation and organise pre-planned patients was raised as a concern by both supervisors and medical educators.

> *ME/Supervisor: "I think sometimes I feel like with the direct observations they're staged. They've got five patients who have called up and say I just want a script or a referral. What's the point of me sitting in here? This is a low value consult. I can't see what this registrar is made of here."*

Direct observation provides greater insight into the way a registrar operates in the consulting room, unveiling concepts that previously might have been hidden. In addition, safety of the registrar in order to practice independently is paramount.

> *Supervisor: "There's been a case of someone actually opening up about an abuse situation and the registrar was too busy focusing on the other things that she completely missed it. Then later on I'm like did you notice what she said about this and she's like, oh, like I didn't even realise."*

> *ME: "Yeah and if I'm not in the room no-one is going to know. No-one is going to be able to assess that because no-one is aware of it."*

Direct observation by supervisor

The use of direct observation by the supervisor appears to be under-utilised as per the feedback from the focus groups. Registrars raised concerns over the lack of direct observation by their supervisor, especially when commencing new placements after the first 12 months of training. Registrars consistently made comments around their desire to have a greater number of observations undertaken by their supervisor to aid learning, particularly in the lead-up to exams, and also to provide feedback on their progress. Medical educators noted there was a disparity amongst supervisors who undertook observation visits.

Supervisors cited both positives and negatives regarding the use of direct observation visits. In regards to the positives, it was seen as a way of protecting their business and ensuring that their registrar was safe and competent, and to help aid in their understanding of the level of supervision that the registrar required.

> *Supervisor: "We've got a business and so there is that to uphold. Because you don't know what happens behind closed doors. You get a sense when you sit down with them."*

> *Supervisor: "Across the practice, you know the worth of your registrar and their knowledge. You pick that up very quickly as well, both in the consult and how they interact across the practice."*

It was surprising to learn that registrars often are not observed at all during their consultations by their supervisor. Thus it was concerning how a supervisor could gain an understanding of the level of supervision the registrar needed without adequate observation of their consultations.

> *Supervisor: "I almost wonder if that should be a mandatory thing because I think it's very useful for our Registrars to have us sit in with them and I think it's a bit uncomfortable for them sometimes and bit uncomfortable for us sometime."*

> *Registrar: "I think once you get to term three, I just struggle to see how your supervisor could assess you, because they just don't...see you work at all."*

Regarding the negatives of direct observation visits, supervisors reported the financial disincentives of observing their registrar's consultations.

> *Supervisor: "I think it has to be funded, because it's impossible to say to the reception staff, I'm just going to sit in with the registrar. It does tend to make you a little bit grumpy that you're missing out on seeing your own patients."*

> *ME: "Supervisors are doing that but it doesn't happen a lot because it's so time consuming and it's so cost ineffective in some ways."*

Supervisors also reported that observing their registrar was often a positive learning experience for themselves as well.

> *Supervisor: "They may have a better way of doing things than me."*

Medical educator direct observation visits

Registrars consistently emphasised their desire for ME direct observation visits, reporting that they were an extremely positive and helpful experience. Some registrars reported initial anxiety prior to the visits when being observed, however these registrars reported how useful these encounters were for their learning and progress.

> *Registrar: "I found them really valuable. I was very nervous about them but I found them really, really good and I actually look forward to them now, which is kind of weird."*

> *Registrar: "Yes, you see the supervisor from time to time. But I felt like the ECTs were such a breath of fresh air, of extra perspective, of extra training. I really wish there was more of that."*

MEs viewed the direct observation visits as a positive exercise where they could assess the registrar on the five domains and see first-hand how they incorporated all the elements into their consultation. The ability to also undertake case-based discussions and random case analysis as part of these sessions was described by many as providing new insights.

> *ME: "Where I see the value potentially of clinical teaching visits is actually you've got a real scenario, a real patient and that's where the communication skills, you can really talk about those in the real world and not just trying to get through an eight minute or a 19 minute station."*

The experience and additional training of the ME was apparent in the registrar's desire for high-quality teaching and feedback.

> *Registrar: "I think they had a more objective approach because they've been doing this as part of their role. They do a lot of them. They know what to look out for and they have a big sample size of experience and people that they've seen."*

> *ME: "The ECT visits, they're just gold. I just think they're so valuable. The feedback I get from the registrars is that sometimes they completely change the way they've gone about doing things or the way they've been directing their training. It would be lovely if there could just basically be a lot more, I think, experienced doctors sitting in and giving feedback because it's relevant, it's what they're actually doing."*

Having someone assess the registrar from outside the practice and provide additional perspective, with an element of pastoral care, was seen as useful. This was mentioned on several occasions in the focus groups.

> *ME: "So, there's that little bit of distance. Which means, often, especially for the pastoral care side of things on an ECTV, is they're struggling with the practice or a supervisor or something of the internal working, it's a safe place for them to discuss it."*

> *ME: "It is interesting when you have students who have been in your group for a while, because you get very different information. So sometimes I have been quite surprised when I've sat in on them and gone, oh, I didn't realise that they were good at that or conversely I didn't realise that that would be such an issue for them, because they haven't come across as having an weakness in that area in the workshop itself."*

External clinical teaching visits (ECTV)

External clinical teaching visits were undertaken by a ME or a trained clinical educator. These visits were seen as a powerful educational exercise for the registrar, as well as providing new perspectives. Registrars valued these visits. Some reported nervousness prior to their visit, however, reported the experience to be positive overall. Of particular note was the general desire to have an 'outsider' or unknown educator undertake these visits, with registrars believing that an assessment undertaken as part of an ECTV would provide an objective insight into an average consulting session.

> *Registrar: "It was good to receive feedback from a third party. It's like the semantics of all the consulting. Nobody else sees your body language and your communication and just these real nitty-gritty details that can be hard to pick up."*

> *ME: "You're coming in fresh. You don't have any particular preconceived notions about this registrar or the way they will perform or their knowledge base or anything. You're just taking it as you find it within the consult and their performance there."*

> *Supervisor: "I always find the external person just puts it all in perspective."*

RTOs appear to emphasise the importance of ECTVs and ME direct observation visits early in training, rather than supervisor direct observation, which is not surprising, given that there appears to be minimal emphasis on formalised supervisor direct observation at many RTOs.

> *ME: "The ECTV, I personally find it usually a really good opportunity to look at the non-clinical kind of things, as I said. How they're using the practice system, how they're doing recall, how they're checking results, how they're using the software, looking at their appointment list and seeing if they're being booked appropriately. You do get that extra information which you don't necessarily get in the supervisor direct observation."*

There was less evidence of bias when the observation was undertaken by someone previously unknown to the registrar. Registrars valued an honest opinion and many preferred having clinical teachers or MEs who were unknown to them or their patients.

> *ME: "I believe it's a more accurate assessment, because I go there without knowing the patient, without knowing the registrar, sit there as an external examiner...it's a more accurate assessment."*

Particularly for rural or regional registrars, the ECTV can provide an additional layer of observation when registrars are undertaking hospital-based work. This allows for an assessment of elements of practice that are dynamic and unpredictable.

> *ME: "We find that reports from CT visitors external clinical teaching visitors tend to be more reliable and part of that is that they sit in for a whole session rather than just - again, with the supervisors, there's variability in how much time they actually spend watching the registrar. So sometimes they're just based on what patients I have seen that my registrar has also seen or what have they written in the notes, which as we know, doesn't necessarily give the full picture of the competency of the registrar."*

Training of the clinical teacher around feedback is important to ensure that feedback is timely and constructive. There was a strong emphasis at the RTO level around training and preparedness of the assessor when it came to undertaking an observational visit, ensuring that the information being returned to the RTO was accurate and accountable.

> *Registrar: "They're an excellent source of feedback as long as you can forget that the person's in the room. Sometimes that's hard depending on the personality of the assessor because some of them do interject a little bit and interrupt your flow."*

Contextual observation

Undertaking an observational visit on the registrar is not just about observing the registrar in action, there is also an important element of observing the practice and the supervisor, allowing an outside perspective to be fed back to the registrar, assigned ME and RTO. Observers (clinical teachers, MEs) should liaise with the supervisor after the observation period to discuss the progress of the registrar and the support they are receiving from the RTO.

> *Supervisor: "There is very little face-to-face communication between the training organisation and each individual supervisor. I think probably it's a good idea that this is made to occur and I think the ECT visit is probably one of the best moments for that to occur."*

> *ME: "It's a really good opportunity when you see some I guess less useful supervision approaches to talk to them afterwards and talk about that and ask if they want extra support. We do also run professional development for supervisors too obviously and... encouraging them to come to some additional sessions."*

The need for context when it came to observation was raised as an important issue and something that would be difficult to gain from a video review or from a non-medical training advisor who visited a practice as part of RTO practice accreditation.

When undertaking general practice training in a community-controlled organisation, particularly one that employs culture-specific healthcare workers, such as social workers in refugee health and Aboriginal Health Workers, it is imperative that these workers have the opportunity to provide feedback on their observations of the registrar. It was highlighted in the registrar focus groups how useful it was when a registrar was observed by one of these workers and provided with insight into how culturally appropriate their behaviour may have been, or perhaps how best to have approached a culturally sensitive issue in a consult. In addition, the use of cultural educator visits was raised as beneficial.

Registrar: "During our Aboriginal health term. They come and it's similar to what an ECTV is. They'll come and watch us consult for three or four hours… and give us feedback on how we could improve our cultural awareness and safety."

Direct observation of supervisor by registrar

Registrars reported a desire to observe their supervisors' consultations, with an opportunity to learn a different style of consulting or gain a new perspective on primary care. There was a desire expressed by registrars to learn consulting skills from someone who had potentially decades of experience in general practice, in comparison to their own early stages. Some practices reported they were undertaking 'wave' consulting, where the registrar and supervisor take turns at observing each other's consultations, but overall there appeared to be minimal observation of the supervisor undertaking consults by the registrar.

Registrar: "Observing a GP who has been doing it forever, the consulting skills that you'd learn from that would be awesome. I've thought that would be a great idea lots of times. Never done it though."

ME: "I think it's a way of looking at how people function in the workplace. In our own practice when we have a new registrar we observe them as well, and we also get them to observe us so that they can see someone more senior doing the job."

In addition, there was a desire to spend time with other doctors and members of the team to better understand the delivery of primary care.

Supervisor: "Some days everything goes well and other days every patient seems awfully complicated …so that role model and the dealing with uncertainty is something certainly that you can't read in a book."

Although this is specifically not a workplace-based assessment, registrars saw supervisor observation as a learning opportunity and a way to improve their consultation skills in the clinic. There was a keen desire from registrars to further develop professionalism, conversational skills and time management, all of which are necessary as part of WBA.

ME: "It's really important for them to be able to watch their supervisors at work as well."

Video review direct observation

Video reviews of consultations are undertaken by some of the RTOs. Video reviews were seen as a way to self-reflect on the consultation style and pick up on mannerisms, such as body language, questioning style, typing skills and repetitive words or phrases that the registrar may use. Registrars reported the tool was useful for self-reflection, but found the equipment set-up, consent, unease of camera and the time showing their supervisor, as burdensome.

Registrar: "A lot of people really like the opportunity to do it once and a lot of people say I noticed I do this, or I noticed that I have this habit, or I look at the computer too long, or I look over my glasses in a weird patronising way."

ME: "From a clinical reasoning point of view, I think videos are actually often better to ascertain what thought processes are flapping through their minds than sitting in."

The video reviews necessitate input from the supervisor, which requires a significant time investment from the supervisor for what is considered to be potentially minimal gain.

> *Supervisor: "I have to confess the registrars were never all that enthusiastic and they were never made to do it. So it was always a little bit ad hoc and fairly time consuming because they'd do the consultation and then they'd watch it with you. Then you'd discuss it. One consultation became nearly an hour's worth."*

> *ME: "The registrars hate them and I hated doing them as a registrar. But I think they're really valuable, because I think after the first couple, the patients switch off and the registrars switch off. I do get a real - I do feel like I get a sense of how they can solve and how they work. I think, as much as they're disliked by the registrars, they're quite valuable."*

Logistically, a video review can be significant work for the practice, registrar and the RTO, especially with issues around legality of recording the consult and patient consent, which varies between states. Feedback is often delayed as a result of time gaps between recording and reviewing consults. Rural registrars reported frustrations with the videos and showed strong preferences for real-time observation and feedback in their clinic, where they are often isolated and working independently.

> *ME: "If I had to remove anything, the first thing I would do, is get rid of the video review. Yep. I would go, it's hard for the practices, it's a technology-heavy, it involves consultations that are recorded and you've got the legality of, I think you're allow to destroy them, or you're supposed to keep them, what do you do with that? It's time-consuming. How does the supervisor get paid? How does the registrar get paid?"*

> *Supervisor: "We've tried the video with one of our registrars that we had a lot of sort of negative feedbacks from patients about."*

The ability to self-select and delete consultations that were deemed to be poor, was raised as a concern by the registrars, supervisors and medical educators. The opinion was that it would be difficult to monitor, without a large number of consults and various presentations being portrayed.

> *Registrar: "You can definitely watch them beforehand and delete any dodgy ones"*

> *Registrar: "You've already learned the lessons from watching a video yourself. I don't think necessarily you need to go through it with your supervisor."*

> *ME: "It's actually setting aside time. Sitting down with the registrar. Using it as a full assessment and teaching tool. Then writing up a quality report with good constructive feedback, after you've done that feedback in person. So I think it's the quality of the assessment using the right tool for the right job."*

One of the advantages of a video review was the ability to use the video as an additional tool when it came to remediating registrars. Registrars can gain insight regarding their behaviour in the consult and reflect on their communication skills, professionalism and overall consultation style. The ability to pause the consult and reflect personally and also with others who are viewing the recording, can be extremely beneficial. Stories were shared in the focus groups from both the registrars and supervisors, of registrars being oblivious to their mannerisms, lack of eye-contact, typing concerns and cultural inappropriateness etc. Thus, the overall viewpoint gathered from the

focus groups was that the videoed consultations were a valuable additional tool in the WBA tool kit, however there was a desire to not make it a compulsory WBA due to the barriers of the tool.

> *Supervisor: "We tried the video with one of our registrars that we had a lot of sort of negative feedbacks from patients about."*

> *ME: "If there's serious concerns and patient safety is a major concern, then I think the video reviews are a great way to go. Because you can take the recording; you can have multiple medical educators review it and make a diagnosis and recommendations. The other advantage is you can sit down - so having done that you can sit down with the registrar and as part of your feedback, review the videotaped session together. So I find that a particularly powerful way of doing it."*

Direct observation of procedures

A request that came initially from the registrars as part of the focus groups, was the desire to have direct observation of procedures, enforcing supervisors to sign off on procedures and dedicate time to observing their registrar. This was mentioned by several registrar focus groups who welcomed the concept of a logbook whereby they could record skills and procedures undertaken. Skills requested included skin excisions, wound management, childhood measurements, spirometry, setting up and reading ECGs, and cervical screening, where knowledge from medical school is relied upon and not necessarily revisited in the general practice setting.
The notion of a procedural logbook was supported by registrars, supervisors and MEs.

> *ME: "You have to have some objective criteria, I guess, a tick list of skills that they've completed. You know, has successfully done pap smears, has successfully done Implanon, has successfully done whatever other criteria you want to include from the supervisor's perspective."*

> *Registrar: "I guess if you were to replace - if you decided a whole bunch of clinical skills were really core, you could certainly sign it off like log book style, DOPS style, and that would be a really good way to assess."*

The concept of encouraging active supervision was raised by a number of registrars, who stressed the desire for more hands-on observation, providing supervision to registrars and assisting in their learning needs. The concept of incentivising supervision, with sign-off was appealing to registrars. A procedural log book or signed-off direct observation of procedures also has the ability to identify registrars who may not be completing tasks, that perhaps the supervisor may be unaware, such as cervical screening. For instance one supervisor discovered that their registrar had yet to undertake a cervical screening test by the end of GPT2.

> *Registrar: "It incentivises the process. So if you've got to tick all these boxes and the supervisors know that you've got to tick all the boxes they're more likely to go and seek them out for you. It makes them need to teach you those things."*

Logistics of observational visits

The setup of observational visits requires additional thought, with the need to have a set number of hours or patients booked in. Thus there is a need to guide the practice manager and the registrar about the expectations of the visit. One concern that was raised in the focus groups was the need to have a set period of time for feedback after one or two patients. It is important that this feedback time is allocated and booked in advance, to ensure that these visits provide the most

value to the registrar. In addition, ideally a variety of patients should be booked, so that the registrar is not undertaking one task during this time, such as a flu clinic. Notification to the practice of a ME or clinical teacher visit should occur ideally more than four weeks in advance, so that patients can be notified prior to booking that there will be an observer and to ensure diversity is factored in when booking patients. Assessments should ideally be undertaken in the registrar's usual environment, to ensure familiarity and confidence. This issue was raised in relation to the Objective Structured Clinical Examination (OSCE) where registrars found it difficult consulting without their usual equipment and room layout.

> *ME: "The other downside is I think one CTV there were no patients booked and we were sitting around waiting for patients. That was a little bit awkward. I think I saw two or three patients but the rest of the time we were just sitting and talking. That's the other thing, just logistical issues."*

It was acknowledged from an assessment perspective, that it would be ideal to have a variety of different presentations, however the observational visits provide significant insight, with a snap-snot of a usual consulting session. Time should also be allocated at the end of the session for the observer to meet with both the practice manager and supervisor, so that mutual concerns, feedback or registrar expectations can be discussed. In addition, consideration should be given to random case analysis as part of these visits and could be undertaken prior to consulting with patients, afterwards, or if there's a patient who does not attend.

Feedback related to direct observational visits

How feedback is delivered to registrars as part of the observational visit was raised as a priority by registrars. There is a need to look at the style of feedback and how it is delivered to registrars to ensure that feedback is professional, non-judgemental, meaningful and actionable. Feedback will be discussed later in this report.

> *Registrar: "I think people with slightly less emotional intelligence aren't particularly good at giving constructive criticism and can sometimes just give criticism. Whereas others might do the complete opposite and tell you how fantastic a job you've done and not give you anything valuable. Then the good ones are somewhere in between."*

Safety Assessment

An interesting concept that was raised by the supervisors and MEs was the need to have a 'safety assessment' when registrars commence consulting in community general practice. Some RTOs already undertake a safety assessment, while other supervisors revealed that they created their own form of safety assessment in the practice, establishing the confidence and level of supervision required by the registrar required as early as possible in their placement.

Emerging from the interviews and focus groups, was the need for the safety assessment to include supervisor observation, assigned ME observation, entrustable professional activities or a similar rating system to establish level of supervision required, multi-source feedback and the commencement of learning goals or a learning log.

> *Supervisor: "Basic GPT1 registrars, they especially come from the hospital system, how they - you know, their general practice, how's their style of practicing, how they approach the patients, how they approach the problems, how they come to the conclusion and make sure they are safe practice."*

When new registrars first enter primary care it is imperative that patient safety is upheld to the highest standard and that registrars have close supervision to ensure a sound base on which to begin their general practice careers.

> *Supervisor: "I think it'd be a good chronological review, I could see an early assessment and then every few weeks or every few months."*

> *ME/Supervisor: "Early in the term I think is a really critical one. Because you don't ever know what goes on in that room until you sit there and watch."*

A safety assessment has the potential to identify early 'flags' regarding a registrar's performance and competency, with the opportunity for MEs and their supervisor to plan for extra supports that may be required to assist with registrar development.

Orientation Assessment

An orientation assessment within the first week of commencing at a new practice was highlighted by registrars as a means to ensure they felt safe and confident to practise in a new location. This assessment would form the foundation for the workplace-based assessments that would follow at the practice. When a registrar commences at a practice there is a need to re-establish their level of safety and thus the level of supervision required to ensure that the registrar is safe and competent when consulting with patients. EPAs could be utilised, which would then be reflected upon as part of the mid and end-term supervisor assessments. Orientation should be undertaken in a structured and thorough manner, providing the ideal platform for registrars to grow. This assessment was viewed by participants as different to a safety assessment, with the orientation assessment occurring when the registrar commences at a new practice, regardless of their stage of training.

A number of supervisors reported that they always observed their registrars prior to them seeing patients on their own regardless of stage of training, however, this is not mandated. In addition, supervisors encouraged, as part of orientation, registrars to be exposed to different doctors, nurses and allied health staff within the practice to allow them to develop their own consulting style and understand practice systems and processes. Spending time with various people who are part of the team was a key aspect to ensuring a well-rounded and competent practitioner. This included spending time with the administrative staff understanding the process of Medicare and the healthcare costs to patients, reviewing new patient information and how this was collected, consent forms and the transfer of medical information in a secure manner, all important as part of RACGP standards. However, it appears the majority of supervisors are not undertaking a structured in-practice orientation.

> *Supervsior: "The direct observation, we do that for a couple of days when they start. I actually get them to observe all my doctors, so they can see lots of different styles."*

> *Supervisor: "Well when they start, like we have a theory that every partner should feel that every registrar is competent. We've got two first year registrars who need a bit more but often you only sit in once or twice and you think, oh they're fine."*

As part of the orientation assessment, supervisors also reported that they reviewed patient consult notes, to ensure that record-keeping was adequate and the registrar safe and their practice secure.

> *Supervisor: "One of the things that we've done, but we've done it for the registrars this year is actually review every case every day in the first week and just spend half an hour just - and it's - the registrars really like it because they feel like they're been paying attention. But for us, it's has this person done the right thing? Have they made the right decision, good decisions? What's their decision making ability."*

The orientation assessment, which could be in the form of tick-box, narrative and sign-off, aimed at increasing the confidence of the registrar when commencing. This would allow for further opportunity to learn prior to seeing patients. In addition, there is a desire by registrars and supervisors that registrars learn about the business side of general practice.

> *ME/Supervsior: "What I normally do is get them to drive the computer, and I sit in the corner and run the consult. Basically so that - I find that a much better way of them learning how to use the software than sitting in the corner and watching it, if they're actually doing that. But I think it also gives them a little bit of an idea of structures to a consult and expectations and things like that."*

> *Supervsior: "So back to the initial assessment, I think when they sit in with you, they pick up a style and I really encourage them to develop their own style, to really critique who they're sitting in with and determine what they don't like and what they do like, because they will mould their own way of doing things… The direct observation, we do that for a couple of days when they start. I actually get them to observe all my doctors, so they can see lots of different styles."*

Entrustable Professional Activities (EPAs)

The ability to quantify and establish the level of supervision that a registrar requires is an important part of training. Using EPAs allows supervisors to document the level of supervision they believe a registrar needs in order to be safe over a range of different competencies, and for registrars to reflect on what they believe is their appropriate level of supervision. This can set the standard for the level of supervision that is required to safely and adequately supervise that registrar. EPAs are undertaken as part of the mid and end-term assessments at some RTOs. After individually completing the EPAs, registrars and supervisors spend time together discussing how the ratings of supervisor and registrar compare, allowing for both verbal and written feedback.

> *ME: "That's why I like the EPAs, because it makes you think differently about the practitioner, trust them to go forward."*

> *Registrar: "They certainly stimulated discussion between me and my supervisor about which areas I found more challenging. Really it was a reflection of, you get a presentation, when do you feel uncomfortable and you need to talk to somebody and how urgently do you need to talk somebody and so I felt like it quite easily identified areas that I wasn't - you just asked a question in a way that made it easy for me to kind of cognitively put myself into the picture and go, yes I would want to call my supervisor in that kind of scenario."*

As useful as EPAs were from the viewpoint of registrars, supervisors and MEs, there were concerns about their utility once a registrar was deemed competent to operate unsupervised. The theme of a saturation point emerged with senior registrars reaching a ceiling, whereby globally the registrar was safe for independent practice.

> *ME: "For the average or lower performing registrar it's got a better place, because it does give you structure to provide useful feedback."*

EPAs encourage active supervision, a key concept that emerged from the registrar viewpoint. In addition, they allow for self-reflection, requiring registrars to critically assess what standard they believed they were at, whether they were in GPT1 or GPT4.

> *Registrar: "The vast majority of the time you're practicing no one's watching you do it, so how are they to know we're doing it well."*

> *Supervisor: "It's also quite a learner-centred tool. It's quite good. If you have a registrar that's quite insightful to the gaps in their knowledge then there's more opportunity for them to bring that up in the EPA form"*

An interesting concept that emerged in the focus groups was the importance of addressing and discussing professionalism as part of the mid and end-term assessments, which could be elicited through using EPAs.

> *Supervisor: "Professionalism starts to be assessed immediately. Like do they turn up on time, are they dressed appropriately, do they do their work, do they leave on time. All that kind of stuff. That's part of the EPAs, but I think it actually gets assessed much earlier and in a much broader context."*

Registrars valued the reflection process and seeing their progress in training based on supervision levels. EPAs also guide the feedback process and encourages supervisors to actively discuss the registrar's competencies.

> *Supervisor: "I marked her quite high on most things but there were a few gaps and she'd definitely improved by the time the end of the year came."*

> *Supervisor: "Probably for the average or lower performing registrar it's got a better place, because it does give you structure to provide useful feedback."*

Feedback

How feedback is delivered is an essential component of workplace-based assessment. As part of this stream, participants were asked about their understanding of feedback and how feedback should best be given to ensure a registrar is safe and competent. Key themes emerged regarding feedback – feedback for excellence, specific, honest, contextualising, respectful, timely and outcome-based, with engagement of both assessor and registrar.

Feedback for excellence

Feedback should continue after exams have been completed whilst awaiting Fellowship, to ensure that continuous learning is occurring. It was apparent that the emphasis for many registrars is simply on passing the exams, after which there is a self-reported lull in their educational drive. For many specialty colleges there is often an additional Fellowship or sub-specialty area that is explored after becoming a Fellow of a College. However, this is not so apparent in general practice. It was discussed that feedback could encourage excellence in registrars post-exam to explore areas of interest in more depth, or to consider partaking in medical education training or teaching.

> *Supervisor: "It's actually then starting to talk about the complexities of the consult, how you structure a consult, the business of general practice. There's a whole lot more nuances that come out. A whole pile of stuff that's not in the exam, that is not in a textbook, but is in general practice every day. I think the registrars are now - at that level, I think they're more open to be able to hear that."*

Registrar: “In third term you feel like no-one is actually watching you and you could be doing anything. So getting some feedback and more being able to tailor your own approach would be helpful.”

ME: “One of the things was, we're not just aiming for okay. We're aiming for the registrar, or whoever, to be the best that they can be. Why not give them some feedback? Even it's a little thing that we've found. Give them that feedback so that there's some value to that assessment for them. And it's not just a tick box exercise.”

Contextualising feedback

Feedback needs to be specific and placed into context, to ensure that it is both actionable and relevant. Assessors and registrars need to understand why feedback is important and understand that critical feedback is an essential part of progression of learning.

ME: “I think being explicit about it is important. Saying to the registrar we have to know how you're going. Us monitoring you is not checking up on you because we want to trap you or trick you, trip you up. It's because we want to know how to help you complete your learning plan, for example.”

ME/Supervisor: “I also think for the registrars who are doing well, there's always room for improvement, and those registrars, you don't just want to pat them on the back and say, oh, good job - you're doing well. So when I sit there with a direct observation with one of my registrars who I know, I'll say to them up front - I'll say, look, I'm not being nit-picky, but I am looking for things to help you improve your practice. So if it seems like the things I'm commenting on are trivial and you've got your way of doing it and somebody else has got their way, and I'm providing an alternative or another suggestion, don't take it as me being picky-picky. I'm just trying to provide some suggestions to improve things - from what I'm seeing today.”

Specific

Where possible feedback should be specific to allow for registrars to understand the exact concerns or issue with their performance. For instance, ‘improve communication skills’ may be written on a feedback form, however this does not provide specificities as to what the assessor is referring to with regards to communication. Specific examples should be encouraged and linking feedback to a patient can be a helpful method of achieving this.

Registrar: “One of the most useful things is being able to ask specific feedback, get resources, communication, timing of consults. So you see it as more of a learning exercise more than anything and not seen as an assessment.”

Registrar: “The issue that I often have with feedback is when it's a tick box or written, it just doesn't ever get communicated to you properly or they rush through it and they just - oh, it doesn't really matter. I think it is so important to get that verbal feedback, where they actually discuss it with you.”

Honest

Feedback should be honest. Registrars reported a desire for honest feedback, to help in their progress towards Fellowship. It is imperative that assessors are trained to give honest and thus reliable feedback to registrars, even if the feedback may be negative.

Registrar: "I got a two-faced feeling from one or two of them, where they said something in the room and they're like yeah, this is this and then they wrote things. You're like I don't remember that being a problem at all at the time. It's just funny that some of them will - they'll say one thing."

Respectful

Feedback should be delivered in a respectful manner and not be discussed with other members of the team for whom it is not relevant. As part of the trust relationship between a supervisor and a registrar it is imperative that feedback is delivered in a respectful and sensitive manner, particularly if negative feedback is delivered. Feedback should not be given in front of others, including patients, unless it is specifically relevant to that situation.

Supervisor: "They need to be supported and it should be constructive verbal opportunities, it's not doing something for the sake of doing it. It's something because it enhances their knowledge and it's useful."

Timely

Feedback should be delivered at the time of the encounter or as close as possible to the assessment. This ensures that the assessment is both familiar in the assessor's and registrar's mind and allows for clarification of any pertinent issues.

ME: "You don't want to be spending hours writing that report. I think a lot of the value in feedback is when it's done at the time and it's relevant and they can think about it directly with the patient they've just seen."

Outcome-based

Feedback associated with WBAs should be outcome-based where possible. As well as the purpose of the WBA, that is whether it is for formative or summative purposes, particular outcomes expected of the WBA should be discussed. These might be assessor-centred (e.g. focussing on time management, body language, computer use) or registrar-centred (e.g. use of appropriate guidelines, consultation structure, communication skills, OSCE-style practice). Outcomes should take into account previously identified gaps, the registrar's learning log, the actions from previous WBAs etc. A discussion at the outset should include how feedback will be given and any consequences that might ensue from the WBA, particularly for registrars who have been flagged.

ME/Supervisor: "What helps in the role of an assessor? I think being explicit about it is important. Saying to the registrar we have to know how you're going. Us monitoring you is not checking up on you because we want to trap you or trick you, trip you up. It's because we want to know how to help you complete your learning plan, for example."

Actionable

Feedback needs to be actionable and have the ability to be incorporated into learning plans or goals. A recurrent theme was that registrars wanted feedback that they could utilise, thus it is important that feedback provided leads to an action that is achievable.

ME: "I try really hard to make comments and I write down lots of notes and I quote what the doctor said and I quote the patients back and try and give them some specific feedback, so that they can develop some learning goals from it."

Observation-based
Feedback needs to be observation-based, often with specific examples. This point is supportive of the need for supervisor direct observation in order for supervisors to give adequate feedback to registrars. Registrars valued detailed reports on their consultation observation periods. Although there is no expectation of detailed reports, there is a desire to have specific observation-based feedback on a particular situation, allowing the registrar to reflect on the encounter and improve.

> *Registrar: "Most of them have actually gone through every single patient that I've seen, a description of how I behaved or what happened in the consult, the good things, the bad things, and how I could have improved on each single patient I've seen. So most of them have been in the room actually typing. So the feedback - or the feedback that I've got on my computer is pages and pages long from each ECT visit which has been really valuable."*

Supervisor engagement
Supervisors gain vital insight into the overall performance of a registrar, thus they are in an ideal position to provide feedback. Supervisors need to be engaged with the importance of feedback and how feedback drives learning and growth. Feedback needs to address the above points, but also needs to be relevant to registrar training and not just about "making the practice more money" as one registrar stated.

> *Registrar: "I think it's tricky when you need an incentive for your supervisor or someone to give you feedback. If things become optional they very quickly fall out of the scope of what's actually going to get done."*

Registrar engagement
Registrars need to be willing to engage in the feedback process, willing to listen and then willing to act on feedback in order to become safe and competent. If registrars don't engage with feedback then this will only result in a disservice towards their progression to becoming a competent and life-long learner.

> *ME: "I find it really hard…it's all so about the relationship and nobody likes negativity and I have got to the end of an ECTV and talked about various sort of consultation skills and made the mistake of saying to a registrar, did you find that useful? To which she turned to me and said, no, we did all that crap in Uni and it just doesn't work for me."*

Case review
The opportunity to undertake review of patient consultation notes, reflection on referral patterns and investigations, and then have subsequent discussions around the management plan, was a key component of reviewing a registrar's level of professionalism.
Intentional review of the notes was deemed useful for both the supervisor and registrar. Case review could either be undertaken as part of random case analysis, remote computer access for those in rural or remote areas, or as part of regular teaching sessions with the registrar.

> *Registrar: "Just to go through the notes and I guess make sure that it was clear what I was thinking, everything had been documented properly, pick up anything that he felt that I'd missed and help me with plans. I found those really useful as well because you find your gaps...You know that you talk to them about safety netting but you didn't document it very well."*

Utilising up-to-date resources and guidelines and thus knowing where to access accurate information online is an important component of modern registrar training.

> *Registrar: "Tying your clinical decision-making to a patient that you've seen helps to retain it in memory. It also really challenges you. Did you follow the guidelines? If you didn't, why? Can you justify it? Would you do the same thing again? If you had that 15 minutes again would you do it differently once you know what the results are? Things like that so really making you reflect on what you're doing."*

Random case analysis

Random case analysis (RCA) was mentioned by the majority of participants as being an exceptionally useful tool for progression and learning. A large number of the stakeholders were already engaged in RCA, all in an informal, formative capacity. RCA was favoured by registrars as an excellent preparation tool for the exams and also as a way to extend thinking in an unpredictable manner. RCA was undertaken in different formats depending on the training of the supervisor or ME. Most commonly it involved changing a detail about a case and adapting a consultation based on different fictitious factors. However, it also involved randomly selecting patients and having a case discussion based on the consultation progress notes. RCA can be a useful tool, as reported in the transcripts, for the over-confident registrar in order to help the registrar to understand their scope of practice and associated weaknesses.

> *Registrar: "Either he'd randomly pick three patients that I'd seen and just go through the notes and I guess make sure that it was clear what I was thinking, everything had been documented properly, pick up anything that he felt that I'd missed and help me with plans and stuff like that. I found those really useful as well because you find your gaps. That's how you find your gaps. You know that you talk to them about safety netting but you didn't document it very well."*

> *Registrar: "I think one of the really useful things and one of the critical things about the random case analysis is taking that one step further and changing one thing about the case and seeing how that changes your approach. So you're changing a presentation from an old person to a young person or a male to a female or something like that. I thought that aligned reasonably well with key features concept of the exam too."*

RCA of the supervisor's patients was also a novel approach to learning. Often RCAs would form part of weekly supervisor teaching sessions. Minimal preparation is required.

> *Supervisor: "Doing it two ways, as well, I encourage them to look at random cases of mine and I look at theirs. But I think that helps to make it sort of level the playing field a bit, because often there are things we could discuss in my consultations as well. Yeah, I've found that really useful."*

Random case analysis as part of the external clinical teaching visits was seen as a way to delve into different scenarios and contemplate the 'what-ifs'. Registrars reported that although the discussions could initially be anxiety-provoking, they found the teaching, learning and reflection a powerful educational tool. In addition, RCA provides an opportunity to review the registrar's consultation notes, test-ordering, time lapse of consult and referrals - all important components of general practice training, which would otherwise not necessarily be reviewed or assessed. There is a need however to ensure that supervisors and MEs are trained correctly in the use of RCA and that registrars understand the context of critical thinking. Engagement of all parties is essential to ensure that tool is utilised to its full potential.

ME: "I guess one of the problems with random case analysis and direct observation is it is a bit dependent on the person who's observing or the person who is doing the random case analysis as to how they delve into the consultation and whether they do pick up on issues. I mean, there's going to be lots of - it's not something that you can standardise because it's going to be completely dependent on what registrars are seeing in practice, but I guess any workplace-based assessment is going to have that element to it anyway."

Multi-Source Feedback – Patient Feedback

Some RTOs mandate collection of patient feedback as part of routine WBAs, whereas others use patient feedback only as part of remediation, particularly if there are concerns regarding professionalism. However the overwhelming comments were that patient feedback was not useful, expensive to collect and required significant resources to collect in a confidential manner. However, informal patient feedback was seen as informative to the supervisor and registrar.

Supervisor: "We rely on what our reception staff noticed - you know, how to assess whether the patient's happy. Our reception staff will tell us very quickly."

ME: "The reception staff get all the feedback from the patients and the feedback from the patients is all, oh, they're lovely - they're really friendly, they're really nice - but their time management is appalling."

Multi-Source Feedback – Staff Feedback

Multi-source feedback from staff members was seen as a valuable tool when it came to both assessing the registrar and providing feedback and reflection for growth. As part of this process, a specific question about patient satisfaction should be asked, such as, "generally do you think patients are satisfied with the registrar". MSF needs to be electronically submitted through an online portal, in order to protect the integrity and confidentiality of the assessor.

Supervisor: "Feedback from the nursing staff and the front desk staff has been incredibly useful because being a GP is about much more than just being a good clinical GP, it's a lot about working with the team to get the best outcome for the patient. So you can't really be a solo person for the best outcome sometimes."

MSF appears to be most useful in the early stages of a registrar commencing practice and in registrars who are flagged or have professionalism concerns.

ME: "I think once we see that a multisource feedback is very positive then it's unlikely to be of continuing benefit throughout the rest of training."

Supervisor: "I do like the multisource feedback because - well, certainly in my clinic, the receptionists, they just say what they actually think. You get fascinating responses."

Feedback from various staff members is important for formulating a global picture of how the registrar operates in a team environment. It should be noted that cultural feedback, as part of their Aboriginal health and refugee health placements, including feedback from Aboriginal Health Workers was appreciated.

Registrar: "I guess in the last one I had, there were a few things that came up that had never been mentioned to me in person and I actually did find that last one helpful."

> *ME: "It does pick up some professional behaviour, that some doctors may not realise that they do. Some of them can be dreadfully rude, probably the only way to put it. They just have this superiority complex that if you really want to get on and work in a wonderful practice, it's not going to suit you well if you treat your staff really badly. So, they're very useful for that. Some of them have been really shocked and hadn't realised and have actually responded to the feedback and didn't realise they were coming across that way."*

Patient Encounter Tracking and Learning (PETAL)

Patient Encounter Tracking and Learning (PETAL) allows registrars to sequentially record their patient encounters, investigations, diagnosis, management and follow-up plans. It is used across a number of RTOs for research purposes, but has the additional value of tracking areas of patient encounters and map the expected presentations against cohorts. It is a valuable reflection tool and allows for review of areas of potential deficiencies.

> *Registrar: "I love the data capture… I like the fact that you can see how you change over time."*

The information needs to be relevant to the learning needs of the registrar and not seen purely as a data-collecting exercise. The ability to reflect on the scope of patients seen is important when it comes to determining whether a registrar is observing the full spectrum of general practice presentations.

> *Registrar: "To be honest I don't know how that fits within the workplace assessment spectrum. I feel like that's more data gathering."*

PETAL also provides value when it comes to reflecting on patient subtypes being seen in general practice, such as paediatric patients. It offers the potential to track paediatric patient numbers, with regards to meeting College paediatric requirements.

> *ME: "So if you're talking about paediatrics it's really nice for them to know how much paediatrics they actually see and what they see in paediatrics as well, so that they can contextualise the information that we're giving them."*

Striving for excellence

There is a need to aim for excellence when it comes to general practice training, in order to produce high quality practitioners with specialist qualifications. It was recognised from the stakeholders that there is a need to have a high standard of training, and concerns were raised regarding whether the training program in its current state strives enough for excellence. There appears to be a significant focus on flagging and remediation, however less on the excellent registrar. It was reported by some in the focus groups that their RTO flagged registrars who were excellent as a way of identifying those registrars who excelled. One supervisor mentioned how their registrar enrolled in an online course in medical education and they completed the modules together, as a way of stretching the registrar's knowledge.

> *ME: "Be competent everywhere, but be super-competent in other areas and almost a mastery of practice in our own little speciality, or just overall keep lifting the bar. Don't just accept satisfactory. We all need to be excellent, excellent, excellent... So, I think, if you put the effort into it, you can help a very good registrar to become even better. But you, as the*

supervisor, have to be prepared to sit down and give them the feedback and give them some information."

The concept of a final sign-off assessment at the end of training was raised as a potential option to ensure that registrars were at a standard worthy of Fellowship. An ME from one RTO reported that it was difficult to fail a registrar with professionalism and safety concerns if they had passed their College exams. There was no other hurdle for the registrar to jump and as such they obtained Fellowship despite the RTO's misgivings. The concept of a statement of awarded responsibility (STAR) from the RTO, related to the ability of the registrar to work independently, was viewed as a means to ensure a practitioner was safe and competent, with the ultimate responsibility belonging to the RTO for a final sign-off before Fellowship is granted.

ME: " I think if I was able to look into my idealistic future and say what do I think training programs should be, it would be what I started off with, which was trying to train the best people to look after me in my old age."

Regional training organisation structure and support

When looking at the utility of WBAs it is important to reflect on the RTO structure and supports and how these interrelate with assessments in the workplace. There was a reported variability amongst the RTOs with regards to who would maintain accountability for completion and collation of WBAs. At some RTOs there was a significant administrative workload required of the MEs when it came to collating WBAs and signing off on College requirements, which was reported as burdensome at times. MEs were keen to focus their efforts on education and assessment, rather than on administrative tasks.

ME: "It's hard to realise the intricacies of the work that the RTO provides, because there is a lot. It's incredibly complex what they do. It's incredibly labour intensive too. You think how hard it could be to organise all of this, but there is a lot of organisation in making it all happen."

RTOs are thought of the as the backbone for the current training model, providing support, both educationally and administratively to registrars, supervisors and MEs. RTOs play an essential role in pastoral care and registrar welfare. Administrative personnel in the RTO, such as those of non-medical training advisors and education support workers, work closely with registrars and build supportive relationships throughout training. The pastoral care role is essential, and the administrative team monitors registrars through their completion of assessments, amount of communication or lack of communication, and overall compliance with the training program.

ME: "One of the big issues that we strike is making sure people are happy in their placements and that the placements are working for them. Most of the time, that happens but probably not always."

When external MEs undertake observational visits, the ME is not only assessing the registrar, but also the context, including the supervisor and practice. This information is often fed back in an informal manner to the RTO and can aid in assisting the supervisor or clinic if there are concerns raised.

Superisor: "I think there is a need to also find out how supervisors are performing and I think this is one of the places you could find out."

> *ME: "There is a lot of academic pedagogy that backs up all of the teaching that they do and all the supporting that they do. There are very, very formalised roles and IT processes in place."*

A point that was raised in the focus groups was the important role that the RTO plays in continuity and establishing links with the practice. There were comments made around fragile and damaged practice relationships during the transition to a new RTO a few years ago.

> *ME: "One of the big issues that we strike is making sure people are happy in their placements and that the placements are working for them."*

The RTO is also responsible for the training of the assessor, a key requirement to ensure that training standards are maintained. RTOs provide education and training for all stakeholders in GP training. There appears to be a significant overlap in these education programs and it was evident that further collaboration between the RTOs would help to standardise training of the trainers.

> *Registrar: "I think that for a lot of the issues that arise in the imbalance is the massive heterogeneity between supervisors' skills."*

Medical educator training

The RTO is responsible for employing and training MEs. There is a need to ensure firstly, that MEs are familiar with GP training and education and secondly, that they truly value training the next generation of primary care doctors. MEs should be hand-picked by the RTO and ideally be familiar with direct observation visits as part of external clinical teaching visits.

> *ME: "Nearly all of our medical educators were taken from our pool of ECT visitors. So, I think that's almost a GPTQ policy. That is how we recruit our medical educators, is we go to the pool of ECT visitors and then interviews are done."*

External Clinical Teaching Visits (ECTV) training

In order to undertake an ECTV, assessors must be trained to a high standard. This is of particular importance with regards to programmatic assessment and ensuring that the opinion formulated is of high value. In addition, clinical teachers need to ensure that they can confidently deliver feedback verbally and in written form, to ensure that registrars can action feedback and progress.

Supervisor training

Supervisor training is mandated by the RTOs, with the majority requiring a minimum of biannual attendance at workshops. Supervisors valued the education provided by the RTO and learning about what other supervisors were undertaking in their practice, including new ways of assessing, such as through RCA or mock virtual patients.

> *ME: "Training the supervisor too and focussing as much on the supervisor, if not more than the registrar, because I think that's where they get most of their information on. That's where the real work goes and that's where the real training happens, it's the day to day stuff in the practice. Actually I think standards for that need to change as well."*

Supervisors expressed that one of the highlights of face-to-face training was the ability to network with other supervisors, describing the environment as collegial.

> *Supervisor: "I think the quality of the workshops that I've been to has been mostly very good, I've left there feeling enthused, feeling that I know the people within the organisation that I would be approaching."*
>
> *ME: "I think that you can't educate supervisors enough. To the point where every single time we interact with supervisors, we mention direct observation. Every single time."*

Supervisors valued learning further about trouble-shooting for the difficult registrar, which appears to be an ongoing concern for supervisors. In addition, WBAs and their purpose needs to be emphasised, to ensure that there is supervisor buy-in.

> *Supervisor: "We need to be educated about why we're being asked to use certain tools."*

Junior supervisors were discussed as well, and the opportunity perhaps for a 'buddy' system to ensure that inexperienced supervisors had additional supports when it came to providing feedback and a safe environment for registrars.

> *ME: "I'd also wonder about like kind of a buddy system for a new supervisor so that if they wanted to - if they had questions, to have a buddy of more a more experienced supervisor. So they've got someone to ring, talk to, ask how do you do this, how do you do that. Rather than trying to have to re-invent the wheel for themselves."*

Registrar support and training

Stakeholders expressed confusion at times over the requirements of WBAs and not understanding the role that these assessments played in their training. It is thus essential that the RTO educates registrars with regards to WBAs and how best to utilise the assessment.

> *Registrar: "It was almost like thrown into the deep end and the CTV visitor had to educate and just prime me a little bit on what was happening. After the first one I kind of knew what was the go. It might have been good to maybe watch a video and just see some examples."*

RTOs are responsible for delivering the education program to registrars through administrators, clinical teachers and medical educators. There appears to be no standardisation regarding the teaching program offered, however it was apparent from the focus groups that registrars valued both the education and the networking opportunities with other registrars offered as part of face-to-face teaching. Overall there was a desire from registrars for training in the utility of WBAs to be strengthened.

Remuneration

There is an expectation to be remunerated for teaching and supervision of general practice registrars. One of the challenges as mentioned in regards to the role of the supervisor, is the concerns around remuneration, and supervisors were often unhappy with the current remuneration offered by RTOs. Participants from four different RTOs specifically discussed the need to remunerate appropriately for education and supervision within the practice. Thus, there is a need for RTOs to review how they are funding supervisors and whether there is a potential for RTOs to increase or change their method of funding practices.

> *Supervisor: "When you are sitting here just losing money doing this, doing it for the good of the country and the good of the community and all this sort of stuff, it wears a bit thin after a while."*

In addition, there were discussions around funding practices for education and training of more senior registrars, for whom there is little, or often no teaching payment.

> *Supervisor: "I guess a bugbear as a supervisor is the lack of teaching payment for more senior registrars. It's kind of like just - I mean, you do do it and it's kind of like you do it for gratis. But it would be nice to just be acknowledged for that rather than just having it assumed."*

> *Supervsior: "If you're getting serious about education and feedback and so on, you got to pay people for what they do."*

In addition, registrars who are underperforming and/or flagged, often require significant supervisor investment that is usually unfunded.

> *ME: If you've got a registrar that you know is struggling but they're not quite on the remediation pathway and you know that supervisors are going to need to be putting in extra time, a lot of that at the moment is done on goodwill, from what I can see."*

Paperwork and assessments

An overwhelming theme to emerge across the stakeholder groups was the burden of technology, and the issues associated with systems that were archaic and non-functional. Users reported that the technology often resulted in increased time being spent on assessment due to platforms being slow and lacking integration.

In addition, there were reports around confusion of the requirements regarding completing forms. Some supervisors would provide minimal narrative, others would provide more specific feedback. The amount of text was often reflective of the type of form, with many reporting they like simple forms with the opportunity to provide written text.

Clarity is needed around the purpose of the form. Specific training should occur at the both registrar and supervisor level as to why the form exists and the underlying intention of the assessment.

> Registrar: "*I think the paperwork associated with doing those assessments was quite cumbersome when we're trying to get consistency often a degree of red tape and paperwork comes along with an already very pushed supervisor".*

Explanation about what happens with the forms once submitted, whether they form part of the assessment portfolio, who will review the forms and how the information will be utilised, should be communicated to both the assessor and the registrar prior to completing the form.

> *ME: "At the moment on the form it just says do you have any concerns about this registrar's professionalism, tick yes/no, that's all. It doesn't say - there's nothing concrete there. There's no guidance document that comes with that. There's no articulation of what constitutes concerns about professionalism and what evidence is required. I think that would be really useful."*

> *Supervisor: "If you asked us to say exactly what these forms are, I can't tell you. We just fill out what they want."*

Number and Quality of assessments

There was no evidence from the focus groups to suggest that the number of assessments was burdensome on the registrars or supervisors. Registrars valued external clinical teaching vists and reflected positively on observation of their consults.

> *Supervisor: "Personally I don't think that's too much for a registrar in training, to be put under the microscope once a month."*

Registrars and MEs both raised concerns around the quality of assessments and the associated feedback being delivered. To ensure quality, there is a need to ensure that the feedback provided is detailed enough to allow registrars to develop their learning log and reflect on the feedback provided.

> *ME: "They'll often write things - boring things - like, with time will develop this skill, or whatever. With more experience. You think, well, that's really useless - because we know that's true for everybody."*

> *ME: "Some of the supervisors feel like it's all a bit too difficult and they - you know, so it's just tick, tick, tick, okay now it's done. It's off my list of things to do, without having that level of quality but some supervisors give extremely insightful and good quality feedback."*

Tick-box vs narrative

As part of progression and learning there is a need to provide information-rich feedback to registrars, which can be measurable, specific and actionable. It was evident across all stakeholder groups that there was a strong desire for narrative feedback in conjunction with tick-box standards. Registrars wanted meaningful feedback and didn't want a tick-box mentality when it came to assessing their progress in the training program.

> *ME: "A number or a tick in a box doesn't tell me the why, as to what that rationale is behind that score or that ranking, but the description will be able to give me that sense of, is this a problem with their competency."*

> *Supervisor: "My particular feelings about them is that I think that they're a bit of box ticking, myself. There's not enough free writing done which I think is far more important."*

Medical educators valued high-quality narrative feedback, which gave more insight into a registrar's competency and concerns. MEs were critical of supervisors who didn't engage with written feedback and discussed the need for written and verbal narrative feedback to the registrar from supervisors. Some registrars reported how they would reflect on their last ECTV report prior to their next ECTV, to allow for reflection on their progress and to ensure they had actioned the feedback. MEs who collate opinions for programmatic assessment reported a desire for feedback that was relevant and detailed, preferring detail over a tick-box.

> *ME: "The things I don't like is when I don't get those verbal comments. So when I just get numbers on a page or ticks on a box and then I'm left scratching my head, going, well, what exactly do they mean by that?"*

> *ME: "I don't find the formative assessments, which are done by the supervisors, necessarily helpful. I think a lot of the supervisors just tick the boxes and just yeah, yeah, I've got to tick all these boxes and that's it. Because you seldom ever see a not-so-well-performing registrar*

on their report. It's very unusual, I think they just tick the boxes and don't look at it very carefully."

Exams

In order to understand the role of WBAs, one of the key aims of this research was to establish attitudes towards the current RACGP exam process. In addition, there was a need to gain perspectives from the registrars, supervisors and MEs regarding the role of written and OSCE-based examination when it came to producing safe and competent registrars. The concept of whether passing an exam translated into an independent general practitioner who was a safe and competent life-long learner was dissected. It was clear from the participants that there needs to be a greater focus on the day-to-day consulting skills of the registrar, in addition to exams.

ME: "I think we can't assume just because they've passed the exam that they're actually fit for practice."

Written exams

There was an overwhelmingly positive attitude towards the written exams from nearly every participant, including the Key Feature Problem (KFP) examination, which addresses critical thinking, a key component in general practice consultations.

Registrars did not believe overall that the KFP was reflective of general practice, however were aware of its importance to ensure a high quality, critically-thinking registrar.

Supervisor: "I think there's an element of devaluing us as compared to other colleges as well, if we are a college that abandons an exam based assessment, no other college is looking at that. I mean I agree, I can see the benefits in doing programmatic assessment with other more global forms of assessment. But I am, yeah, 100 per cent opposed to I guess, yeah, not having an examination of any sort, it just seems wrong to me." [S]

Written exams were seen as a key driver to gaining further knowledge. The ability to look up information during the consult has changed the reliance on learned knowledge in general practice, however clinicians need to be able to think critically to decipher the history, examine, and solve a patient's issues. These are essential skills for general practice despite some claims that the internet and big data are moving in this direction.

Registrar: "The whole process was absolutely painful but retrospectively…three months of buckling down and studying and learning something about everything is actually a useful thing today because the way I practice today has changed."

Supervisor: "Because basically the assessment drives the learning doesn't it. So you've got to have something to drive the registrars to learn."

An overwhelming theme emerged of the need to have both written and 'in-practice' assessments to adequately assess trainees.

Registrar: "There needs to be some independence but I think exams are too blunt a tool… I'm one of those people that suffers as a result of that sudden do or die. Whereas if you gave me 20 CTVs I perform quite well in them."

Supervisor: "I think it assesses them in the actual working environment whereas a test exam, whether it's an OSCE or a multiple choice, it's an artificial environment and I think to see how

> *someone interacts with patients and sometimes things will get a bit - things won't go as planned and so how to deal with things as they come up, how to interact with different personalities, how to deal with things that become somewhat emergent or urgent. I think those are all things that'll show what kind of doctor they are. More so than a written test or a staged test."*

In addition, a concept of credibility was raised by a number of groups and individuals, regarding the importance of a knowledge-based written exam to ensure a high level of knowledge amongst registrars, when it came to achieving a specialist qualification.

> *ME: "We need some credibility in the medical world and I think if we get rid of our exams we show less credibility in comparison to other specialties."*
>
> *Registrar: "I would worry about having less rigorous assessment practices."*

Objective Structured Clinical Examination

The OSCE, which can be sat after passing both written exams, didn't translate into the same level of satisfaction from participants as the written exams. There was a desire from a small subset of individuals who believed the OSCE was a necessary part of the exam portfolio, who believed the OSCE provided a standardised approach to assessing communication skills and professionalism, with set cases in breaking bad news, Aboriginal health and ethical issues. However, registrars saw the exam as a tick-box exercise, displaying practised empathy and elaborate examinations that were not reflective of their usual consultation style.

> *Registrar: "The OSCE's a bit of a driving test to put it bluntly. It's a good test. In terms of clinical medicine, I think it's great and it's valuable but you do have to perform to a certain extent."*
>
> *ME: "An OSCE is - what's the word - it's pretending - a pretend clinical assessment."*

There were both positive and negative feelings towards the OSCE. There was a strong desire for 'real-world' medicine and the need to assess the competency of the registrar within the practice. Although the OSCE allows for standardisation, there were concerns raised regarding how necessary the OSCE might be if the programmatic curriculum was strengthened across the RTOs.

> *ME: "Certainly, it seems that there are not many people that fail the OSCE that haven't failed another exam segment first. So, I can certainly see why there's talk about is the OSCE even worthwhile. I think if there was no OSCE, then there's got to be some other observation component in training to replace that."*
>
> *Registrar: "I think if you're comparing an OSCE versus a workplace assessment, I think this is better because you're in your environment."*

A strong preference emerged for WBAs over the OSCE, to allow for real-world scenarios to play out and offer the potential for evolving presentations to be observed. Seeing how a registrar interacts with patients, how they might approach a patient who mentions chest pain as they leave the consulting room and how they deliver bad news to a patient with real emotions is a true reflection of the professionalism and communication skills of the registrar.

> *Supervisor: "I think exams can be very artificial. The trouble is, in an exam where you've got a scenario you're trying to put yourself into the examiner's head as to what they're trying to do.*

What's the patient trying to tell you? Whereas if you've got a real patient there it's usually much more obvious how the flow goes and what things you need to rule out and rule in. Because you've got a person who might be breathless or might be pale or whatever. You certainly don't get the end of the bed look at somebody in an exam situation that gives you a few clues."

Registrar: "I'd appreciate that for someone to judge my practice and go yeah, you're worthy of fellowship based on the way that you see patients. It's real life as well. It's probably the best way we should be assessed within clinical practice."

Summative and formative

There appeared to be confusion in the focus groups on the use of the terms summative and formative and how they relate to WBAs. The concept of an assessment counting towards gaining Fellowship would indicate an assessment is summative, whereas formative would indicate that the assessment has no weighting when it came to awarding Fellowship.

Registrars valued formative ECTV and preferred a low-stakes approach to direct observation, using the visits as a learning experience. This is despite the reality that ECTVs are used as a 'flagging' tool and as such may be viewed as a summative assessment by some individuals, as flagging has the potential to influence the training pathway and potential outcomes.

Registrar: "One of the nicest things about it is it isn't formally assessed, despite the feedback. Is that something that you feel that if you have someone sitting there giving you feedback and it's sort of relaxed environment, that you tend to perform better or feel more comfortable seeking feedback? Certainly it's much more comfortable and having the ECTs assessed will change the tone of it dramatically."

There is a request for clarity around the terms 'summative' and 'formative' due to confusion, with preference given to 'low, medium and high-stakes' in order to understand the importance of the assessment with regards to expected outcomes.

Programmatic assessment

Programmatic assessment allows for snapshots of information to be obtained and collated, formulating a global picture of a registrar's competency. There was overwhelming support for the use of programmatic assessment in general practice training across the stakeholders. Programmatic assessment is based on WBAs collected over a registrar's training period, with MEs ultimately responsible at all RTOs for the collation and sign-off of these assessments. The current assessment model at the majority of RTOs already utilises this model to some degree with regards to flagging and meeting RACGP standards, however, it is apparent that there is significant scope to strengthen the process and create a robust program that formulates a more accurate picture of a registrar's true competencies.

Registrar: "We're building our skill all the way along. Do we really need to have such a high pressure situation where we could have lower value assessments earlier that reflect our progress and skill as we learn?"

ME: "I'm looking for good quality day to day assessment information, so that we can put together an overall picture of that registrar, through programmatic assessment and get an idea of how that registrar is progressing through their training and what kind of additional support they might need. Where their placements might need to be in the future, where are their strengths, where are their weaknesses, so we can fill in the gaps."

The collection of information from multiple data points was viewed favourably, however there were concerns raised around the funding of supervisors, particularly if their requirement for being an assessor increased.

> *ME: "I think it would have to be theoretically quite a lot of assessment data points. And obviously you'd be looking at, over the course of their training, at presumably a portfolio of how they've gone through assessment."*

> *Supervisor: "If there is talk about getting supervisors to be doing more assessments and more complex assessments we need more time to do that and that really means the money to do that."*

As part of programmatic assessment multiple view-points from multiple assessors can be collated, increasing inter-rater reliability. One-off assessments in the workplace, which are dependent on the context and assessor, tend to provide a snapshot of competency, but do not provide information about how the registrar performs overall. With multiple opinions the validity increases, thus it is important that DOVs are undertaken by different assessors, and that registrars move to different practices, with different supervisors.

> *Registrar: "You are assessing people from different aspects and you're really trying to find that you've got a very well rounded practitioner who's able to represent your college in your specialty in a robust and meaningful way."*

The concept was raised that safety is not only related to knowledge, but in relation to how a GP performs in the workplace.

> *ME: "The AHPRA statistics show that the problems that we face are not knowledge problems. They make up three per cent of complaints. Most of the problems are related to clinical reasoning issues. They're dependent on people having accurate and good clinical records. I think we have no way of assessing the clinical records that people keep at the moment, not properly and not by an exam system."*

Integrating core skills, such as procedures, into the programmatic model was discussed.

> *Supervisor: "It needs to be programmatic. It needs to be progressive along the way. I actually think that there should be some hurdles along the way that you actually have to successfully complete. Don't ask me what those hurdles are. But I think there needs to be hurdles that we think are core skills, you need evidence for the core skills, that you would then say, sorry, just spend another month or so doing that before we have another go at that."*

Programmatic assessment affords the ability to assess and quantify areas that are not covered in exams, which may factor into whether a registrar is safe to practice and at what level of supervision.

> *Supervisor: "We all know whether a doctor is any good or not, in the broader sense or are safe to practice. It's a case of trying to give enough evidence to support that intuitive view. Or alternatively, if we know they're not, which is probably the more difficult decision to make, having enough evidence to support that intuition that this person just isn't up to speed."*

> *ME: "Programmatic assessment gives you all those little bits of information."*

For programmatic assessment to form the basis of the assessment model in general practice training, changes to the current WBA model would need to occur, in order to increase the collection and collation of various view-points.

> *ME: "I don't think our programmatic assessments, although they are improving enormously, I'm just not sure that they're robust enough."*

Registrars were incredibly supportive of a programmatic model and emphasised their desire for robust feedback as they progressed throughout their training in the lead up to exams.

> *Registrar: "I feel like it actually would be really good for our mental health as well leading up to the exam, so that we consistently know where we're at throughout the year, gauge where we're at rather than it comes to the end and there's this massive exam."*

> *Registrar: "You have to have a strong robust way of assessing their abilities from multiple different angles."*

Programmatic assessment helps to inform an assessor of areas to focus on when it comes to assessing a registrar, with the ability to reflect on their portfolio of previous assessments.

> *ME: "Well, I think the programmatic assessment gives you all those little bits of information. That means when I go and do a direct observation visit, I'm not just going in, in isolation, going, okay, well, I'm just sitting in on some random registrar, who I've never met, who I know nothing about, who I don't know if they've got particular deficits in knowledge base or language or communication or whatever it is, and I - my job is to come in and sit and observe them at this point in time and that is it - which doesn't give me that big picture. Whereas the programmatic assessment is that pooling of all that information, so when I come in and observe them, I'm doing it with all this other knowledge."*

Programmatic assessment and collation of WBAs is usually undertaken by MEs, allowing tailoring of learning and education towards deficiencies that may emerge after a WBA.

> *ME: "All clinical teaching visits for my registrars, the reports on them are all sent to me and I have to sign them off. So, you can see how others do their clinical teaching visits. A lot of them are excellent and I often pick up pointers from them, actually. You can see if there are registrars who are struggling a bit and might need a bit more input."*

The role of programmatic assessment versus exams

The need to supplement examinations with workplace-based observation was raised as an important factor, with overwhelming support towards the usage of a robust programmatic assessment model in conjunction with written knowledge-based exams.

All stakeholders saw significant value in written, knowledge-based exams, with registrars acknowledging the importance of a formal written examination to ensure a high quality practitioner. However, there was less support for the OSCE, with participants acknowledging that the OSCE had the potential to be replaced with a less artificial form of assessment, in the form of a strong programmatic WBA model.

> *Supervisor: "I actually think it's a really good idea because the exams are so artificial compared with what really happens in general practice. Plus, it would allow to have a*

number of different opinions. You could have a whole range of different consultations and then a range of different doctors looking at those different consultations. It also allows you to assess the doctor patient interaction which would be much more realistic than what you get in an exam."

ME: "It is about collection [of] information. It's about giving honest, constructive feedback to the registrar. But also, and then to help them kind of move on to the next stage of their learning. But also to have that information to fed back into their overall portfolio with the training organisation."

Although the OSCE is a standardised assessment that offers a variety of simulated patients, there are varying attitudes to its utility in conjunction with a programmatic assessment model.

Registrar: "The OSCE is perfect medicine. Like they're actors who have been trained to give you answers if you ask the right questions, which is great in terms of testing clinical knowledge. But in real life, people have different education backgrounds, different socioeconomic status, and so I think that's a real advantage as well to having work placed assessment is how do you actually treat the person in front of you? Not just textbook medicine."

ME: "Everyone is so assessment focussed they're forgetting about the actually a real patient needs some really good communication skills, which are very difficult to replicate in an eight minute OSCE station."

An emerging theme from participants was that a safe and competent practitioner isn't necessarily a registrar who has passed the three exams, rather it's one who performs well in the work-place, with continuous assessment deeming the doctor to be safe to practice independently.

ME: "Does the exam actually mean this person's fit for practice, I'm not entirely certain about that."

The changing nature of medicine towards a greater reliance on technology for prescribing, guidelines and patient education, has resulted in the need for registrars to be assessed in how they operated in a modern GP environment.

Registrar: "I don't know that completely doing away with exams will ever be possible but the world has changed in the last 10 or 12 years and the way we practice medicine is different and the breadth of our knowledge is different."

ME: "The GP of today is totally different to the GP of 30 or 40 years ago. The expectations now are totally different. So the GP 40 years ago was expected to have a good grounding, knowledge and understanding whereas the GP nowadays, it's a little bit different. He's got instant access to data. So there's more of an emphasis on clinical reasoning and care taking, communication skills and things of that nature. I just wonder whether the college assessment is actually - how can I say - probably needs to be a little more up to date on that"

Registrar: "I think that the exams are important but I would like to see more workplace based assessments. I think in this day and age, patients have access to written information and the internet. They can look things up. Realistically, whilst I'd love to have the skill and ability to practice regardless of location, access to resources, it's very common for me to look up things that I use all the time, double check them in eTG, and use that information or check, you

> *know, monitoring in the AMH. Yes, while I can remember all of those things, at this point, I don't really find it's that useful."*

Overall a strong desire exists for the Applied Knowledge Test (AKT) and Key Feauture Problem (KFP) written exams to remain.

> *Registrar: "I think it's a good replacement to OSCEs but personally I think AKT and KFP should remain the way it is."*

There was no consensus regarding whether programmatic assessment should reduce the overall weighting of the exam and how much percentage should be given to the WBA programmatic model. However, an overall sign-off of the programmatic model, with no changes to the written exams, appears to be a supported concept.

> *Registrar: "I still think that the written component is important. Maybe it shouldn't just be the be all and end all 100 per cent marker of whether you're a safe and competent doctor."*

> *Supervisor: "I think there's a benefit from combining the data that you could get by more formal assessment through an ECTV for those intangible more nebulous things that might be difficult to pin down with an OSCE…. so just one of the things that [our RTO] has done in the last one to two years is they've replaced the final ECTV with a slightly different one where the medical educator sits in and just watches the registrar do, say, maybe half a dozen cases without the usual conversation about each case in between and then does some random case analysis with the registrar. So there's a certain sort of particular ECTV, which I think would be more useful in assessing what the registrar is capable of."*

Professionalism

Professionalism is an important skill for a safe and competent practitioner, yet professionalism can be 'performed' in an exam environment, and thus is difficult to assess. As part of professionalism there is a need to evaluate a registrar's work ethic, which is assessable through DOV and MSF.

> *Supervisor: "I don't want someone who's got 100 per cent and can't talk, can't interact, can't - some of the most frustrating things, one - the last unsatisfactory registrar I sat in with, she could not round off a consultation to save herself. The patient almost sort of said, can I go now."*

Difficult registrar

When it comes to assessing a registrar, it is important that the approach is multi-factorial, taking into account the supervisor or practice perspective. External clinical teaching visitors or external MEs observing the registrar interacting in the workplace can further help understand registrars who may be failing to perform. The desire of the registrar to truly want to undertake training appears also to be a factor, establishing whether there may be resistance to learning. One aspect of registrar training is to ensure that registrars are supported from a pastoral care perspective.

> *Supervisor: "I think you could almost say that's what makes a difficult Registrar, is that resistance, that not wanting to learn, not thinking you've got anything to teach them, not wanting to be there and making it obvious, and yes it is hard and then you've got to watch every step because you don't know what they're going to do either."*

> *Supervisor: "If you got an absolutely hopeless Registrar who shouldn't be there, what can you do? And some of them aren't hopeless forever but they're going through a bad time right now, and need some kind of pastoral support or something like that, but the options there are pretty limited."*

It is apparent that some supervisors need additional support in order to further support the needs of their registrar, including the need for additional financial support when registrars require additional support in the workplace.

> *Supervisor: "Where you're handed somebody that you think's either totally out place or who's experiencing difficulties where they're seriously underperforming at the moment, that sort of thing. That's the hard stuff that I've found."*

> *ME/Supervisor: "So when you've got a bit of a dud registrar who's dragging the chain and not necessarily fulfilling obligations. One, they're time consuming, they're demanding, they're difficult, they're taking up a lot more resources than a functioning registrar. Then looking at the fact that you may be doing all of this work and time and having no - potentially no funding because they haven't submitted the required stuff, I found that to be quite challenging."*

Flagging and remediation

Flagging and remediation as a result of a WBA was a theme that emerged from the focus groups and interviews. Registrars appeared unfamiliar with the term flagging and did not have a grasp of the concept of remediation. Supervisors had greater insight into the flagging process, and saw a red flag as an opportunity for the RTO to step in and provide additional support, often with the aim of passing exams.

> *Supervisor: "You get flagged by the training body. You don't - well, okay. Sorry. I say fail - it's not in the sense of fail you kicked out of the program. It's fail like you get extra support. So you may get more site visits, so they might schedule more ECTVs. They may get the education coordinator to do more one on one training or see what their focus is. So it's more intensive."*

Medical Educators were familiar with the diagnostic process of flagging and reported on the grading of flags. The process appeared complex and as such was explored in further detail in Stream 3d.

Literature Review

A formal literature review was undertaken as part of Stream 1 of this project, however certain concepts emerged from participants that warranted a further exploration of the literature. As such a grounded theory approach was undertaken further exploring the above themes, allowing for reflection on the empirical research in order to guide the literature review.

Entrustable professional activities (EPAs) have a very strong evidence base in medical education, with a stepwise progression that potentially bridges a gap between the theories of competence and clinical practice (ten Cate, 2007). A STAR is a 'statement of awarded responsibility', which can be assigned once a certain level of competency is reached, such as being able to practise unsupervised, taking into account different assessor viewpoints (ten Cate, 2013).

There was limited evidence in the literature regarding review of videoed consultations providing a standalone view-point in programmatic assessment, with implementation difficulties noted and a general dislike regarding their usage (Nilsen, 2005). However, when used in conjunction with

feedback from experts with a specific purpose video-reviews were seen in a more positive light (Hammoud, 2012).

Random case analysis (RCA) was not part of the original WBA audit, however emerged as an important component of WBA. There have been significant developments in recent years with training of supervisors and medical educators in the delivery of RCA. Evidence regarding RCA as a robust tool to assess clinical reasoning has been published (Morgan, 2013). In addition, adding random case analysis to direct observation visits by external clinical teachers (ARCADO) has gained momentum in Australian general practice training as an additional tool, with an increasing body of research supporting its utility (Ingham, 2016). Using random case analysis in conjunction with audit and feedback methods is seen as a useful addition to the WBA toolkit, as exampled in a published pilot study looking at TRAFk, 'test result audit and feedback', to monitor and reflect on registrar test ordering (Morgan, 2016).

Procedural skills logs are not currently utilised in RACGP general practice training, however there is solid evidence that a signed-off logbook allows for staged learning and observation to ensure a satisfactory level (Wearne, 2011). There has also been additional research in Australia regarding which procedures should form part of a procedural skills curriculum (Sylvester, 2011). It is apparent that skin cancer medicine skills are increasingly important in general practice, with basic surgical techniques required in general practice and as such warranting assessment (Tng, 2018). Logbooks also promote registrar empowerment, encouraging an increased level of supervision, by mandating the sign-off of skills after observation (Gouda, 2016).

Learning plans, when performed correctly whereby there is planned learning, as a prospective tool, with actionable outcomes, has a strong evidence-base in the literature. Mandated learning plans are considered bureaucratic and of minimal usage to the user and the assessor (Garth, 2016). Learning plans should take into account training needs and goals, ensuring that the high-achieving registrar is identified early, allowing for the development of their strengths and areas of special interest (Zaharias, 2018).

Patient encounter tracking and learning tools (PETAL), have been discussed in the literature for over two decades, with an emphasis on the importance of monitoring patient encounters and reflecting on the data (Mulloy, 1995). The Registrar Clinical Encounters in Training (ReCEnT) study, which is undertaken across a number of RTOs, is a paper-based PETAL that longitudinally maps the nature and associations of consultation-based clinical and educational experiences of general practice registrars (Morgan, 2012). The key component of PETAL, including ReCEnt, in relation to WBAs appears to be the prompting for self-reflection on individual feedback reports, reflecting on patient encounters, practice and training progression (Magin, 2015).

Early assessment, such as a safety assessment, allows the supervisor to determine the level of supervision and support that the registrar may require, with subsequent assessments charting progression towards competence (Beard, 2011). Although the term 'safety assessment' is a new concept to the literature, there is evidence that an early assessment can assist with flagging GP registrars to ensure extra assistance is provided (Gladman, 2011). Orientation with a 'survival skills' approach has been shown to help ease the transition into the general practice consulting environment (Morgan, 2014).

Discussion

Rich qualitative data was gathered from the interviews and focus groups, with registrars, supervisors and medical educators providing their insight and opinions on the role of WBAs in general practice training. The results were grouped into key themes that emerged in the coding process and explored from the viewpoints of all stakeholder groups interviewed. The above themes have provided significant guidance to the proposed implementation example and overall framework.

Revisiting assessments such as a procedure log was raised as a positive change that registrars would like to see incorporated into the WBA portfolio. The concept of a procedural skills log, whereby skills would be observed and signed off by an assessor (supervisor, ECT, ME), allowing for standardisation of skills-based competencies in general practice was reported across the registrar stakeholder group. In addition, there was a desire from registrars to encourage more active supervision in the work-place with procedures being an avenue for doing this (Gouda, 2016).

There were significant negative reflections around the use of learning plans across all stakeholder groups. Registrars appear to be using self-initiated learning logs whereby they record their learning needs electronically as a checklist, often updated multiple times a day whilst consulting. A set learning plan per training terms appears to have minimal utility in a modern training program. An electronic, dynamic learning log, which can be assessed at a direct observation visit with a reflection by the registrar appears more in line with registrar learning requirements. There is also the additional need for the registrar to reflect after each WBA and ensure that the feedback can be translated into meaningful outcomes and goals for the registrar. This was strongly supported by registrars who utilise PETAL as a patient encounter and tracking tool for learning and reflection.

Registrars value narrative-style, in-depth, timely feedback and prefer balanced feedback incorporating negative feedback and ways to improve. This ensures that feedback can result in action and progression of learning level. It was as part of the feedback discussions that registrars found benchmarking unhelpful and preferred a user-rated scale based on level of supervision, as offered by entrustable professional activities. Direct observation was a highlight for the majority of registrars, and despite anxiety that may be associated with having an assessor in the room, was highly regarded. The feedback and teaching associated with these visits, structured into the observation visits, emphasised why in-person observation is so helpful to a registrar's progress.

Videoed consultations, while beneficial for identifying professionalism concerns, appeared to be useful for self-reflection and in cases of remediation only. Otherwise they were seen as burdensome and time-consuming for the registrar, practice and supervisor. Supervisor direct observation was performed at a minimum, with significant concerns around remuneration and time lost raised as reasons why supervisors are reluctant to spend time observing consults. MEs and registrars reported their eagerness to increase supervisor direct observation, especially to assist with ensuring safety to practice.

Flagging and the need for remediation emerged as key concepts in the role of the ME, as well as their role in pastoral care. The supervisor was seen as the day-to-day person who could help guide and nurture the registrar's progress, whilst maintaining a safe environment for learning and developing competencies. The external clinical teaching role is imperative when it comes to independent observation, with registrars valuing honest and non-biased feedback. The relationship of the supervisor and registrar and the tension that can exist emerged, which

emphasised the need for multiple viewpoints when it came to assessing a registrar, including multi-source feedback from staff members.

Overall there was a desire to strengthen the workplace-based assessment curriculum into a programmatic model and look at whether all three exams components were necessary as part of the Fellowship requirements, once a more robust programmatic model came into effect. Surprisingly, there was a desire from all stakeholders for the two written exams (KFP, AKT) to continue, with registrars emphasising the importance of a high-standard knowledge-based written exam to both drive learning, and to ensure that Fellowship is maintained at a high standard. The OSCE had significantly less support with stakeholders, and a majority concluded that the OSCE could be superseded by a more structured workplace-based assessment model, with multiple data-points throughout training, allowing for real-world, in-context assessment that truly reflects whether a registrar is safe to practice independently.

Conclusion

It is evident that there are considerable differences in the implementation and delivery of workplace-based assessments across Australia, however there were a number of common elements. There is a strong desire for more active supervision, providing numerous snapshots of registrar progression. Increased supervision, facilitated through DOVs, procedural logbooks, EPAs, learning logs, RCA and a safety assessment, all appear to have a robust role in the workplace. In addition, fundamental features included the value that participants placed on the importance of a written based exam, and that this was the viewpoint of both assessors and registrars. Based on the stakeholder interviews and focus groups, the utility of the OSCE appears low, with work-place based programmatic assessment model providing a more robust and collaborative assessment of real-world primary care.

References

Beard, J. (2011). Workplace-based assessment: the need for continued evaluation and refinement. *the surgeon*, *9*, S12-S13.

Garth, B., Kirby, C., Silberberg, P., & Brown, J. (2016). Utility of learning plans in general practice vocational training: a mixed-methods national study of registrar, supervisor, and educator perspectives. *BMC medical education*, *16*(1), 211.

Gladman, G. (2011). The identification of the general practice registrar needing assistance. *The Australasian medical journal*, *4*(6), 308.

Gouda, P. (2016). The need for logbooks to evolve in the undergraduate medical setting. *Perspectives on medical education*, *5*(1), 65-65.

Hammoud, M. M., Morgan, H. K., Edwards, M. E., Lyon, J. A., & White, C. (2012). Is video review of patient encounters an effective tool for medical student learning? A review of the literature. *Advances in medical education and practice*, *3*, 19.

Ingham, G., Fry, J., & Ward, B. (2016). Adding random case analysis to direct observation (ARCADO)-Updating the external clinical teaching visit to improve general practice registrar assessments. *Australian family physician*, *45*(12), 918.

Krippendorff, K. (2004). Reliability in content analysis. *Human communication research*, *30*(3), 411-433.

Magin, P., Morgan, S., Henderson, K., Tapley, A., Scott, J., Spike, N., ... & van Driel, M. (2015). The Registrars' Clinical Encounters in Training (ReCEnT) project: Educational and research aspects of documenting general practice trainees' clinical experience. *Australian family physician*, *44*(9), 681.

Morgan, S., Chan, M., & Starling, C. (2014). Starting off in general practice-consultation skill tips for new GP registrars. *Australian family physician*, *43*(9), 645.

Morgan, S., Ingham, G., Kinsman, L., & Fry, J. (2015). Clinical supervision using random case analysis in general practice training. *Education for Primary Care*, *26*(1), 40-46.

Morgan, S., & Ingham, G. (2013). Random case analysis: A new framework for Australian general practice training. *Australian family physician*, *42*(1/2), 69.

Morgan, S., Magin, P. J., Henderson, K. M., Goode, S. M., Scott, J., Bowe, S. J., ... & Van Driel, M. L. (2012). Study protocol: the registrar clinical encounters in training (ReCEnT) study. *BMC family practice*, *13*(1), 50.

Morgan, S., Saltis, T., Coleman, J., Tapley, A., & Magin, P. (2016). Test result audit and feedback (TRAFk) as a supervision method for rational test ordering in general practice training. *Australian family physician*, *45*(7), 518.

Mulloy, J. V., Leuschen, M., & Rowe, B. H. (1995). Computer-based patient encounter tracking. Development of a system for family medicine residents. *Canadian Family Physician*, *41*, 1742.

Nilsen, S., & Baerheim, A. (2005). Feedback on video recorded consultations in medical teaching: why students loathe and love it–a focus-group based qualitative study. *BMC medical education*, *5*(1), 28.

Sylvester, S., Magin, P., Sweeney, K., Morgan, S., & Henderson, K. (2011). Procedural skills in general practice vocational training: what should be taught?. *Australian family physician*, *40*(1/2), 5.

Ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *Journal of graduate medical education*, *5*(1), 157-158.

Ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: can we bridge the gap between theory and clinical practice?. *Academic Medicine*, *82*(6), 542-547.

Wearne, S. (2011). Teaching procedural skills in general practice. *Australian family physician*, *40*(1/2), 63.

Zaharias, G. (2018). High-performing general practice registrars, the gifted and the talented: Helping them to reach their potential. *MedEdPublish*, *7.*

Stream 3c
Evaluation of Entrustable Professional Activities

Assoc Prof Jill Benson AM (ModMed Institute)
Dr Nyoli Valentine (ModMed Institute)
Dr Jon Wignes (ModMed Institute)

MODMED INSTITUTE

In collaboration with

GPEx

general practice training tasmania

WAGPET

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

This research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

The researchers would also like to acknowledge:

- ModMed Institute for authoring this report and contributing their existing knowledge and expertise.
- Project management support provided by Ms Michelle Pitot (GPEx).
- The Working Group for this stream including: Assoc. Prof Jill Benson AM (ModMed Institute/ GPEx), Dr Nyoli Valentine (ModMed Institute), Dr Jon Wignes (ModMed Institute), Mr Michael Bentley (General Practice Training Tasmania), Dr Colleen Bradford (Western Australian General Practice Education and Training) and Ms Michelle Pitot (GPEx).
- GPEx, General Practice Training Tasmania and Western Australian General Practice Education and Training for their support and contributions to the research and this report.
- Special thanks to Dr Kaye Atkinson (ModMed Institute) for her contribution to the narrative data analysis in this report.
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.

In collaboration with

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Introduction

In the mid-2000s Olle ten Cate developed Entrustable Professional Activities (EPAs), an innovative and practical method to assess competency in medical education (ten Cate, 2007). Since then, EPAs have gradually been accepted by medical training organisations throughout the world as a way of bridging the gap between clinical practice, including the concept of the expert judgement of supervisors, and educational theories such as that of outcome-based training and competencies (ten Cate, 2007). 'Entrustment' is a global assessment about the level of supervision needed to ensure patient safety. 'Trust' implies, not only current work, but also the registrar's ability to handle future patients and situations that may pose new challenges (Peters, 2017). Thus, entrustment also includes features of self-reflection and life-long learning (Rekman, 2016).

Curricula and competency frameworks identify the areas in which a trainee (registrar) is expected to be safe to practise unsupervised at the end of training (Shaughnessy, 2013). Entrustment decisions are based on observation of multiple activities within the workplace that involve a complex integration of knowledge, skills, attitudes and values that can be mapped to those competencies (Schultz, 2015). EPAs enable assessors such as supervisors to combine their own intuition with observed assessments such as multi-source feedback, case review, direct observation and skills assessment (ten Cate, 2015). The complexity of general practice is far greater than the measurement of competencies, and EPAs better reflect this complexity as they require proficiency in multiple competencies simultaneously (Shaughnessy, 2013; ten Cate, 2015).

The progression of building trust will be graded from 'Needs frequent direct in-room review by supervisor', the expectation at entry into the programme, to 'Safe to practise unsupervised' in all areas, at finalisation of training. Registrars will progress with increasing expertise at different speeds through the various EPAs depending on their exposure to the issues, the complexity of the EPA, their self-motivation for learning, their self-confidence, the teaching they receive and their personality (ten Cate, 2005). In some EPA literature, a 'statement of awarded responsibility (STAR)' is awarded as an acknowledgement of the ability and right to practise without supervision, either in a particular EPA or overall (ten Cate, 2015). At that stage the usual EPAs could be superseded by a fifth level of achievement of 'Safe to provide supervision to junior learners' (ten Cate, 2015). At this level autonomous practise continues to be in collaboration and interdependence with other health professionals, not as supervisors, but as colleagues (ten Cate, 2016).

Training in the use of the assessment tool, engagement of the supervisor in the process of assessment and teaching, a self-reflective registrar, and an appropriate context, are all essential features if EPAs are to be successful for the purposes of ensuring public safety and registrar learning (ten Cate, 2013; ten Cate, 2016). As well as clinical competence, EPAs can assess reliability (stable, predictable behaviour and a conscientious way of working) and humility (recognition of own limitations and willingness to ask for help if needed) (Peters, 2017). They can assess the ability to work effectively in teams and to navigate complex health care systems (Valentine, 2019). EPAs can also assist trainees in requesting targeted feedback from their supervisors (Bonnie, 2019).

Sharing EPA assessments across the span of a registrar's training enables longitudinal evaluation of the trust conditions of 'competence, reliability, and honesty' across different contexts (ten Cate, 2016). Narrative descriptions of entrustment are able to be more clearly interpreted than a comparison with peers or 'at expected level' without a benchmark of what that might be, and this narrative is a rich source of information about the progress of the trainee (Rekman, 2015). Focusing on an end-goal of 'safe to practise unsupervised' rather than a grade will assist in finding those who are not on track and has the potential to 'increase the amount of well-constructed

actionable feedback' as early as possible (Rekman, 2015). On the other hand, an entrustment decision made carelessly as a 'tick-box' exercise, or a situation where adequate supervision is not possible, can jeopardise patient safety (ten Cate, 2016).

The EPAs used in this study were developed in South Australia (SA) using evidence-based principles (ten Cate, 2013; O'Dowd, 2019). In particular they complied with the recommendations by O'Dowd et al (2019) using 'Broad EPAs that in turn link to multiple competencies and reduce the complexity of the curriculum enable a holistic view of the learner and therefore lend themselves better to implementation'. A study outlining the development and use of these EPAs as an assessment tool in general practice in South Australia has recently been published (Valentine, 2019). It concluded that 'in addition to having a strong theoretical basis, EPAs are a valid assessment tool in a real-world workplace-based training environment' (Valentine, 2019). This current research builds on the previous study by Valentine, expanding the quantitative and narrative component to two other RTOs in Western Australia (WA) and Tasmania, and exploring the opinions of registrars, supervisors and Medical Educators (MEs) about EPAs.

Aim and research questions

The aim of the present study was to determine if EPAs are useful as an assessment tool for general practice registrars, to map that utility in the three regional training organisations (RTOs) currently using them, and to garner the opinions of supervisors, registrars and MEs who are currently using EPAs as part of their workplace-based assessments.

This study was based on the following questions:

- Are EPAs a robust assessment tool in a general practice workplace-based training environment?
- Are EPAs effective tools for stimulating registrar self-reflection?
- Do EPAs facilitate teaching and learning?
- Are EPAs able to demonstrate competency progression through different stages of the Australian General Practice Training (AGPT) program?

Methods

Quantitative data

Study Cohort

GPEx has been using EPA assessments for the mid and end-term assessments of all registrars since January 2017. In second semester 2018, both Western Australian General Practice Education and Training (WAGPET) and General Practice Training Tasmania (GPTT) piloted the same EPA assessments in small cohorts and in 2019 both RTOs will be using EPAs for all of their registrars. Twice each semester (i.e. four times a year) supervisors completed EPA assessments for registrars, and registrars undertook a self-assessment, considering their entrustment levels of themselves.

There are 13 EPAs mapped to the RACGP curricula, domains and competencies of the Royal Australian College of General Practitioners (RACGP) and the Australian College of Rural and Remote Medicine (ACRRM). The content headings are:

1. Take a comprehensive history and perform an examination in all patients
2. Identify common working diagnoses and prioritise a list of differential diagnoses
3. Manage the care of patients with acute common symptoms and diseases across multiple care settings

4. Manage the care of patients with common chronic disease and multiple morbidities
5. Manage gender related health issues
6. Manage mental health issues
7. Manage the care of children and adolescents
8. Follow screening protocols and primary prevention guidelines
9. Lead and work within professional teams
10. Communicate effectively and develop partnerships with patients, carers and families
11. Demonstrate time management and practice management skills
12. Demonstrate attributes that are expected of a general practitioner (GP)
13. RURAL ONLY: Demonstrates an ability to manage emergency on call and inpatient care

Each EPA is graded as:

1. Needs frequent direct in room review by supervisor
2. Needs onsite supervisor available in the practice most of the time
3. Safe to practise with phone access from supervisor
4. Safe to practise unsupervised

EPAs 1-8 and 13 are clinical and EPAs 9-12 non-clinical. EPA 13 is only relevant for rural registrars. Supervisor and registrar pairs were expected to discuss the EPA assessments and subsequently develop learning and teaching outcomes.

Data relating to all general practice registrars working in SA for the two semesters of 2017, and relating to the WAGPET and GPTT registrars in the pilot group for the second semester of 2018 were analysed. Registrars were grouped according to their length of time in community-based training, with those in their first 12 months considered junior registrars (GPT1 and GPT2), and those with more than 12 months experience considered senior registrars.

Ethics approval from Flinders University (#7882) for data collection, focus groups and interviews had previously been granted for SA and was modified to include WA and Tasmania.

EPA assessment process

The EPA assessments were developed and refined by the ModMed Institute, and the article by Valentine outlines the development and assessment of the validity of those EPAs (Valentine, 2019). Data for GPEx has previously been collated and results published (Valentine, 2019). This data is included as part of this project.

Prior to Valentine's study, no EPAs were available for general practice registrars within Australia, though Australian EPAs have been developed for psychiatry, physician and surgical training, and internationally for family medicine. The EPAs used by GPEx, WAGPET and GPTT have been mapped to the curriculum, domains and competencies of the RACGP and ACRRM. Valentine's article shows that EPAs are a valid assessment tool to map growth in entrustment levels over 12 months, a difference in entrustment levels between senior and junior registrars, and an increase in concordance between registrar self-assessment and supervisor assessment over two semesters.

Entrustment levels for all 13 EPAs were assessed by supervisors and by the registrars' self-assessment at four three-monthly time points in 2017 at GPEx and at two three-monthly time points in 2018 at WAGPET and GPTT. Assessments and comments were uploaded to an online learning management system as part of usual education for GPEx and for the pilot groups in WAGPET and GPTT the EPA assessments replaced the usual mid and end-term assessments. The entrustment level data and comments were extracted from the online system at GPEx and WAGPET for analysis and GPTT sent de-identified collated data to GPEx.

Statistical analysis of quantitative data

Entrustment levels were ordinal, but not interval, so non-parametric statistics were used in the analysis. For each EPA, to show the difference between junior and senior registrars, the levels of entrustment (1-4) as rated by supervisors, were recorded as frequency distributions and tested using a Mann-Whitney U test at the three month and 12-month distributions. EPAs were presented as clinical or non-clinical. A one-tailed p-value for an increase in entrustment for each EPA was produced using the statistical package GraphPad Prism. A Bonferroni correction was applied to the expected levels of significance to compensate for the number of individual comparisons; consequently, the required p- value for significance was set to .0038. The ratio of the registrars who were safe to practise unsupervised was compared with those who were not safe to practise unsupervised between junior and senior registrars using a chi-squared test with 95% confidence intervals determined by the Koopman asymptotic score.

Growth in entrustment scores was determined over a nine-month period for GPEx registrars. At each time-point, supervisor entrustment levels for clinical and non-clinical EPAs were pooled for each training level to determine the ratio at the time-point that were safe to practise unsupervised. A Fisher's exact test was performed to compare the ratios to the three-month levels, with 95% confidence intervals determined by the Koopman asymptotic score.

To calculate concordance between supervisor and registrar levels of entrustment; for each EPA the difference was summed to obtain a net concordance difference (score of 0 meant that the registrar and supervisor agreed, negative ratings indicated the registrar scored higher levels of entrustment, and positive indicated the supervisor scored higher levels of entrustment). These were visualised as frequency distributions. The sums of the differences in scores were compared between senior and junior registrars with an independent samples t-test.

Comments associated with supervisor and registrar EPA assessments

Each EPA included a comment section for both supervisor and registrar. The purpose of this was to expand and explain any feedback from the supervisor, and for the registrar to present a justification and/or self-reflection on the EPA score they had given themselves.

From the GPEx supervisor comments in 2017, 90 were coded – 51 giving feedback to junior registrars and 39 to senior registrars. There were 43 WAGPET supervisor comments coded for junior and senior registrar feedback and 10 GPTT supervisor comments, which were not separated into junior and senior registrar feedback comments. The same number of registrar comments were also coded but not matched to those of their supervisor pair.

The GPEx comments were coded by one coder, and the WAGPET and GPTT comments coded by a second coder using the first coder's themes. Each statement in the comments that provided a single feedback item was coded as a unique utterance. For example, a statement 'consistently displays thorough approach to History and P on review of notes and requires only infrequent phone access for problem cases' was coded as two utterances.

Utterances were coded into nine discrete themes: Specific Competency (Competency); Specific Advice (Teaching); Experience (Exposure); Specific Gaps/Weaknesses (Gaps); Confidence; Requirement for Supervision (Supervision); Behaviour; Performance, and; Progress.

The comments were rated as high, moderate or low-quality feedback or reflection. The criteria rating for the comments used the criteria for effective feedback, based on Jackson's research: Quantifiable, Specific, Actionable, Balanced, Objective, Based on goals, and Behavioural (Jackson, 2015). EPA comments that include none of these characteristics were rated as low-quality,

feedback comments meeting at least one of these criteria were rated as moderate and meeting more than one of these criteria was high-quality feedback.

Focus groups and interviews

The qualitative arm of the present study involved audio-recordings of focus groups and individual interviews in all three RTOs. These were face-to-face in SA and Tasmania and by video-conference in WA. Participants in WA and Tasmania were invited through the staff member of the RTO who sat on the Steering Group for the overall research grant. In SA, supervisors were invited to participate as part of a GPEx Supervisor Workshop rotation, MEs were invited as part of an ME meeting and then volunteered individually for more detailed interviews, and registrars were invited by email through the Registrar Liaison Officer at GPEx. Overall there were 19 registrars, 17 supervisors and 8 MEs who took part in the research.

Signed informed consent was gained from each individual and recordings were transcribed confidentially off-site.

Interviews and focus groups were approximately half an hour in length. They were semi-structured and informally followed the questions below:

Focus Group and interview questions

- How are you finding Entrustable Professional Activities?
- What are the benefits of EPAs?
- What are the limitations of EPAs?
- How do you determine levels of entrustment? (Supervisors only)
- Do they stimulate feedback discussion with the registrar? (Supervisors only)
- Do they help facilitate your teaching? (Supervisors and MEs only)
- Do they help you identify a struggling registrar? (Supervisors and MEs only)
- Does the registrar self-evaluation help you with your feedback discussions? Do you review this prior to completing your assessment? (Supervisors only)
- Are they straightforward to review? (MEs only)
- Do they help facilitate your teaching? (MEs only)
- Do they help you coach the registrar? (MEs only)
- Are they useful in developing learning plans? (Registrars only)
- Are they useful in helping you identify your learning needs? (Registrars only)
- How could they be improved?

Transcripts were coded in NVivo Version 12 for Windows (QSR International, 2016). This coding was informed by the research questions about benefits, challenges and limitations, registrar self-evaluation and how to determine entrustment. Further themes of training, ceiling, feedback, narrative, and the tension of being an assessor and educator emerged.

Results

Quantitative data

Across three RTOs, EPA data was available for analysis from 337 registrars (172 junior, 165 senior). The number of junior registrars entrusted to perform EPAs without supervision ranged from 21.3% to 43.3% for clinical EPAs, and 37.4 to 65.2% for non-clinical EPAs. The range for senior registrars who were entrusted to perform all EPAs without supervision, was from 67.6% to 93.3% for clinical EPAs and 80.2 to 94.1% for non-clinical EPAs. This shift was significant (one tailed Mann-Whitney,

p<0.001 for each EPA). The results are shown in figures 1 and 2 and is consistent with the results previously published from GPEx (Valentine, 2019).

Figure 1 below shows the difference in levels of entrustment between junior and senior registrars in clinical EPAs. For each EPA, the distribution of entrustment levels for junior and senior registrars was significantly different (Mann-Whitney, p<0.0001).

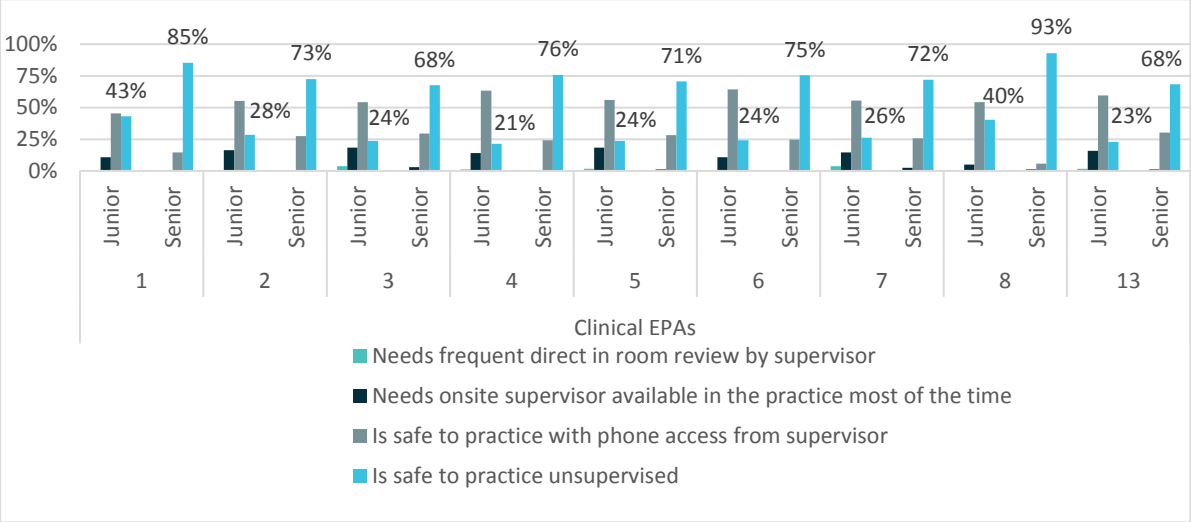


Figure 1. Differences in levels of entrustment between junior and senior registrars in clinical EPAs.

Table 1. Ratio of proportions of junior and senior registrars entrusted as being safe to practise unsupervised by their supervisor at three month assessment.

Clinical EPAs	*Junior registrars*	*Senior registrars*	*Rate Ratio (95% CI)*
EPA 1	71/165 (43%)	146/171 (85.4%)	1.984 (1.662-2.411)
EPA 2	47/165 (28.5%)	124/171 (72.5%)	2.546 (1.984-3.322)
EPA 3	39/164 (23.8%)	115/170 (67.6%)	2.845 (2.143-3.840)
EPA 4	35/164 (21.3%)	128/169 (75.7%)	3.549 (2.644-4.856)
EPA 5	39/164 (23.8%)	120/170 (70.6%)	2.968 (2.243-3.999)
EPA 6	40/165 (24.2%)	129/171 (75.4%)	3.112 (2.370-4.162)
EPA 7	43/164 (26.2%)	123/171 (71.9%)	2.743 (2.108-3.633)
EPA 8	66/164 (40.2%)	159/171 (93.0%)	2.310 (1.930-2.824)
EPA 13	13/57 (22.8%)	52/76 (68.4%)	3.000 (1.881-5.042)

Figure 2 below shows the difference in levels of entrustment between junior and senior registrars in non-clinical EPAs. For each EPA, the distribution of entrustment levels for junior and senior registrars was significantly different (Mann-Whitney, *p*<0.0001).

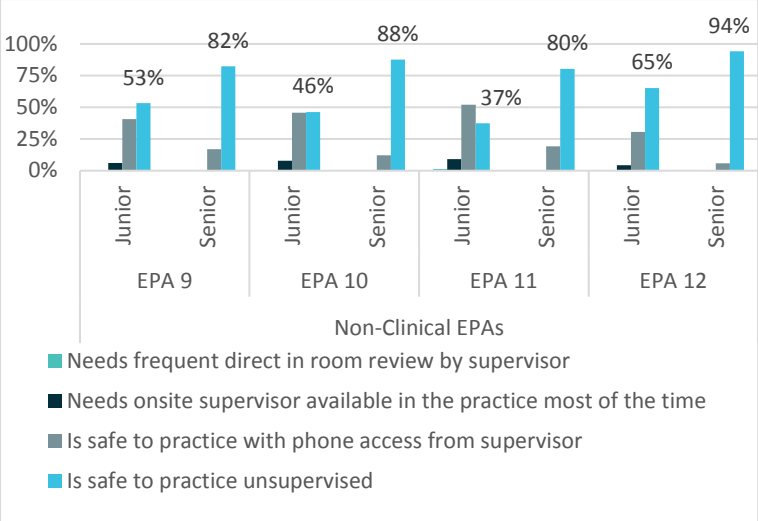


Figure 2. Difference in levels of entrustment between junior and senior registrars in non-clinical EPAs.

Table 2. Ratio of proportions of junior and senior registrars entrusted as being safe to practise unsupervised by their supervisor at three month assessment.

Non-Clinical EPAs	*Junior registrars*	*Senior registrars*	*Rate Ratio (95% CI)*
EPA 9	88/165 (53.3%)	141/171 (82.5%)	1.546 (1.328-1.826)
EPA 10	76/164 (46.3%)	150/171 (87.7%)	1.893 (1.605-2.273)
EPA 11	61/163 (37.4%)	138/172 (80.2%)	2.144 (1.750-2.672)
EPA 12	107/164 (65.2%)	160/170 (94.1%)	1.443 (1.293-1.639)

Valentine (2019) showed the proportion of EPAs on which junior registrars were judged to be able to practise unsupervised increased significantly over 12 months training in a general practice setting (Figure 3). The proportion entrusted to perform clinical EPAs unsupervised increased from 26% at 3 months, to 35% at 6 months (1.37 times increase, 95% CI: 1.15-1.63, p=0.0003), 50% at 9 months (1.92 times increase, 95% CI: 1.64-2.26, p<0.0001) to 69% at 12 months (2.684 times increase, 95% CI: 2.322-3.118, p<0.0001). Similarly, for non-clinical EPAs the proportion entrusted to practise unsupervised significantly increased from 50% at 3 months, to 56% at 6 months (1.13 times increase, 95% CI: 0.97-1.32, p=.076), 72% at 9 months (1.44 times increase, 95% CI: 1.26-1.66, p<0.0001) and 83% at 12 months (1.66 times increase, 95% CI: 1.47-1.90, p<0.0001).

Figure 3 shows the percentage of junior registrars at GPEx at each entrustment level for clinical and non-clinical EPAs over a 12-month period (from Valentine, 2019).

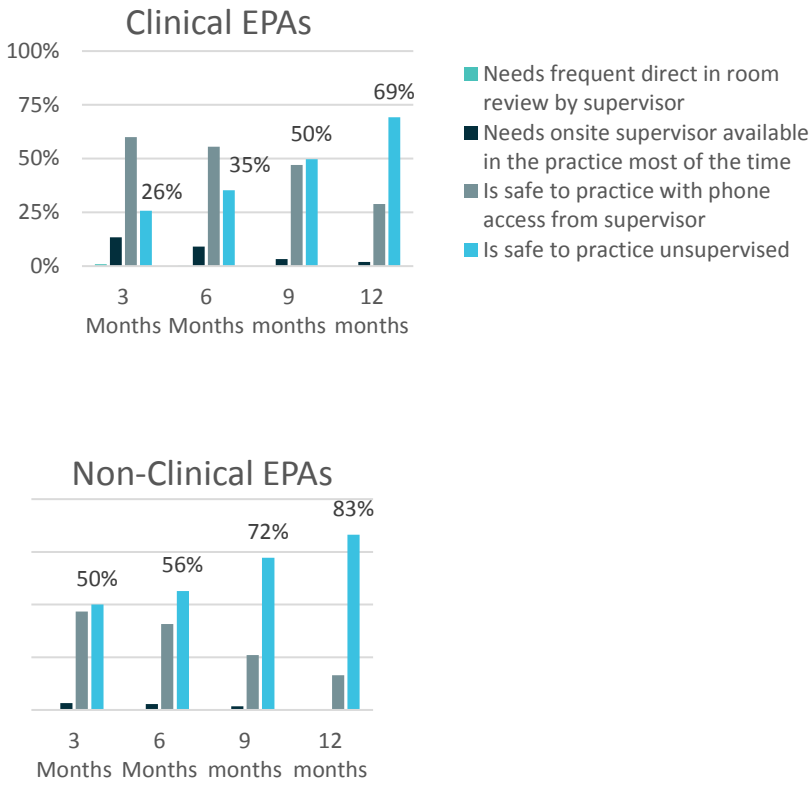


Figure 3. Percentage of junior registrars at GPEx at each entrustment level for clinical and non-clinical EPAs (Valentine, 2019).

Concordance between levels of entrustment for supervisor assessment and registrar self-assessment at GPEx three months is shown in figure 4. The mean differences in entrustment ratings between supervisor and registrar decreased significantly from 5.52 for junior registrar to 2.93 points for senior registrars (t=4.126, df=169.146, p<0.0001).

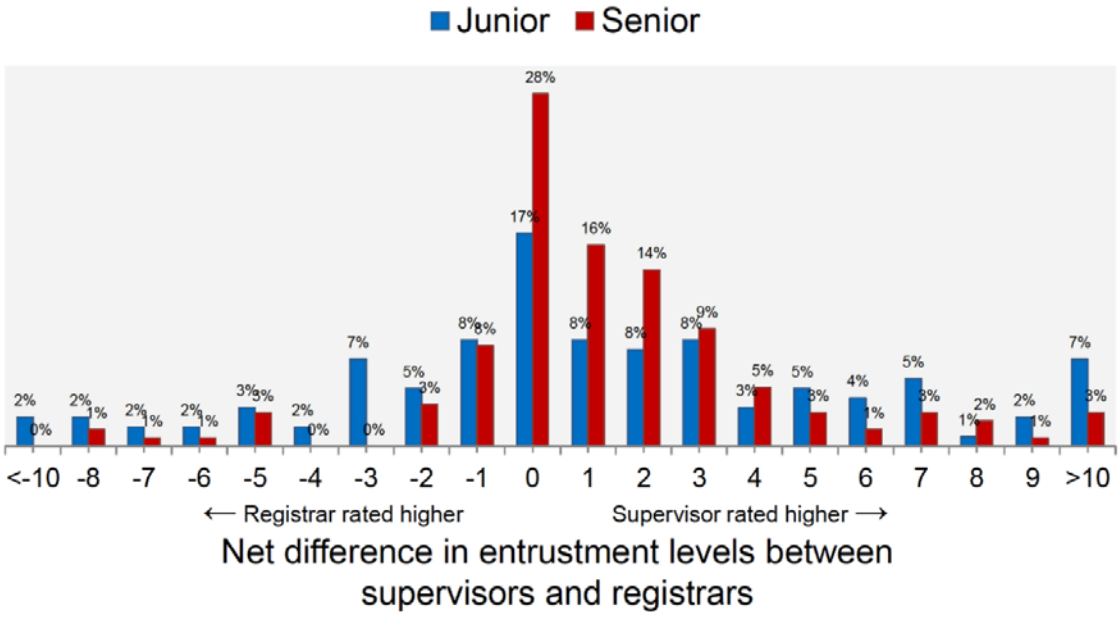


Figure 4. Concordance of levels of entrustment between junior and senior registrars at GPEx (Valentine, 2019).

Comments associated with supervisor and registrar EPA assessments

There were 90 general practice registrars at GPEx whose supervisor comments were qualitatively coded. There were 1307 unique utterances by general practice supervisors: 28% were about competency, 22.6% were about performance, 11% were about supervision and 11.7% were about the provision of specific advice designed to improve performance. Only 5% were about specific gaps or weakness.

Comment themes

Looking at the EPAs for GPEx registrars in Year 1 (junior) and Year 2 (senior), supervisors provided comments across the nine themes for Year 1 and Year 2 registrars, with more comments based around Advice and Experience for Year 1 registrars.

Table 3. Coding for utterances for feedback from GPEx supervisors

Theme	Junior (Year 1)		Senior (Year2)		Total	
	Count	%	Count	%	Count	%
Specific Competency (Competency)	143	22.14%	161	36.76%	***304***	***28.04%***
Performance (Performance)	121	18.73%	124	28.31%	***245***	***22.60%***
Specific Advice (Teaching/ Advice)	100	15.48%	27	6.16%	***127***	***11.72%***
Requirement for Supervision (Supervision)	84	13.00%	35	7.99%	***119***	***10.98%***
Experience (Exposure)	76	11.76%	34	7.76%	***110***	***10.15%***
Progress (Progress)	43	6.66%	21	4.79%	***64***	***5.90%***
Specific Gaps/Weaknesses (Gap)	30	4.64%	24	5.48%	***54***	***4.98%***
Behaviour (Behaviour)	37	5.73%	0	0.00%	***37***	***3.41%***
Confidence (Confidence)	12	1.86%	12	2.74%	***24***	***2.21%***
Total	***646***	***100.00%***	***438***	***100.00%***	***1084***	***100.00%***

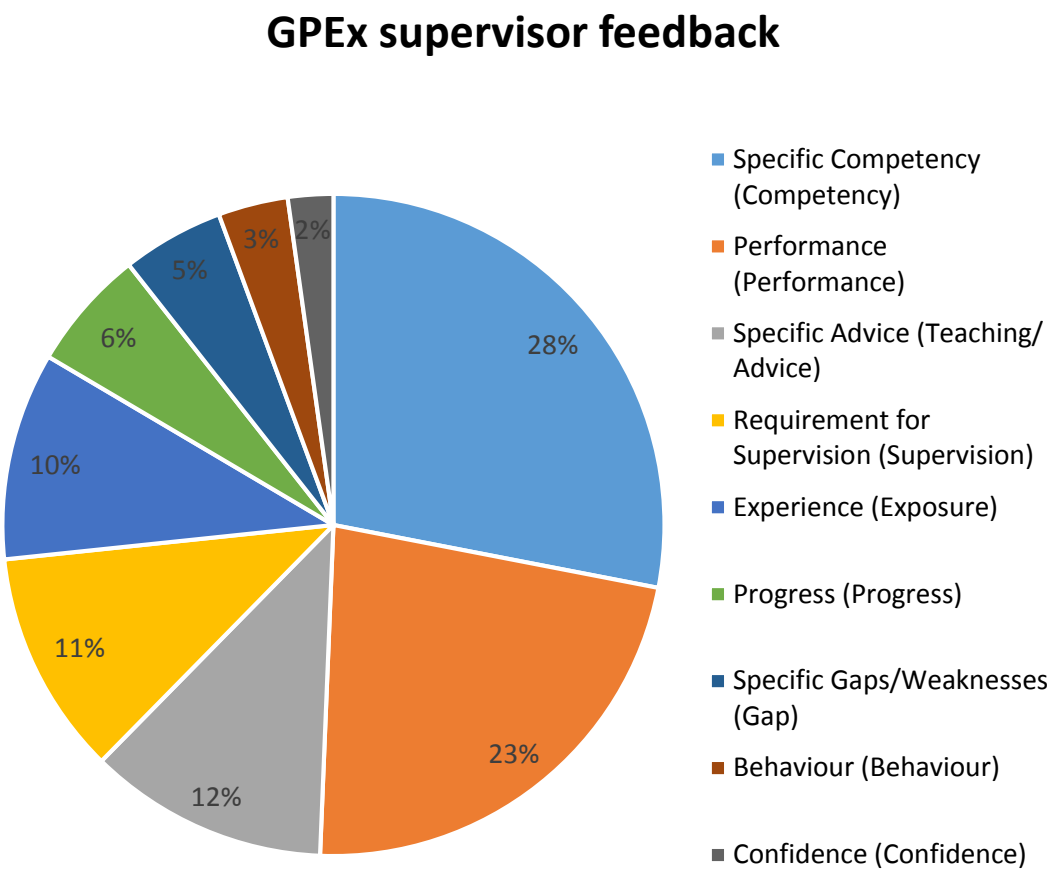


Figure 5. GPEx supervisor feedback to all registrars.

Similar to the GPEx registrars, feedback from WAGPET supervisors to junior registrars focussed on specific competency (15%) and performance (24%) with significant numbers of comments related to specific advice (14%), progress (13%), and specific gaps and weaknesses (12%). 26% of the feedback related to specific advice or gaps.

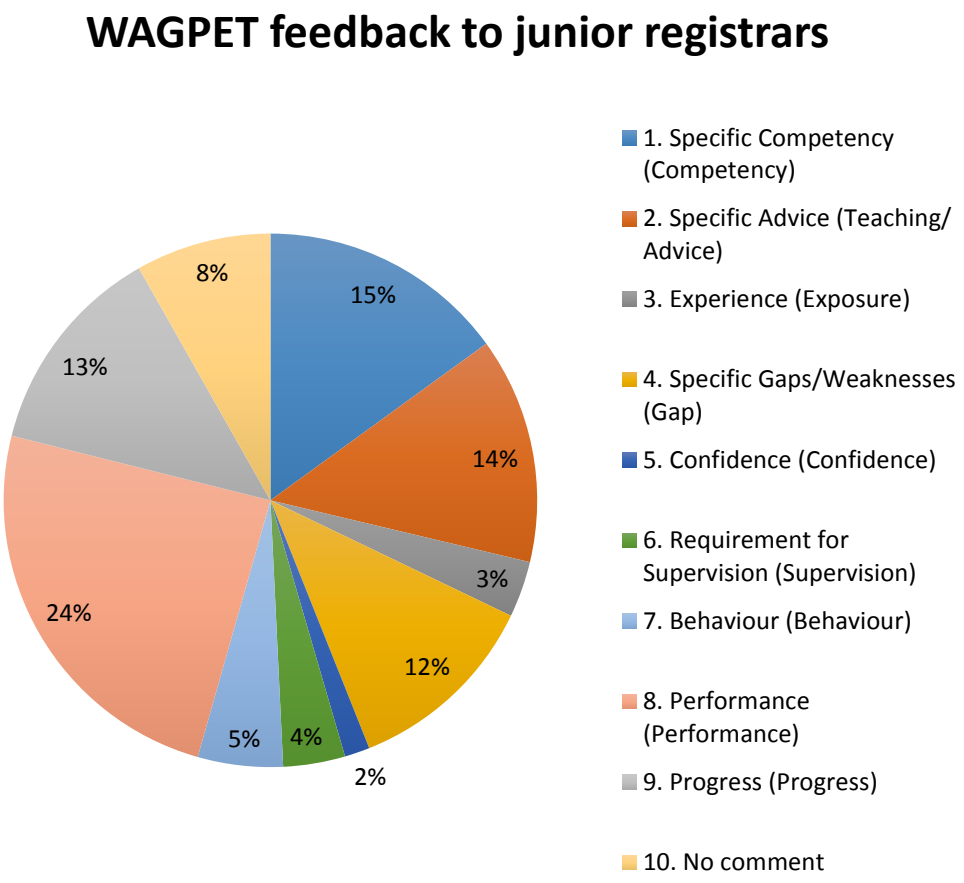


Figure 6. WAGPET supervisor feedback to junior registrars.

By comparison the feedback to senior WAGPETregistrars focussed more on Performance (52%).

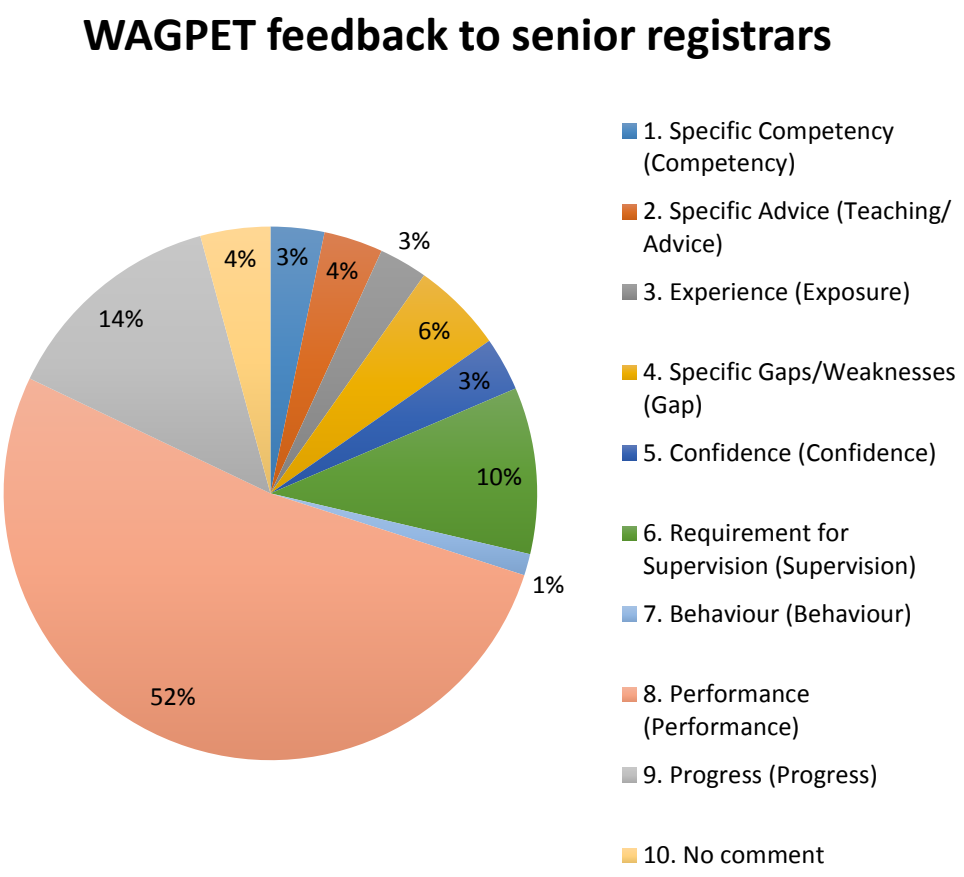
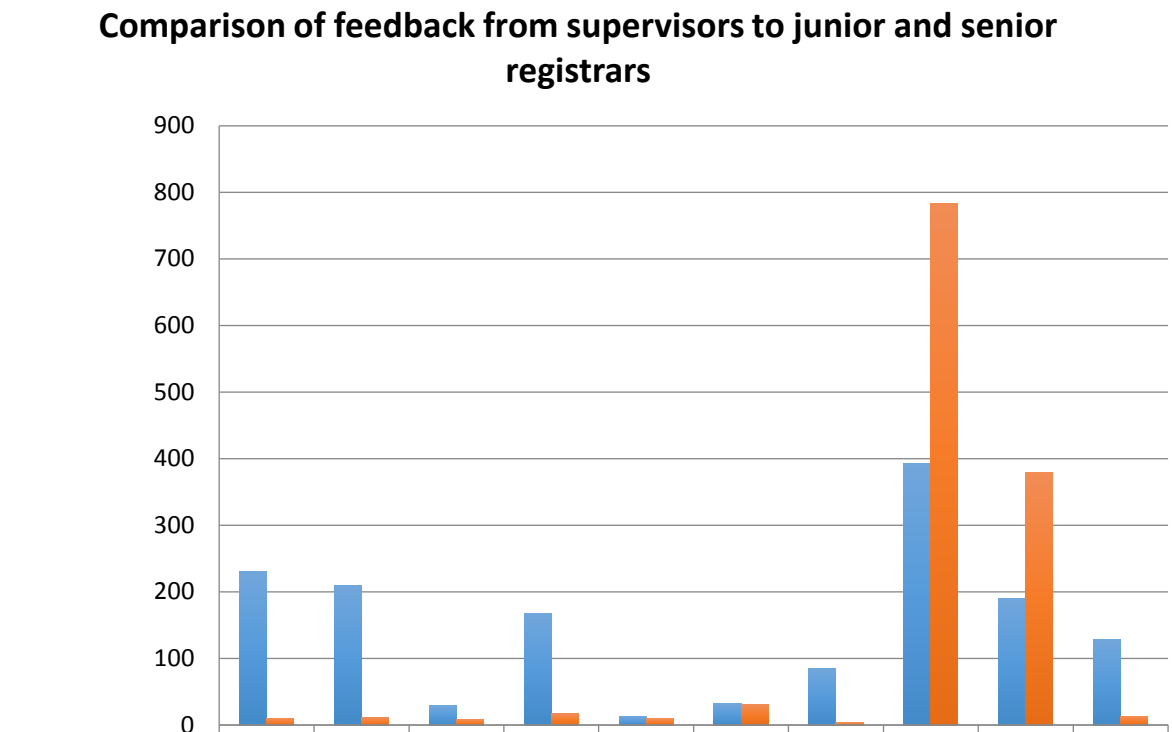


Figure 7. WAGPET supervisor feedback to senior registrars



	1. Specific Competency (Competency)	2. Specific Advice (Teaching/ Advice)	3. Experience (Exposure)	4. Specific Gaps/ Weaknesses (Gap)	5. Confidence (Confidence)	6. Requirement for Supervision (Supervision)	7. Behaviour (Behaviour)	8. Performance (Performance)	9. Progress (Progress)	10. No comment
Junior registrars	230	210	29	168	13	32	85	393	190	128
Senior registrars	10	11	9	17	10	31	4	783	379	13

Figure 8. Comparison of feedback from supervisors to junior and senior registrars at WAGPET.

The comments from GPTT could not be sub-grouped into those from junior and senior registrars.

When all of the supervisor comments from GPEx, WAGPET and GPTT were collated, it can be seen that Specific Competency (20%) and Performance (20%) were the most likely content of the feedback from supervisors to registrars. 15 of the comments overall related to Specific Gaps/Weaknesses and Specific Advice.

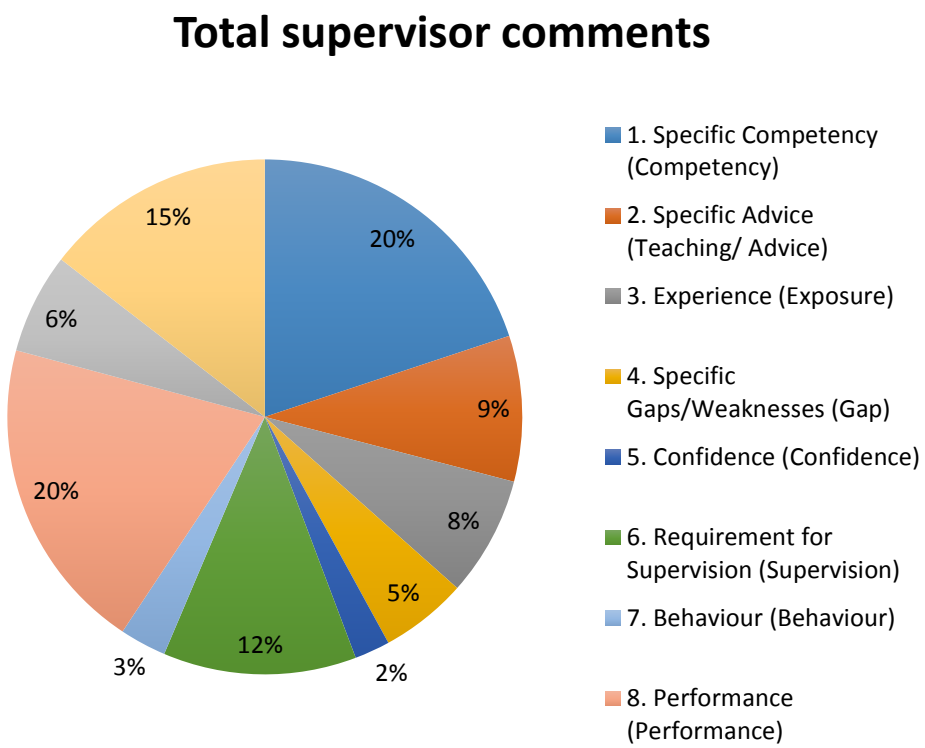


Figure 9. Collation of all supervisor comments to registrars from GPEx, WAGPET and GPTT.

Registrars from GPEx, WAGPET and GPTT were all required to complete a self-reflection using the same EPAs for their mid and end-term assessments. The GPEx registrar self-reflections were not available at the time of writing. The GPTT self-reflections could not be sub-grouped into junior and senior registrars but are included in the total below. The WAGPET registrar self-reflections were sub-grouped into those from junior and senior registrars as well as being collated with those from GPTT to give an overall picture.

The striking difference about the self-reflections by the WAGPET Junior registrars is that 34% of them related to Specific Gaps/Weaknesses. As with the supervisor comments, 22% related to Competency and 11% to Performance, but registrars also reflected more on their lack of Experience (13%) and Requirement for Supervision (13%).

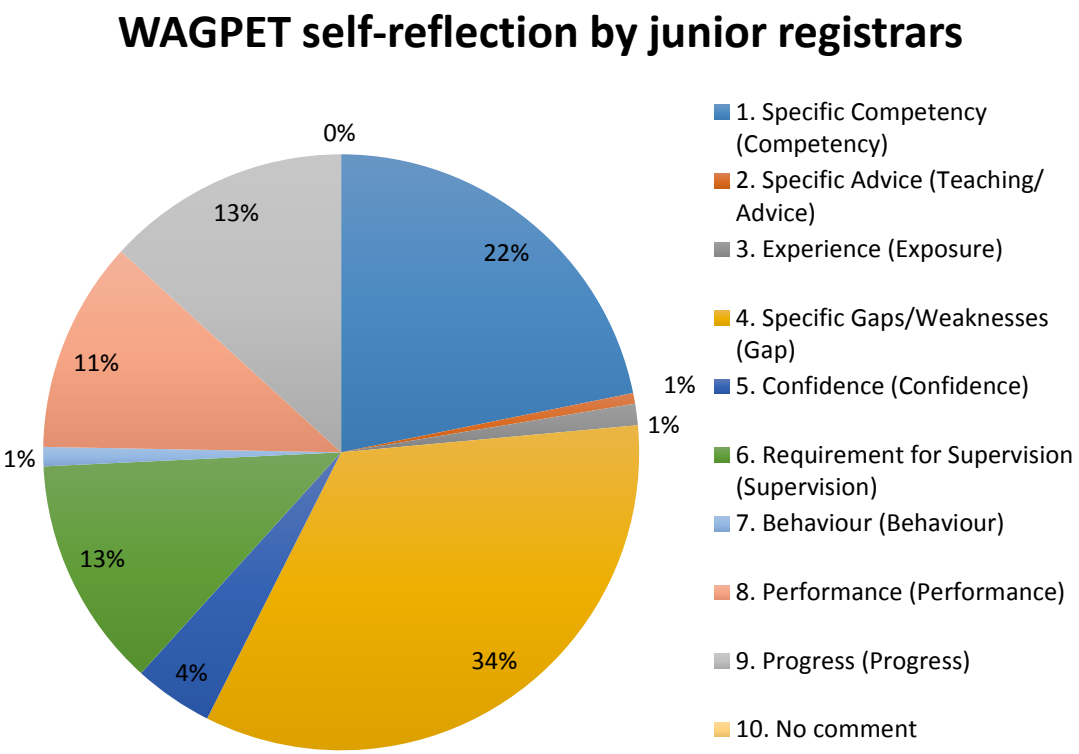


Figure 10. WAGPET self-reflection by junior registrars

The WAGPET senior registrars tended to mainly comment on Performance (44%) and Specific Competency (17%), with 11% commenting on their Progress and 10% making No comment. 7% reflected on Specific Gaps/Weaknesses.

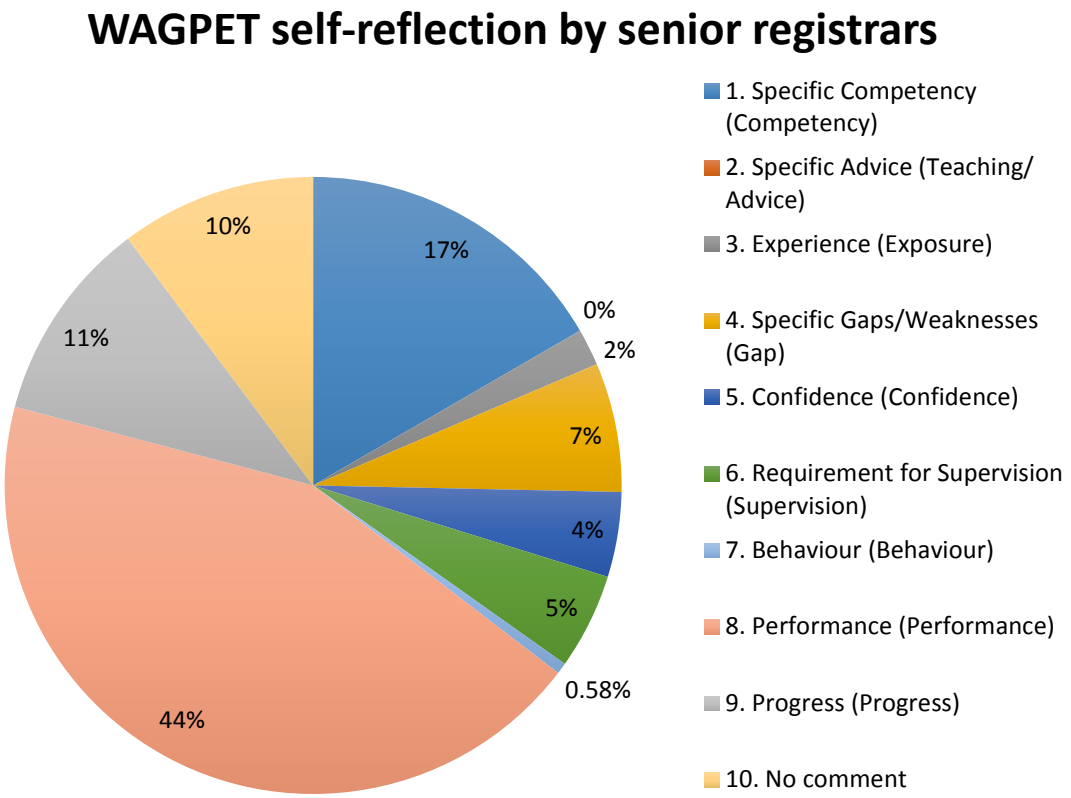


Figure 11. WAGPET self-reflection by senior registrars

When the comments from senior and junior registrars are compared, it is obvious that junior registrars used the EPA template as an opportunity to reflect on their learning needs which they would then discuss with their supervisor.

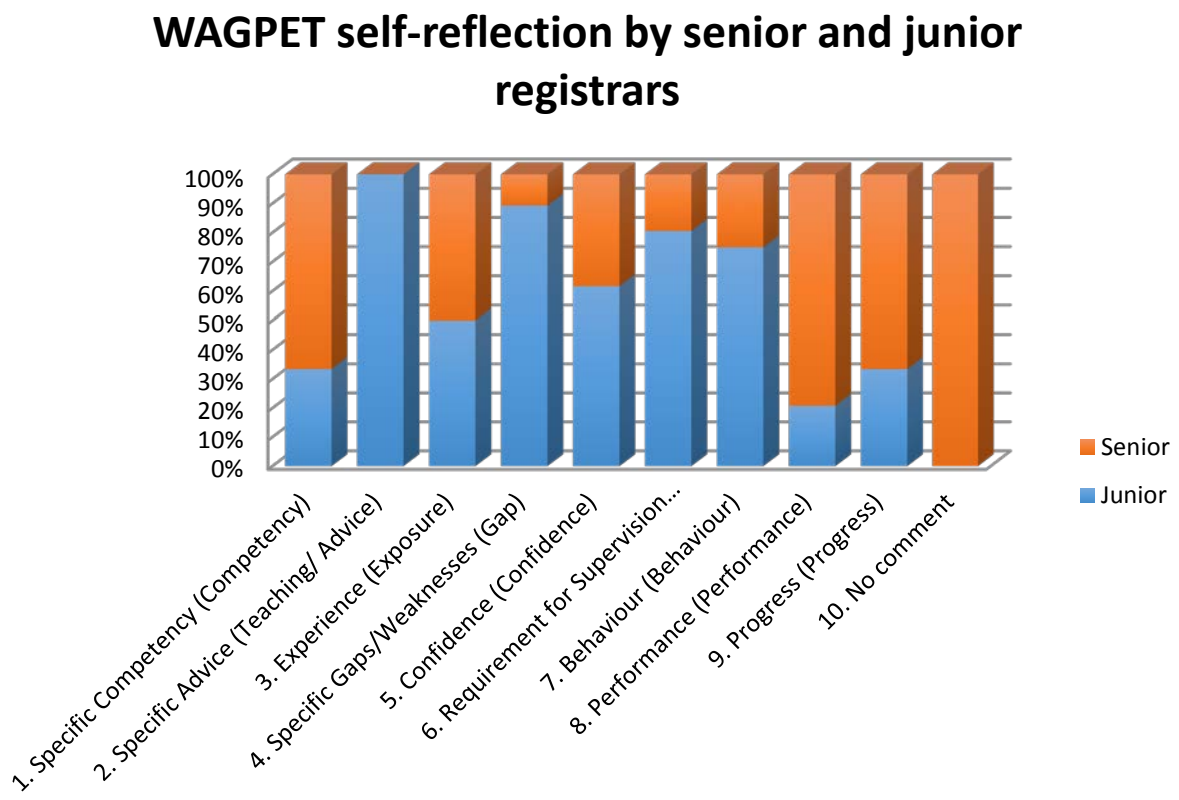


Figure 12. Comparison of self-reflection by senior and junior WAGPET registrars.

When the registrar self-reflections from all the WAGPET and all the GPTT registrars are collated, almost half (49%) of the comments related to Specific Competency or Performance. 16% of the comments identified Specific Gaps/Weaknesses and a further 16% reflected on their Progress.

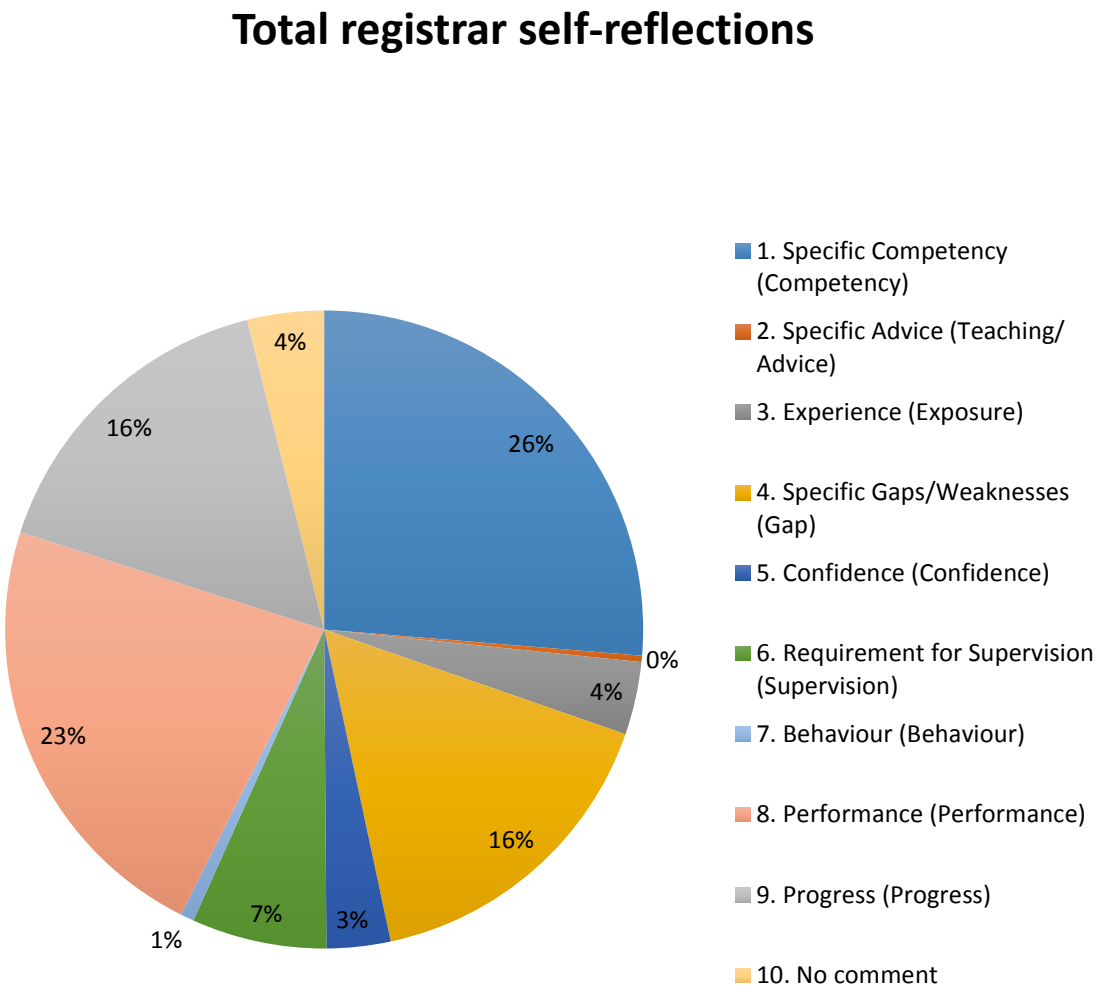


Figure 13. Total registrar self-reflection for WAGPET and GPTT.

When the feedback comments are compared with the registrar self-reflections it can be seen that registrars are more interested in exploring Specific Gaps/Weaknesses and value guidance from their supervisors.

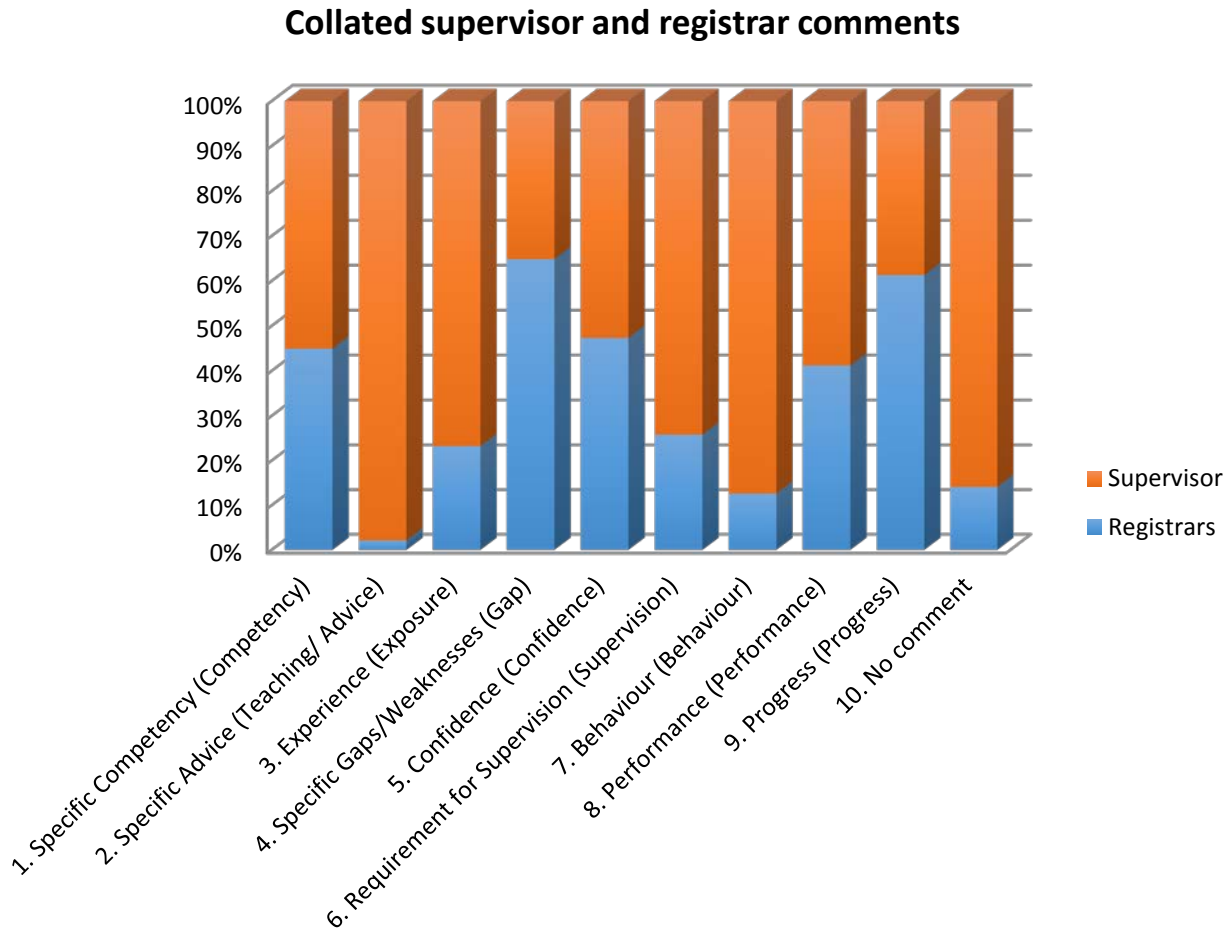


Figure 14. Comparison of total collated supervisor feedback comments and registrar self-reflection.

Specificity and quality of comments

A comparison was made of the specificity and quality of the feedback comments.

37% of GPEx supervisor comments were non-specific in nature, with a greater majority of GP supervisors providing non-specific comments for senior (Year 2) GP registrars (41%) than for junior (Year 1) registrars (22%).

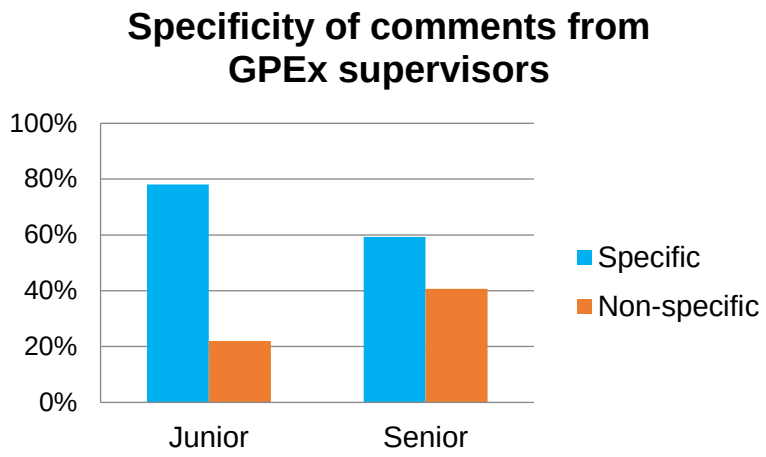


Figure 15. Specificity of comments for GPEx supervisor feedback.

This was similar for WAGPET. The self-reflections from registrars were also more likely to be specific than the feedback from supervisors.

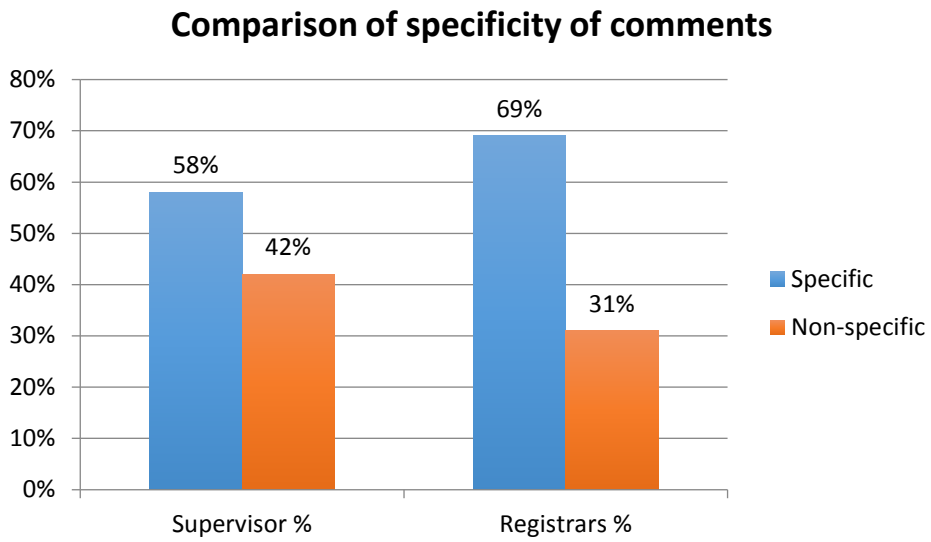


Figure 16. Percentage of collated comments for registrars and supervisors from WAGPET and GPTT that were specific as compared to non-specific.

Comments were also coded in terms of whether the statement was high, moderate or low-quality feedback.

23% of the EPA comments from GPEx supervisors were rated as low-quality in terms of feedback, with greater than two-thirds (77%) of supervisor comments being rated as moderate to high-quality

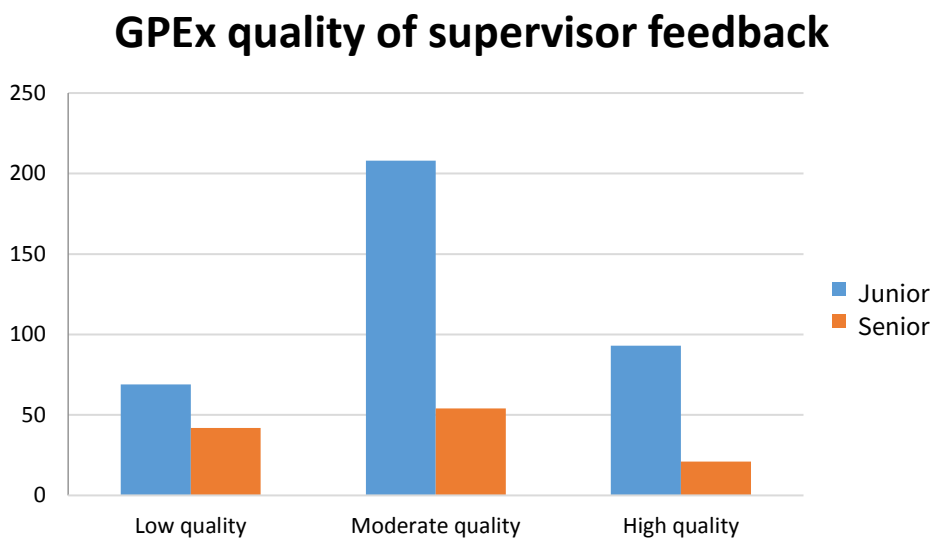


Figure 17. Quality of feedback from GPEx supervisors for junior and senior registrars.

The collated results from WAGPET showed a similar pattern to that of GPEx with more high-quality feedback for junior than for senior registrars.

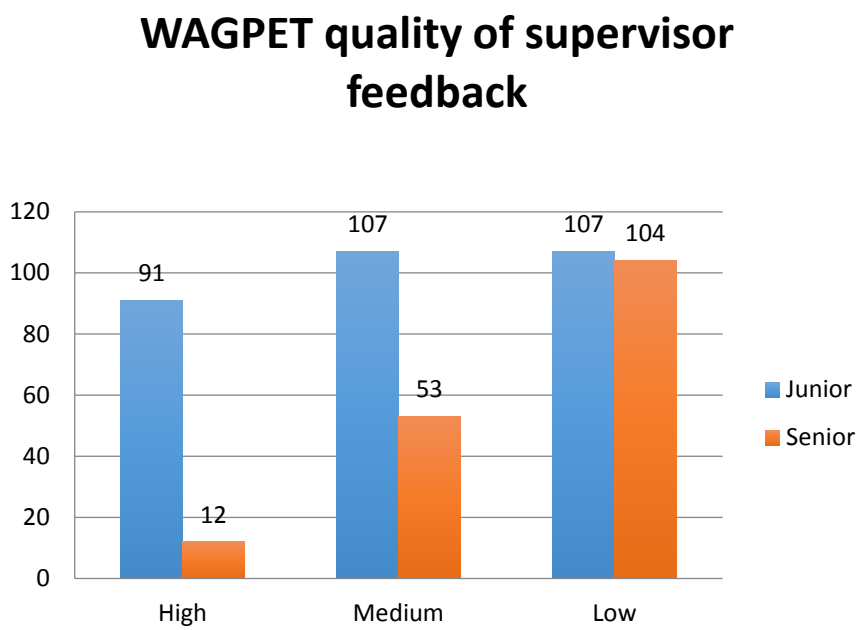


Figure 18. Quality of feedback comments from supervisors for junior and senior registrars at WAGPET

The supervisor and registrarcomments from GPTT could not be sorted into junior and senior registrars.

The registrar comments from WAGPET showed that junior registrar's self-reflection comments were also mostly (76%) of medium to high-quality.

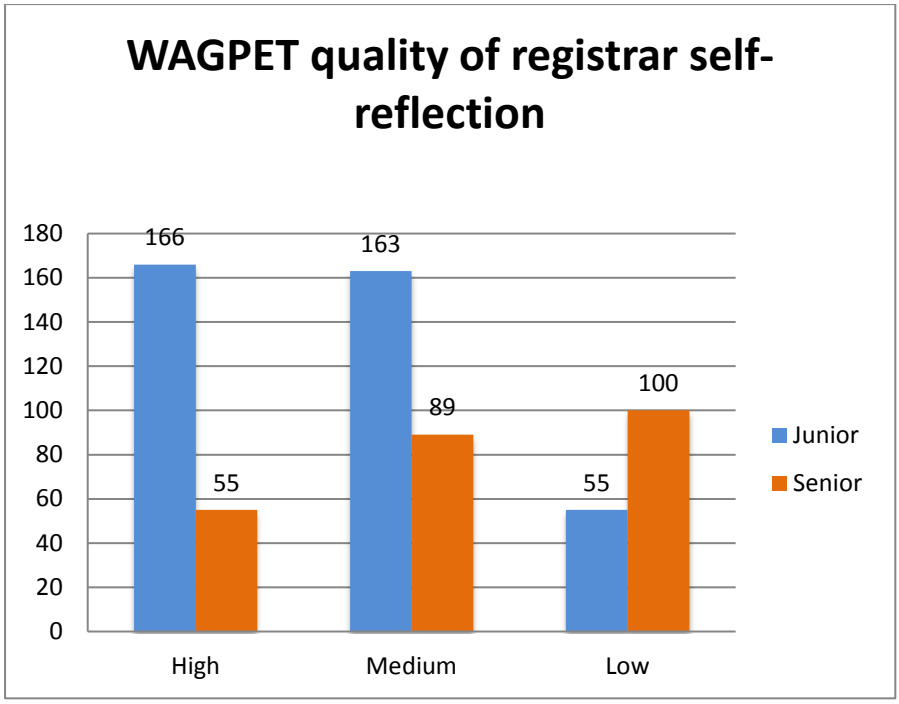


Figure 19. Quality of self-reflection comments for junior and senior registrars at WAGPET.

When the collated comments for WAGPET and GPTT for supervisor feedback and registrar self-reflection are compared it can be seen that the registrars are taking their self-reflection very seriously and are providing medium to high-quality comments in order to identify gaps and plans for improving in their weaker areas.

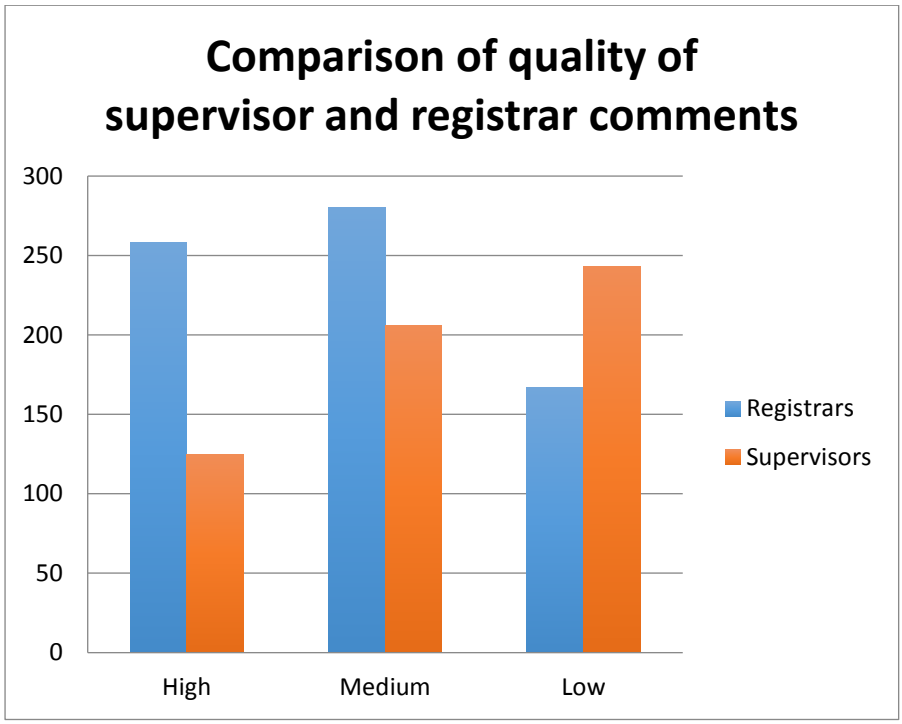


Figure 20. Collated comments from WAGPET and GPTT comparing supervisor feedback and registrar self-reflections on EPAs.

Focus Groups and Interviews

The focus group participants and interviewees generally expressed positive views about the EPAs, especially in comparison to previous assessment forms.

Benefits

Most supervisors, registrars and MEs were able to grasp the concept of global trust and apply it to the way that registrars progressed through their training.

Sup: "*The trust isn't necessarily just about the knowledge level. It's about the ability to ask for help at the appropriate level, right, so that's the whole thing of how close your supervision needs to be. That's not saying I trust you because you know lots or not, it's knowing what you don't know and being able to ask for help at the appropriate point.*"

In summary the benefits of the EPA workplace-based assessment tool were described as:

- Empowering for both the registrar and the supervisor
- Detailed, specific, realistic, clear, comprehensive, easy to navigate and reflecting the RACGP domains and curriculum.
- Providing a more global view of how the registrar is going.
- Helping to gather intuitive thought processes and formulate them into words
- Encouraging clear comments and a narrative about why the registrar has been rated the way they have been.
- Individualising the learning pathway for each doctor.
- Identifying if there was a significant differential between knowledge-based competencies, confidence and access.
- Providing the ability to assess professionalism and working in a team as well as knowledge.
- Opening up opportunities for conversation between registrar and supervisor about which areas the registrar found more challenging and needed more support or teaching in.
- Enabling the supervisor to identify gaps, particularly knowledge gaps, and where to direct their energies so as to facilitate more learning opportunities.
- Assisting in mapping registrar progression.
- Helping in identifying what additional support the registrar needs to progress – where their placements might be in the future, where their strengths are, where their weaknesses are.
- Being particularly useful in identifying struggling registrars who might even have passed their exams but in other areas were not assessed as being safe.
- Helping registrars self-reflect on where their weaknesses are.
- Assisting registrars to feel emboldened to call on supervisor for assistance if they didn't feel confident with a patient.
- Assistance with concordance between the supervisor and registrar about the level of entrustment of the registrar.
- Identifying registrars who lacked insight and thought they had greater skills than they actually did.
- Helping training organisations to identify gaps so they could steer the registrar to a practice in the future that might help fill that gap e.g. children, women's health, mental health
- Helping to diminish paperwork and administrative processes
- Assisting a group of supervisors in developing inter-rater reliability and peer support about how to discern when a registrar is trustworthy enough, in supervisor training.
- Providing potential for the EPA concept to be used in Direct Observation Visits

EPAs were especially identified as improving the ability to map progression in a way that previous forms were unable to. This involved training both supervisors and registrars that it is expected that registrars begin by needing more supervision.

ME: "*Why would we accept somebody, a surgical trainee, who'd been in a position for three months, why would we expect them to be a surgeon? We wouldn't. So, as GPs, why do we feel bad, if we are assessing our registrars as only a one or two on the entrustment scale? I don't think we're bad, I think we're honest…we're not expecting them to get threes and fours early in their training."*

Both registrars and supervisors identified that once registrars had reached the 'ceiling' or 'flatline' of being able to practise unsupervised, then the EPAs became repetitive and redundant. Several participants suggested that once they reach this stage there could be another level for more specialised areas or they could be assessed as ready to teach and cease doing EPAs.

The giving and receiving of feedback is an essential component of training. As well as feedback about gaps and progression, EPAs were identified as providing prompts and structure to encourage useful written and verbal feedback. Some supervisors said they were more comfortable writing things down than talking about them, and some more comfortable talking than writing. MEs additionally found it helpful when reading assessments if there was adequate narrative written.

ME: "*It's the words that give me, as a medical educator, that big picture. A ticking the box, or a number, or a score or whatever, gives me the bare bones, but to be able to see the descriptions - that helps to flesh out that big picture."*

Challenges and limitations

The wording of the non-clinical EPAs did not seem to suit some people, but most embraced the concept and could see that consistency was important. Registrars generally do not like doing learning plans and did not think that EPAs were useful in documenting learning plans, but did generally think they were useful in identifying learning needs for both themselves and the supervisors in planning their teaching. Some registrars, supervisors and MEs thought the EPA assessments were more time-consuming but did acknowledge that this was because they were more comprehensive.

ME: "*They do take more time to fill in for the registrar, for the supervisor and then for them to be able to sort of have that discussion. From the ME point of view, they do take longer to mark, to actually go through. But you do get a lot of feedback you have to read and so it was quite obvious if there was something you had to follow up on.*"

ME: "*That's why I like the EPAs, because it makes you think differently about the practitioner, trust them to go forward. So, that's why I'm bit of a fan of them and it takes you longer to write a good report."*

Because of the relationship between the supervisor and registrar, supervisors were sometimes reluctant to mark the registrar with a 'low' score.

Supervisor: "*I don't know how good an assessment it is because you know they are reading it and you are working with them, so they're colleagues. You are hardly going to have a crack at them. In an assessment that's going to some anonymous agency, you are going to fill it out. If you've got a real concern, you go and talk to them. You are not going to put it on an assessment form."*

This tension of being an assessor as well as an educator was seen as a difficulty in registrars and supervisors. Some thought that the EPAs helped them resolve this tension better than previous forms that emphasised a comparative standard, benchmark or expectation. Partly this was due to an increased ability to generate discussion and written narrative as feedback.

Supervisor: *"I think people get too much criticism in their life. But if somebody isn't performing at the standard I have to say why. I have to think very carefully how that's put and why I think that. Then I also feel that I must have a solution to the reason why I think that that person isn't performing at their standard."*

Some supervisors struggled with filling in the assessments online and asked their registrars to help them. Registrars were well aware that it was not just about the form, that the supervisor's training and engagement with the process was important. Supervisors themselves also acknowledged that unless they were interested in teaching that any form was likely to just become a tick-box.

Supervisor: "*My feeling is that no matter what method you use you've got to remember that as a supervisor you, in my opinion, should be addressing everything critically so that it's a real learning experience for the registrar. Otherwise you might as well just tick it and just say everything's all right."*

In outlining the challenges of using the EPAs, one of the main limitations was a lack of training for both registrars and supervisors. Supervisors who began using the assessments without training were reluctant to discontinue their previous practice of "*marking them all as fours because they didn't want to upset them*" (ME). Others discussed a "trial and error" learning curve that might have been avoided if they had been trained in both the educational philosophy as well as the practicalities of using the assessment.

Supervisor: "*I think if we knew that this was the way the registrars were going to be assessed and the reasons why and the educational reasons why it's a powerful tool, then after two or three runs through, we would have it under our belt and we will become quite comfortable with it".*

The registrars also need training on what they should expect from the EPA tools.

ME: "*And I think reminding the registrars that we expect them to start at that less competent level, and to progress. And that we're not saying that they're a bad person, or they're a bad doctor. But that, you know, they're developing those skills and that confidence. As long as they are progressing."*

Registrar self-evaluation

The registrar EPA self-evaluation (the same as the supervisor's) was mostly well-received. Registrars use it to guide their own learning and to open up discussion with the supervisor. Most acknowledged that they were more critical of themselves than the supervisor. MEs discussed how they could use the EPAs as a way of identifying a mismatch of supervisor and registrar assessments so that the underconfident or overconfident registrar can have support and strategies put in place to assist them.

Some registrars, as well as some supervisors, were reluctant to be honest in the assessments because of the registrar/supervisor relationship, or the desire to continue working in the practice – the supervisors did not want to upset the registrars and the registrars did not want to expose their weaknesses. Both were also concerned that if there were an assessment that the registrar needed more supervision in certain areas, that this would not be logistically possible. Some felt there was a medico-legal risk with saying that more supervision was needed and then not being able to

supply this supervision. Similarly, some registrars were reluctant to say that they needed more supervision for medico-legal reasons.

Registrar: "*…if we say the registrar needs more supervision it's not like the supervisor can come into your room and make sure you're doing that. There's no result from that."*

How to determine levels of entrustment

For many supervisors how an entrustment decision was made was not an intuitive process and many registrars thought that their supervisor did not know enough about how they were working to make an entrustment decision.

Registrar: "*…how accurate can a supervisor comment on so many areas of general practice when in actual fact… I don't think my supervisor has seen me even consult in one of the 100 consultations… So I think in order for a supervisor to comment accurately on the EPA, they actually have to spend a substantial amount of their time to actually see us consult and determine whether we're making the right decision or not. If they can't do that then it's not an accurate EPA. It would be based on how they feel about the registrar. It would be based on how many times the registrar has asked them questions about certain cases and so on and so forth."*

Supervisor: "*I think the most difficult thing is to actually know really whether the registrar can actually do those things. Their experience may not yet have covered some of those things and just knowing what the registrar has been doing or not doing is very difficult to know. You aren't sitting in with them long periods of time. You aren't analysing case-by-case what's happened with all their patients. It's too arduous and there is no time for that practice to occur."*

There was sometimes a lack of understanding on both sides that entrustment is a global assessment collated from many different sources. Others embraced this concept eagerly and could articulate the various ways they gathered information and made their entrustment decisions, for instance by direct observation; the types of questions the registrar asks and how often; review of patient files; discussions with other GPs in the clinic, receptionists, nurses and patients; and formal teaching time.

Supervisor: "*If you expose yourself to a situation for a reasonable amount of time, like watching a registrar or talking about a registrar or thinking about what a registrar has done, you are gathering multiple data points constantly."*

ME:" *I think one of the differences also as a supervisor, you've got a longitudinal basis and you've got the whole week. So you know how many times they ring you, you know what they're asking, are they asking you the same question twice…I think you've got a better understanding of the depth of knowledge and the clinical reasoning…"*

Supervisor: "*Just because we're intuitive people, you get a good idea within just going through that registrar straight away what sort of person they are. What sort of doctor they are. How they practice, their personality type. That will then inform you that you need to keep hovering over them or just let them go a little bit."*

Supervisor: "*I will often tell the registrar at the end is I would trust you to look after my family or my children and I do, I use the word trust a lot."*

Registrar: *"They certainly stimulated discussion between me and my supervisor about which areas I found more challenging…it was a reflection of, you get a presentation, when do you feel*

uncomfortable and you need to talk to somebody and how urgently do you need to talk to somebody and so I felt like it …asked a question in a way that made it easy for me to kind of cognitively put myself into the picture and go, yes I would want to call my supervisor in that kind of scenario."

Discussion

The goal of general practice training is to educate registrars to be ready to provide safe, unsupervised, professional care by the completion of their training. In this study, the trust of the supervisor progressively increased with time, allowing registrars to progress through levels of entrustment whilst still in a supervised, training environment. From a risk management perspective, the use of EPAs provides guidance regarding tangible parameters for supervision of certain activities.

The mid and end-term assessments by supervisors are an opportunity for a GP who has day-to-day contact with a registrar to reflect on the registrar's safety and learning needs, communicate this to the registrar and to the RTO, and to monitor their progression through their training. Adding to this a registrar self-assessment and reflection, also gives the registrar an opportunity to assess their need for supervision, to reflect on their specific gaps, and to notice their improvement in both safety and learning. Supervisors are prompted by the EPAs to combine their intuition with informal and formal observational data about the registrar. Registrars similarly can view their practice from the perspective of patients, supervisors, and other members of the practice team, as well as reflect on the gaps they have found in their daily patient load. The discussion of the EPAs empowers the supervisors to openly discuss performance and competency without fear of 'failing' the registrar, and the registrar is also empowered to ask for supervision and guidance.

As well as showing a difference in supervision requirements between junior and senior registrars and a progression of need for less supervision over time, there was a significant difference between the concordance of registrar and supervisor assessments over training time. Junior registrars tended to assess themselves as needing less supervision than senior registrars, there being more concordance between the registrar and supervisor assessments in the senior year. This may be that the junior registrars did not yet 'know what they didn't know' – sometimes called 'unconscious incompetence'. As they progressed through training, registrars were more likely to become aware of the gaps in their knowledge and skills and so became 'consciously incompetent'. As their skills and knowledge did then improve, their self-assessment became more realistic.

This research has confirmed that EPAs are indeed a valid global assessment tool to assess supervision needs and progression. The additional data collections from WAGPET and GPTT showed that there is a significant difference in the entrustment levels between junior and senior registrars, growth in entrustment levels over time, and an increase in concordance between supervisor assessment and registrar self-assessment over time. This suggests that initially registrars were unconsciously incompetent, not yet aware of what they did not know. As registrars develop expertise, these unconscious gaps in competency decrease. This also allows for the identification of overconfident registrars, and for learning strategies to be put in place if this is the case.

As the EPAs are mapped to the curriculum, domains and competencies, benchmarking completion of training as needing (at least) safety to practise unsupervised in all EPAs allows standardisation and a clear final outcome of training. The expectation that registrars begin their training needing closer supervision, and complete their training as independent practitioners, decreases the fear of 'failing' and marking registrars as 'below the expected standard' of previous assessments. EPAs

are also able to assess the more difficult areas of the curriculum such as professionalism, time management and team-work, and the feedback and self-reflections for these EPAs identified weaknesses in these areas for both supervisors and the registrars.

One difficulty is the possibility of a 'ceiling' being reached, as high performing registrars will achieve global safety to practise unsupervised at an earlier stage. Adding a fifth level where registrars are encouraged to take on a teaching role as 'safe to supervise junior learners' may go some of the way to generating more engagement with these registrars. As registrars progress in their training they are more realistic about their need for advice from colleagues, and even excellent registrars may assess themselves as still needing supervision as they have not thought through the difference between collegial practice and supervision.

These EPAs are also valid tools for generating rich feedback from supervisor to registrar and useful reflection for the registrars. The majority of the feedback and self-reflection was specific and of high quality. This was more so for the junior registrars when both supervisor and registrar seem more aware of their specific learning needs and gaps. Registrars appear to have taken their task of self-reflection very seriously and seem to value the chance to reflect on their learning needs. However, supervisors may need more training on how to give feedback, particularly for senior registrars where the learning gaps are not so obvious.

Registrars and supervisors commented that the EPA format encouraged comments, discussion and narrative feedback, particularly in comparison to previous assessments. The ability to identify a struggling registrar who might be 'flying under the radar' was identified. Finding learning gaps, in women's or men's health, paediatrics or mental health for instance, can facilitate targeted teaching. More importantly, registrars who are recognised as requiring closer supervision can be flagged. It is important if this is the case that the training organisation supports the supervisor to provide the supervision needed.

Conclusion

This comprehensive research across three Australian states has confirmed that the mid and end-term assessments using EPAs that have been used at GPEx since 2017, are a valid assessment tool in Australian general practice. Not only do they facilitate assessment, teaching and learning, they also demonstrate competency progression throughout training, are an effective tool to stimulate registrar self-reflection and are well-accepted by both registrars and supervisors. Their ability to globally assess safety collating multiple modalities, and their prompting of self-reflection and identification of gaps for registrars, means they are ideally placed as a robust workplace-based assessment tool in identifying whether registrars are safe, competent and self-reflective GPs.

References

Bonnie L, Visser M et al (2019) Trainers' and trainees' expectations of Entrustable Professional Activities in a primary care training programme. Education for Primary Care 30:1, 13-21

Jackson J, Kay C, Jackson W, Frank M (2015) The Quality of Written Feedback by Attendings of Internal Medicine Residents. Journal of General Internal Medicine 30:7, 973-978

O'Dowd E, Lydon S, O'Connor P, Madden C, Byrne D (2019) A systematic review of 7 years of research on entrustable professional activities in graduate medical education, 2011-2018. Medical Education 53: 234-249

Peters H, Holzhausen Y et al (2017) Twelve tips for the implementation of EPAs for assessment and entrustment decisions. Medical Teacher 39:8, 802-807

Rekman J, Goften W et al (2016) Entrustability Scales; Outlining Their Usefulness for Competency-Based Clinical Assessment. Acad Med. 91:2, 186-190

Schultz K, Griffiths J, Lacasse M (2015) The Application of Entrustable Professional Activities to Inform Competency Decisions in a Family Medicine Residency Program. Acad Med 90:7, 888-897

Shaughnessy A, Sparks J et al (2013) Entrustable Professional Activities in Family Medicine. Journal of Graduate Medical. March 2013: 112-118

ten Cate O and Scheele F (2007) Competency-Based Postgraduate Training: Can We Bridge the Gap between Theory and Clinical Practice? Acad Med 82:542-547

ten Cate O(2013) Nuts and Bolts of Entrustable Professional Activities. Journal of Graduate Medical. March 2013: 157-158

ten Cate O, Chen H et al (2015) Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. Medical Teacher 37:11, 983-1002

ten Cate O, Hart D et al (2016) Entrustment Decision Making in Clinical Training. Academic Medicine 91:2, 191-198

ten Cate (2018) A primer on Entrustable Professional Activities. Korean J Med Educ. 30:1, 1-10

Valentine N, Wignes J et al (2019) Entrustable Professional Activities for workplace assessment of general practice trainees. MJA 210:8, 354-359

Stream 3d
An exploration of registrar flagging models and their association with RACGP exam performance

Mr Shaun Prentice (GPEx)
Associate Professor Jill Benson AM (GPEx)
Professor Lambert Schuwirth (Prideaux Centre, Flinders University)
Dr Emily Kirkpatrick (GPEx)

In collaboration with

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

GPEx would also like to acknowledge:

- The Working Group for this stream: Mr Shaun Prentice (GPEx), Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr Cat Kirby (Eastern Victoria GP Training), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx) and Ms Michelle Pitot (GPEx).
- The collaborating Australian general practice training organisations: Eastern Victoria GP Training (EV), General Practice Training Tasmania (GPTT), Murray City Country Coast GP Training (MCCC), Northern Territory General Practice Education (NTGPE), Western Australian General Practice Education and Training (WAGPET), Remote Vocational Training Scheme (RVTS), GP Synergy and General Practice Training Queensland (GPTQ) for their support and contributions to this research.
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.

GPEx In collaboration with

GP training · Flinders University · general practice training tasmania · MURRAY CITY COUNTRY COAST GP TRAINING · NORTHERN TERRITORY General Practice Education

RVTS REMOTE VOCATIONAL TRAINING SCHEME LTD · WAGPET · GP SYNERGY · GPTQ

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Introduction

Given the considerable investments made by a variety of stakeholders in the education of medical trainees, helping trainees to thrive throughout their training is an important endeavour. In the Australian general practice context, this involves helping registrars to develop into safe, independent practitioners who are lifelong learners. The major barrier to the culmination of this process is the series of final exams registrars must sit, which can be seen as a final decision point to determine whether they are ready for safe and independent practice. A critical role of the education process is therefore the early identification of any issues that may threaten a registrars' chances of successfully completing the training program. If such early warning signs are detected, remediation programs can be implemented to help the registrar improve during training. This 'preventive medicine' approach to medical education, which examines early warning signs should therefore be a priority.

We are aware of two published studies examining predictors of RACGP exam performance. In one study, Patterson et al. (2016) identified that pre-selection assessments (the situational judgement test, multiple-mini interview and overall selection score) yielded positive correlations (*r* range = .12 to .54) with each of the three RACGP exams, the Applied Knowledge Test (AKT), Key Feature Problem (KFP) and Objective Structured Clinical Examination (OSCE). In hierarchical regression analyses, these pre-selection assessments explained between 3% and 29% of variance in exam performance. Similarly, a retrospective cohort study by Stewart et al. (2018) identified that pre-training variables (i.e. selection band and pre-commencement assessment) were significantly associated with passing all exams in univariate, but not multivariate, analyses. Furthermore, these pre-training variables were also predictive of KFP and OSCE performance in univariate and multivariate analyses. Interestingly, the two during-training variables examined (external clinical teaching visits and term reports) demonstrated no relationship with exam performance. Conversely, another study identified that 'in-vivo' assessments (i.e. workplace-based assessments – WBAs) during training were significant predictors of the need for formal in-training remediation but that 'in-vitro' assessments were not (Magin et al., 2017). Collectively, these findings suggest that pre-admission assessments are, to an extent, predictive of RACGP exam performance, while WBAs are predictive of the need for remediation.

Before the point of formal remediation, though, there are systems of 'flagging' within Australian Regional Training Organisations and the Remote Vocational Training Scheme (hereafter 'RTOs'). These systems take a preventive approach by identifying registrars in difficulty and providing them with necessary supports to help them overcome these difficulties. Over the course of this ERG, it became evident that there is great variety in how 'flagging' is operationalised and implemented across Australian RTOs, however the effectiveness of these systems is unknown. Understanding whether flagging is predictive of exam performance would therefore be both theoretically and practically informative. From a theoretical perspective, flagging is a broad approach designed to identify problems in a registrar's progress using information from formal and informal sources. It therefore may include information unavailable within the realm of formal assessments both prior to, and during, training. From a practical perspective, understanding the effectiveness of flagging systems may help to further improve them, thereby improving the capacity for RTOs to prepare registrars to pass their exams and help them develop into safe, independent practitioners who are lifelong learners.

Originally, this stream of the project was intended to be a systematic review and meta-analysis of published and unpublished literature exploring the relationship between WBAs and RACGP exam performance. However, literature searches identified that there was insufficient literature to meaningfully pool data. Additionally, given the aforementioned findings of the literature, it became questionable whether relationships would emerge between WBAs and exam

performance. Considering the potential for important relationships to exist between the broader system of flagging and exam performance, it was decided to refocus this stream on flagging. Accordingly, this stream of the project sought to explore flagging in the context of Australian RTOs. Given the great diversity in flagging systems across Australian RTOs, the first aim of this stream was to develop an understanding of the nature of flagging in different RTOs and to identify similarities and differences between organisations. The second aim was to pool data from Australian RTOs to explore relationships between flagging and exam performance.

Methods

The present project comprised two inter-related components. The first involved qualitative analysis exploring the models of flagging in Australian RTOs. The second involved a meta-analysis of data examining the relationship between flagging and exam performance. Exploratory subgroup analyses and meta-regressions in the meta-analysis were informed by the findings of the qualitative analysis. The details of each component of this study are detailed below.

Qualitative Analysis

Data collection

When GP training organisations were asked for data examining the relationship between flagging and exam performance, difficulties were identified in producing a sufficiently homogenous dataset. Simultaneously, as data was collected as part of Stream 3b, the diversity of flagging systems in training organisations across Australia became evident.

Given the importance of understanding the diversity of flagging systems, interviews were conducted with a Senior Medical Educator or Director of Education/Training in each GP training organisation in order to better understand the flagging procedures in that organisation. Interviews were unstructured, guided by the objective of understanding the flagging system in each GP training organisation. Interviews were audio-recorded and transcribed. Information regarding flagging was also extracted from the data collected in Stream 3b from registrars, supervisors and MEs.

Data analysis

This data was then analysed using inductive thematic analysis using a grounded theory approach to build a richer picture of conceptualisations and processes of flagging from the perspectives of the GP training organisations, the medical educators (MEs), the supervisors and the registrars. Additionally, features of RTO flagging systems were identified and categorised. These features were mapped against the exam results from seven RTOs to allow statistical analysis in the meta-analysis arm of this stream.

Meta-Analysis

Data collection

Data reporting the flagging history[8] and exam outcomes for registrars who sat any RACGP exam in 2018 were collated by seven RTOs. This data was sent to the lead author who then transformed it into a series of 2 x 2 tables to represent the following twenty-eight relationships in each training organisation:

[8] Flagging history data excluded minor flags and flags for exam failure

	Failing any exam	Failing the AKT	Failing the KFP	Failing the OSCE
Flagging (at any stage of training)	Relationship 1	Relationship 2	Relationship 3	Relationship 4
Receiving a flag in multiple training terms	Relationship 5	Relationship 6	Relationship 7	Relationship 8
Receiving the first flag in GPT2	Relationship 9	Relationship 10	Relationship 11	Relationship 12
Receiving the first flag in GPT3	Relationship 13	Relationship 14	Relationship 15	Relationship 16
Receiving a flag in GPT1	Relationship 17	Relationship 18	Relationship 19	Relationship 20
Receiving a flag in GPT2	Relationship 21	Relationship 22	Relationship 23	Relationship 24
Receiving a flag in GPT3	Relationship 25	Relationship 26	Relationship 27	Relationship 28

Categorical or continuous data for each GP training organisation based on features identified from the qualitative analyses and Stream 2 were recorded alongside the 2 x 2 tables to permit subgroup analyses.

Data analysis

All data was entered into Comprehensive Meta-Analysis (V3.0; Biostat Inc., 2014) from which Relative Risks were meta-analysed using a random-effects model (Borenstein et al., 2009). Heterogeneity was assessed using Cochrane's *Q* test, T (as an estimate of τ) and I^2 (Borenstein et al., 2009). Meta-analyses were run on all 28 possible relationships. Subgroup analyses with relationships 1-5, 9, 13, 17, 21 and 25 were run with all categorical features to ascertain whether any of these features explained differences between training organisations. The predictive strength of these features were assessed using meta-regressions. As a substitute for subgroup analyses, meta-regressions were run on continuous variables.

Results

Qualitative Analysis

Interviews were conducted with 11 Senior Medical Educators or Directors of Education and Training across 9 GP training organisations specifically for the purpose of this stream (coded as RTO). Furthermore, as part of Stream 3b interviews and focus groups were conducted with 58 registrars, 41 supervisors, and 28 MEs across nine training organisations. From the analysis of these data, the below themes emerged.

Early flagging

A common theme from all groups of participants was the importance of using flagging to identify registrars in difficulty as early as possible:

ME: "The earlier we know about problems the more likely we are to get people passing the exam."

Several RTOs have implemented screening strategies to identify these registrars, including pre-training Multiple Choice Questions (MCQs) and the capacity to flag pre-training issues, such as hospital records or issues arising from the application interview. Specifically, some RTOs have implemented a formal safety assessment in the first 4-10 weeks of training designed to assess a registrar's safety, supervision requirements and need for additional training input to provide a

basis for future monitoring. The assessment usually consists of some combination of a series of questions for the supervisor; an MCQ; a Direct Observation Visit by the supervisor or ME; and a discussion with practice staff. Usually these results are discussed with a group of MEs or senior staff:

RTO: "...so part of that is safety, you know are they actually fit to be left in a room by themselves with the supervisor down the hall..."

Participants felt that Training Coordinators (TCs) who liaise with supervisors, practice managers, MEs and registrars are well positioned to provide information about professionalism issues from an early point in training, particularly regarding paperwork compliance, and that these issues could be effective early warning signs:

RTO: "...they already find early on that administration staff have people on their mind that they are not going to do well and they are often right..."

Although MEs and supervisors discussed the utility of an early assessment, more often they talked about registrars who reached the end of their training with issues that they felt should have been identified earlier. Some MEs and supervisors spoke about registrars whose difficulties were flagged early and went on to have good outcomes, such as exam success, due to additional input, for example closer supervision and further WBAs.

In light of these advantages, some participants discussed factors that affect whether a registrar is flagged. Participants reported that, if the RTO has a good relationship with the supervisors and practice managers, they are more likely to informally flag early. Supervisors and MEs discussed how supervisors can be 'too nice' and are reluctant to flag people early, despite the costs to patient safety of registrars in distress not being flagged and a registrar's risk of exam failure and inability to complete of training.

ME: "I think that all RTOs are probably screaming, please tell us the truth. Tell us early, tell us often. There is always that reluctance."

Diagnostic process

Several RTOs have a formal 'diagnostic process' for flagged registrars, similar to the diagnostic phase in medicine. When a flag is raised, the ME who has followed the registrar throughout their training (assigned ME) will usually first discuss the issue with the registrar and, when relevant, the supervisor. Sometimes this is enough at the time, and the flagged registrar is 'monitored' by the ME (usually the assigned ME). For more serious flags, the ME will 'gather data' from various sources, including previous WBAs, flags, learning plans, and relevant people involved in the registrar's training, for example the training coordinator, supervisor and sometimes the practice manager.

The diagnosis will look at what is underlying the registrar's problems, potentially categorising these:

RTO: "...it's individualised, to work out are we dealing with knowledge, are we dealing with professionalism, interpersonal skills or clinical reasoning or is it a personal health issue."

After the 'diagnosis' is made, the ME will usually have a discussion – much like a 'ward round' – with senior MEs and together they will develop an intervention plan. In most RTOs this is a formal committee that meets solely to discuss flagged registrars.

ME: "Then I present the case starting from this is a GPT1 registrar in such and such an area. This is why she's been flagged. This is what's happened so far. This is how I plan on managing this…Then they kind of act as consultants making recommendations in terms of how this particular registrar should be managed from here."

The ME will then discuss the intervention plan and the WBAs that will be used to monitor the outcomes with the registrar. Monitoring and assessment is usually in a programmatic way using several different WBAs and the ongoing plan continues to be tailored to the registrar's needs.

Remediation ME

Some training organisations have a specific ME who undertakes all internal remediation for registrars who are flagged. This is either for the whole organisation or for a particular sub-region. RTOs and MEs themselves discussed the pros and cons of having a Remediation ME role as compared to the remediation being carried out by the registrar's assigned ME.

Those who are against having a specific remediation role discussed the importance of the ongoing relationship between the assigned ME and the registrar:

ME: "So it's a dual purpose. You get to be better by taking it on yourself and learning more about how to do that but also you think that the relationship with the registrar is more useful for their remediation."

RTO: "...each ME has however many registrars and they follow them through beginning to end and they manage everything..."

These MEs can be supported with resources such as templates with guidance about, for example, using additional WBAs, online courses, closer supervision, clinical teaching visits and communication programmes. As well as having the ongoing relationship with the registrar, they are more likely to have the advantage of a global context of the registrar's training. Such factors contributing to this global context may include having heard from the TCs about any issues; knowing if the registrar has had too many days off work; and having read the registrar's previous WBAs and learning plans.

The support from an ME with whom they have a relationship is also seen by the registrars as one of the most important factors. For some registrars, they built up this supportive relationship with the Remediation ME as well:

ME: "...when we do our completion of training I talk to some of my remediation registrars. They basically say that the number one thing that they really valued was the genuine support and the sense that we've got medical educators who are really trying hard to do things for them and that we genuinely believe in their capabilities of becoming good GPs."

Flagged-issues group or committee

A consistent feature across RTOs regarding flagging was the existence of a group or committee of MEs who dealt with all, or at least all serious, flags. The frequency with which these groups/committees meet varies. In some training organisations, MEs and TCs meet quarterly to discuss all registrars, with fortnightly meetings between Senior MEs. Others meet monthly or bimonthly where the assigned ME discusses flagged registrars with the Senior ME group.

One RTO had previously held regional meetings, but decided to combine these to ensure equitable remediation and consistent flagging standards across regions:

RTO: "...we wanted to make sure that registrars in each region were being treated equitably and that meant really discussing the sorts of issues that were being flagged and the sorts of plans that were in place for remediation…So it's partly sharing best practice and it's partly about benchmarking and making sure things are equitable..."

Participants reported that remediation templates and flowcharts were useful tools for helping assigned MEs to effectively handle remediation. This approach also allowed the committee to predominantly focus on registrars with more serious flags, but with the flexibility to discuss flagged registrars not specifically brought to their attention.

When reviewing flagged registrars, all committees collated WBAs and multiple opinions to inform decisions and plans. Committees that also discussed every registrar at quarterly meetings collated information in the same way.

RTO: "So the kind of information we will be reviewing in that meeting is basically all the WBAs the registrars have done … includes the feedback about their placements, the feedback from the supervisor about them, the feedback from their medical educator about them. Any tasks that they have done… from the administration team with their compliance with paperwork…if we have any concerns about their interactions with any staff..."

Grading flags

There is great diversity in the systems used by training organisations to classify flags. Many do not grade flags, although in these RTOs there is an ad-hoc system whereby different flagged issues involve different interventions. One RTO differentiated 'monitoring' and 'action' flagging grades. A 'monitoring' flag requires additional WBAs to contribute to a programmatic view of whether the registrar needs 'action' or not.

RTO: "The action list is a group of registrars that are actively discussed at each … committee every month where there may be a serious issue in their training and the monitoring list is one where minor issues have occurred … that also looks at practices and supervisor..."

Some RTOs use formal online grading systems with green, amber and red flags. One of these RTOs had just used green and red flags, but added an amber flag to provide supervisors with an intermediate level for less serious issues. This aligned with a common theme of assessors 'failing to fail' a registrar, that is, not flagging a registrar for fear of upsetting them. Participants also reported that sometimes supervisors were more comfortable to flag a registrar verbally with an ME or Regional Head of Education rather than formally document their concerns.

Some training organisations use graded flagging systems in conjunction with an online check-list; any item checked on this list will raise an amber flag and require the supervisor to comment. In some RTOs, the decision to flag a registrar is made at a senior level or by the flagging committee itself on the basis of WBA written feedback.

RTO: "...we make a decision based on the feedback whether they're really a green flag - in other words we think they are good to go, a yellow flag means there are a few things that have come up but we are not really sure and at that time we normally discuss interventions, and the interventions might be as simple as just asking the ME who is doing their visits to monitor them a bit more closely… and the red flags are where we get some more formal intervention…and at that point we bring in a focussed learning plan which is on a template..."

Some MEs reported keeping their own list of '*soft flags*' that is not part of the RTO system, as they don't quite meet the criteria. Conversely, all training organisations described '*nuclear bomb'* red flags that are escalated to the Head of Education or Director of Training immediately. Some participants also indicated that flagging frequency and severity can vary between practices, MEs and external visitors.

The relationship between flagging and exam performance

Some RTOs were of the opinion that flagged registrars should be more likely to pass their exams because of the increased input they received in the form of WBAs and feedback. Others said that flagging does not help with exams, as supervisors were reluctant to flag registrars in difficulty. Still others suggested that flagging is not usually about knowledge and so, while flagging should be unrelated to AKT and KFP performance, hopefully flagged registrars would demonstrate higher OSCE performance.

RTO: "...in reality you should actually expect a lower failure rate for those that are flagged if you believe there has been effective remediation. So my concern is that it's really looking at different things, I think the OSCE is more related to the ECTV [External Clinical Teaching Visit] because that's actually looking at the process of running consults..."

Registrars were also of the opinion that flagging would not actually help people pass the exams, as they believed that passing exams focuses on knowledge and exam technique rather than the usual issues that are flagged such as professionalism, communication skills, time management or ethical issues. All RTOs have a flagging system for registrars who fail their exams. For some this is the majority of their flagging workload[9].

The need for training

Supervisors talked about a lack of training in what should be flagged and some thought they would like a checklist that would guide or prompt them. RTOs and MEs expressed concern that External Clinical Teaching Visitors, supervisors and registrars were not fully aware of what happens when someone is flagged and of the consequences of failure to flag. Some supervisors felt insecure documenting flags and would prefer to talk to someone with '*more experience in medical education'*, such as an ME, before they would flag a registrar.

Supervisor: "We are hoping to discover where registrars have weaknesses and then help them grow. If you are not looking for it, then how can you figure out where they need to be helped and where they need to grow?"

Training of all stakeholders should be undertaken about the importance of flagging and the various pathways that can be taken in order to flag a registrar. The RTO relies on the people on the ground to tell them about the problems.

RTO: "Our biggest challenge is getting to the unknown, so the supervisors who are really good at flagging and telling us they have problems, they are almost always going to fix the problems, it's the supervisor we don't hear from are the biggest challenges and the registrars that fly under the radar..."

[9] Registrars who were flagged for failing an exam were excluded from the meta-analysis arm of this research, as the aim was to ascertain whether flagged registrars are more likely to pass exams and such flags would seriously confound any results.

Practice-level intervention before training organisation involvement
MEs and supervisors discussed some of their reasoning behind the reluctance to flag registrars to the RTO. Some tried to address the issue themselves but had their own criteria about when they would escalate.

ME: "...if I'm seeing a problem that I don't think is life-threatening for the patient, I'm much more likely to start talking about … changing the behaviour or flagging the behaviour and seeing what the response of the registrar to suggestions of behaviour change are – if it's a positive response, well, obviously I'm going to be … a lot more optimistic and I'm probably more likely to start discussing it then, but if I'm getting a stone-walling and a negativity about it, I'm probably more likely to go, okay, I'll write about these things in the report and then talk to the ME in charge of the ECTVs about how we might then plan to change this behaviour."

As a way to work with this, some RTOs initially contact a flagged registrars' supervisor and ask them if they think they can remediate the issue themselves in their usual teaching time. This seemed to be particularly the case for registrars who had difficulty settling into general practice from hospital medicine. They would always offer support and review their progress with the supervisor regularly. While supervisors may find this extra load difficult, some RTOs pay supervisors and MEs for extra WBAs and teaching time.

Use of additional WBAs for flagged registrars
Although RTOs and MEs agreed that flagged registrars need additional WBAs, the specific types of WBAs required depended on the problem. For instance, patient complaints about time management may prompt an MSF of patients. However, extra Direct Observation by an experienced ME is usually one of the additional WBAs. Furthermore, some RTOs used video review as a way of observing multiple issues in a flagged registrar with the advantage that multiple assessors could view it:

ME: "Especially if there's serious concerns and patient safety is a major concern - then I think the video reviews are a great way to go. Because you can take the recording; you can have multiple medical educators review it and make a diagnosis and recommendations. The other advantage is you can sit down with the registrar and as part of your feedback, review the videotaped session together."

Focussed Learning Intervention Plans (FLIPs)
FLIPs are designed and used by every training organisation, but each one uses them differently. There are, however, some core principles, mostly around a contract with the registrar about the additional WBAs and self-reflection on what they will do differently. Additional elements in contracts may include the expected outcomes; timelines; who will perform the WBAs; how feedback will be given; and how the FLIP will be resolved and the registrar no longer flagged.

RTO: "So in terms of flagging this is again organic intuitive get to know the person, it's a relational thing for me and my flags are very much influenced by that… there's specific activities that they are expected to do and that's a contract. So even though that's not formal remediation, these learning plans are signed and agreed to by the registrar so it's also I guess the first step into letting them know there is a problem and formally agreeing that these are the issues addressed and these are the outcomes we expect them to do."

Even though RTOs have templates, all FLIPs are individualised to the registrar's particular issues. Templates will guide an ME about what the options are, for example a communication course, additional one-hour tutorials, or an extra DOV every month. This also assists the RTO in keeping

track of the FLIP and monitoring the outcomes as to whether the ME needs more assistance from the RTO or the registrar intervention needs to be escalated further.

Who flags and how they flag

For most training organisations, there is both a formal flagging process and an informal flagging process, although there is great variation in this, ranging from lists or tick-boxes or coloured flags on the computer, to the ME reading all the reports and flagging a registrar based on an overall view, to the supervisor phoning the ME or the TC – 'whoever they have the best relationship with'. One RTO put it very succinctly about whom they would like to hear about:

RTO: "...any registrar that comes across pretty well anyone's radar where there is a moderate to significant concern or even a minor concern actually..."

Some RTOs do not have a formal flagging process as they think the bureaucracy might be a barrier and registrars are flagged in many different ways.

RTO: "...it's more a discussion or an email, or different types of discussions but there isn't formally a flagging form or an online form they put that through, I suppose a bit like nasty results we want to be called up, we want contact we want it to be highlighted in big flashing lights to us that this registrar has a flag to us and needs some help. We don't want it to be lost in a load of paperwork that might come our way."

What is flagged

Flags can be raised for an extensive array of issues. Usually, a diagnosis of the problem will be the first step in this procedure. Some RTOs flag particular high-risk groups, for example registrars in rural areas away from their family. Supervisors and MEs thought it was important that a new practice should know that a registrar has been flagged so they can ask for more support if necessary. It is important that the registrars understand that flagging is not failing, but that flagging is about everyone working together to try to get the best outcome for the registrars.

RTO: "...knowledge issues, performance issues but there is a whole lot of messiness with practice issues to try and work out whether it's a practice problem or a registrar problem. In my experience it is never cut and dried..."

Decision to ask the registrar to leave training

The final decision to ask a registrar to leave training because remediation has not been successful is a difficult one. Such a high-stakes decision is made by a group of senior MEs and senior staff based on an overview of the registrar's training progress and remediation. The final decision is usually that of the Director of Training. According to participants, that this is a possible end-point should be communicated to supervisors as well as registrars. Supervisors value their profession and need to have confidence that the RTO is upholding standards.

Supervisor: "You get flagged essentially…So it's more intensive, but there is a finite point. There comes a point when it - seeing that the registrar is struggling or not absorbing or unwilling to engage with the change, then they're advised to leave."

The decision is even more difficult if the registrar has passed their exams but the training organisation still thinks they are not an 'effective' practitioner. Documentation and justification for a decision to leave training must be clear and transparent. One RTO discussed how difficult this is, as WBA assessments tend not to carry enough weight and that it is usually compliance with training requirements or remediation that becomes the final arbiter.

RTO: "If they're not progressing, then at what point do they get pulled out of training for compliance reasons. We haven't had to pull anybody out for competency for a long time. It's mainly compliance…because to prove that they're not competent, can be extremely difficult. Because they can still be safe, but they may not be an effective practitioner."

Quantitative Analyses

Included Data

In total, data from seven GP training organisations was included, covering six Australian States and Territories. The total sample size was 643 GP registrars who sat at least one RACGP exam in 2018. On average, approximately half of all registrars who sat an exam in 2018 were flagged at some point in their training (M = 49.79%, range = 32.05% to 64.15%), with the proportion of registrars who had ever received a flag increasing over the course of training (refer Figure 1). The average incidence of flagging increased from GPT1 (19.07%) to GPT2 (25.72%), but then levelled off at GPT3 (25.91%; refer Figure 2). On average, the highest proportion of first-time flags were issued to registrars in GPT1 (M = 19.07%), although the proportion of registrars being issued first-time flags in GPT2 and GPT3 were similar (M = 15.09% and 14.06%, respectively; refer Figure 3). Table 1 includes further details.

Table 1: Flagging rates across included RTOs

		Flagging incidence			*New Flags*	
RTO	Overall prevalence of flagging	GPT1	GPT2	GPT3	GPT2	GPT3
I	64.15%	39.62%	34.91%	24.53%	12.26%	12.26%
II	52.69%	11.83%	18.28%	22.58%	18.28%	22.58%
III	40.00%	14.29%	28.57%	8.57%	22.86%	2.86%
IV	56.41%	12.82%	43.59%	38.46%	30.77%	12.82%
V	32.05%	2.56%	8.97%	27.56%	3.85%	14.74%
VI	60.00%	22.22%	24.44%	37.78%	11.11%	26.67%
VII	43.20%	30.18%	21.30%	21.89%	6.51%	6.51%
Average	49.79%	19.07%	25.72%	25.91%	15.09%	14.06%

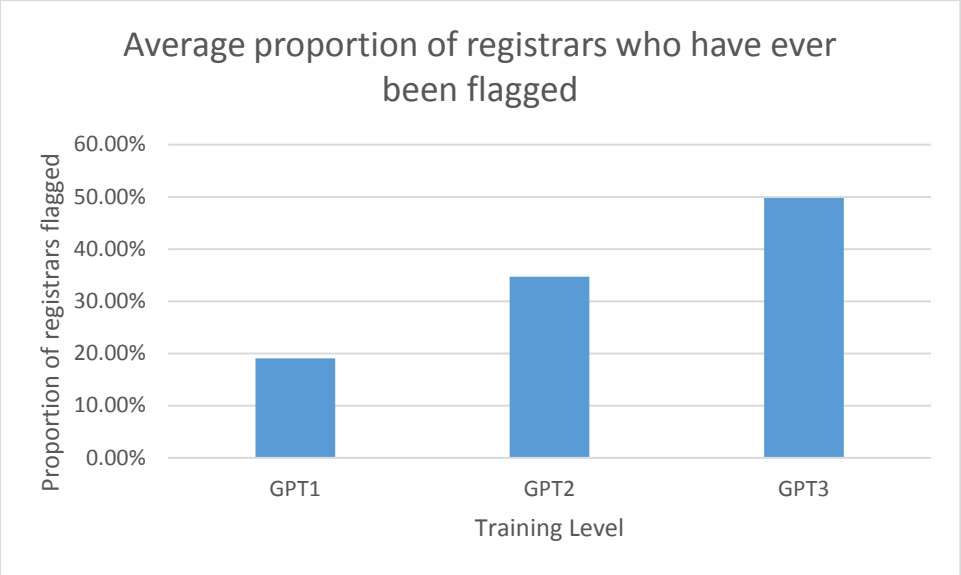


Figure 1: Average proportion of registrars who have ever received a flag across training levels

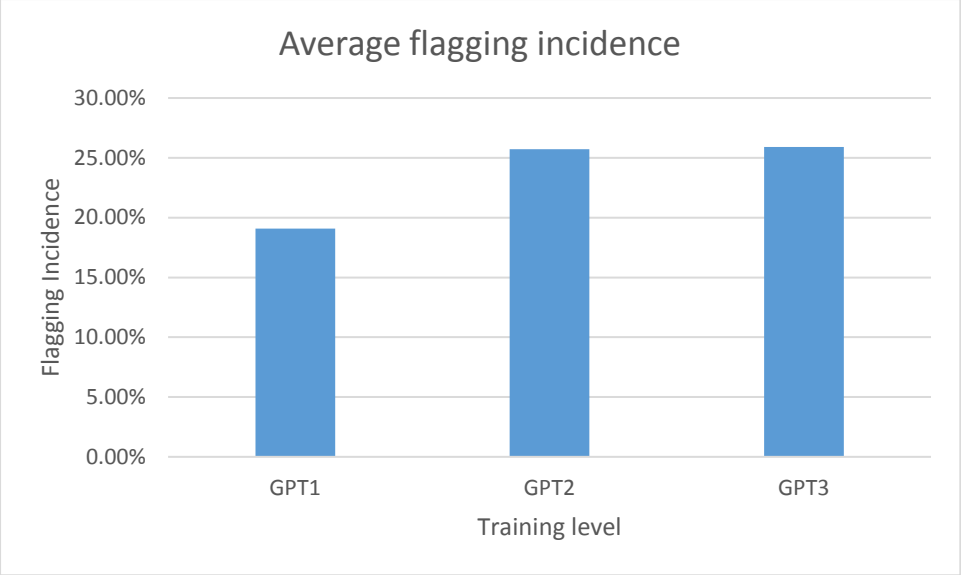


Figure 2: Average flagging incidence by training level

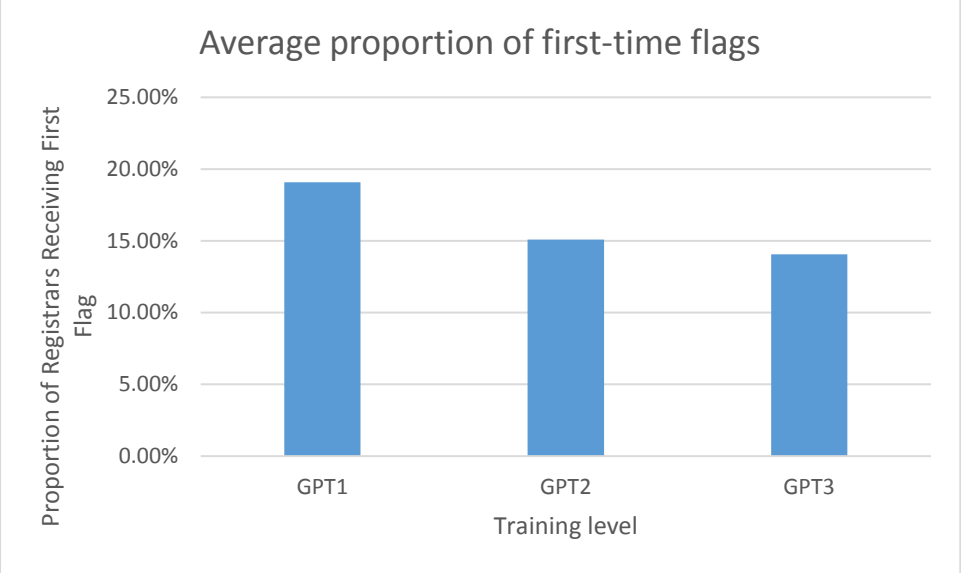


Figure 3: Average proportion of registrars receiving a new flag by training level

Meta-Analyses

Flagging at any stage of training

Registrars who were flagged at any stage of training were not at a significantly greater risk of failing any, or each, exam. However, the relationship between a flag at any stage and OSCE performance was trending to significance (RR_w = 1.748, p = .084). The heterogeneity analyses showed that for all of these analyses (except for the relationship with the OSCE) there was significant, large heterogeneity (i.e. high variance in the effect size between RTOs). Conversely, registrars who were flagged in more than one training term (i.e. flagged in GPT1 and GPT2, GPT1 and GPT3, or GPT2 and GPT3) were two-to-three times more likely to fail any, and each, exam than registrars not flagged in multiple terms (RR_w range = 1.895 to 3.161, $p \leq .002$). Again, heterogeneity data for these analyses suggested the presence of moderate-to-large variance for all but the relationship with the OSCE. Sensitivity analyses of the above relationships yielded no meaningful changes in significance or magnitude.

Flagging in each GPT level

Registrars flagged in GPT1 were found to be over 2½ times more likely to fail their OSCE than non-flagged registrars ($RR_w = 2.646$, $p = .003$). This relationship exhibited low, non-significant heterogeneity, suggesting consistency across RTOs. All other relationships with GPT1 flagging were non-significant, but most showed moderate-to-large significant heterogeneity.

There were no significant relationships between flagging in GPT2 and exam performance, however most of these analyses also showed large and significant heterogeneity. Yet, registrars who were flagged for the first time in GPT2 were over two times more likely to fail any exam than other registrars ($RR_w = 2.188$, $p < .001$). Furthermore, this relationship exhibited minimal heterogeneity, suggesting consistency across RTOs.

Perhaps the starkest effects were evident for registrars flagged in GPT3, who were significantly more likely to fail any, and each, exam. Registrars flagged in GPT3 were over two times more likely to fail any exam ($RR_w = 2.283$, $p = .006$); their AKT ($RR_w = 2.549$, $p = .006$); and KFP ($RR_w = 2.282$, $p = .015$) than registrars not flagged in GPT3. Furthermore, these registrars were four times more likely to fail their OSCE than registrars not flagged in GPT3 ($RR_w = 4.060$, $p < .001$). Additionally, this last relationship showed minimal, non-significant heterogeneity, suggesting this effect is consistent across included RTOs. Conversely, the remaining relationships exhibited significant, moderate-to-large heterogeneity. Similar effects were identified for registrars flagged for the first time in GPT3. These registrars demonstrated an increased propensity to fail any exam ($RR_w = 1.805$, $p = .028$); their AKT ($RR_w = 1.953$, $p = .011$); and their OSCE ($RR_w = 2.847$, $p = .001$) relative to other registrars. Yet, there was no significant relationship with respect to registrars first flagged in GPT3 and KFP performance. Furthermore, the relationships with KFP performance and failing any exam demonstrated significant, high heterogeneity.

Sensitivity analyses of the above relationships yielded no meaningful changes in significance or magnitude. Further details regarding the main meta-analyses are reported in Table 2.

Table 2: Main meta-analyses

Predictor	Outcome	N_{RTOs}[a]	95% CI			Heterogeneity				N_{FS}[b]
			RR	Lower	Upper	*p*	*Q*	*p*	Tau	I^2	
Overall flagged	AKT performance	7	1.783	0.626	5.081	.279	35.611	<.001	1.211	83.151	46
	KFP performance	7	1.544	0.618	3.855	.352	67.188	<.001	1.117	91.070	17
	OSCE performance	7	1.748	0.928	3.291	.084	3.635	.726	0	0	74
	Overall exam performance	7	1.661	0.750	3.676	.211	80.147	<.001	0.987	92.514	18
GPT1 flagged	AKT performance	7	1.655	0.926	2.955	.089	12.228	.057	0.538	50.931	69
	KFP performance	7	1.205	0.738	1.966	.456	14.829	.022	0.496	59.539	24
	OSCE performance	7	***2.646***	***1.388***	***5.046***	***.003***	7.372	.288	0.375	18.609	132
	Overall exam performance	7	1.359	0.916	2.015	.127	22.200	.001	0.44	72.973	52
GPT2 flagged	AKT performance	7	1.517	0.731	3.149	.263	17.445	.008	0.754	65.607	77
	KFP performance	7	1.178	0.687	2.018	.552	19.943	.003	0.591	69.915	21
	OSCE performance	7	1.457	0.743	2.859	.273	4.379	.626	0	0	48
	Overall exam performance	7	1.205	0.744	1.954	.448	25.888	<.001	0.556	76.823	39
GPT3 flagged	***AKT performance***	7	***2.549***	***1.316***	***4.940***	***.006***	17.491	.008	0.684	65.697	126
	KFP performance	7	***2.282***	***1.176***	***4.430***	***.015***	31.959	<.001	0.755	81.226	117
	OSCE performance	7	***4.060***	***2.187***	***7.537***	***<.001***	4.316	.634	0	0	195
	Overall exam performance	7	***2.283***	***1.262***	***4.130***	***.006***	44.580	<.001	0.708	86.541	106
Flagged for the first time in GPT2	AKT performance	7	1.095	0.460	2.606	.838	13.985	.030	0.842	57.096	62
	KFP performance	7	0.926	0.464	1.850	.828	18.884	.004	0.740	68.226	34
	OSCE performance	7	0.760	0.308	1.877	.552	1.720	.944	0	0	31
	Overall exam performance	7	***2.188***	***1.609***	***2.976***	***<.001***	1.422	.964	0	0	106
Flagged for the first time in GPT3	***AKT performance***	7	***1.953***	***1.163***	***3.281***	***.011***	9.779	.134	0.420	38.645	90
	KFP performance	7	1.730	0.852	3.514	.130	26.139	<.001	0.779	77.046	101
	OSCE performance	6	***2.847***	***1.517***	***5.342***	***.001***	2.134	.830	0	0	123
	Overall exam performance	7	***1.805***	***1.067***	***3.054***	***.028***	31.273	<.001	0.601	80.814	81

Predictor	Outcome	N_{RTOs}[a]	RR	95% CI		p	Heterogeneity				N_{FS}[b]
				Lower	Upper		Q	p	Tau	I^2	
Flagged in more than 1 training term	***AKT performance***	6	***2.583***	***1.444***	***4.620***	***.001***	9.871	.079	.486	49.344	116
	KFP performance	6	***1.895***	***1.195***	***3.005***	***.002***	13.001	.023	.434	61.542	81
	OSCE performance	6	***3.161***	***1.633***	***6.116***	***.001***	3.405	.638	0	0	136
	Overall exam performance	6	***1.947***	***1.309***	***2.895***	***.001***	18.006	.003	.404	72.232	93

*Note: **Bolded** figures are statistically significant relationships (*p < *.05). [a]Note that some analyses required the exclusion of some RTOs, as there were instances of insufficient data to calculate Relative Risks; [b] Number of RTOs with a mean RR of 1 required to reduce results to $RR_w = 1.050$*

Subgroup Analyses

Using the information gleaned from the qualitative component of this research stream, the following features of RTOs' flagging procedures were identified:

- Whether flagging protocols are clearly articulated to stakeholders
- Whether different types of flagging are distinguished (e.g. professional vs health)
- Whether flags are formally graded (e.g. green, amber, red)
- Whether the assigned ME is involved in remediation
- Whether flags are recorded in an online system
- Whether benchmarking is used as a means for flagging
- Whether prospective registrars can be flagged during selection
- Whether there are dedicated Training Coordinators (TCs)
- Whether there is a formal diagnostic procedure following flagging
- The Medical Educator to Registrar ratio
- The TC to Registrar ratio

A cautionary note about subgroup analyses

Subgroup analyses involve dividing groups (in this case GP training organisations) on the basis of certain group-wide characteristics (e.g. whether flags are graded) and running separate meta-analyses on these groups. These effects can then be compared to see not only if there is a significant effect *within* each group, but also *between* the groups. If there is a significant between-groups difference, this suggests this variable may moderate the relationship being examined. Yet, in dividing groups on the basis of this factor, it is also possible that the groups differ on other factors as well. Given the inherently high risk of subgroup analyses being confounded, results must be interpreted very cautiously. Any significant findings arising from subgroup analyses highlight factors which *may be* associated with differences between groups rather than identifying the cause of such differences. Significant subgroup analyses highlight areas for future primary research; they are not a basis for recommendations for major changes.

Clear flagging protocols

In the overall analysis (i.e. the relationship between being flagged at any stage of training and failing any exam), in RTOs where there are clear flagging protocols (i.e. where flagging protocols are communicated to stakeholders), flagged registrars were nearly three times more likely to fail any exam than non-flagged registrars ($RR_w = 2.891$, $p < .001$). This effect was consistent across almost all other relationships (RR_w range = 1.869 to 4.113), with the exception of OSCE performance, although this too was approaching significance ($p = .055$).

In RTOs that do not have clearly articulated flagging protocols, the only significant relationship identified was that registrars flagged for the first time in GPT2 were over two times more likely to fail any exam than other registrars ($RR_w = 2.132$, $p < .001$). This effect mirrored that for the corresponding relationship in RTOs with clear flagging protocols. The only significant between-RTOs difference was in the relationship between GPT2 flagging and failing any exam ($Q(1) = 5.696$, $p = .017$). The heterogeneity for RTOs with clear flagging protocols was consistently non-significant. Conversely, RTOs without clear flagging protocols had significant heterogeneity.

Flags distinguished by type

In RTOs where different types of flags are distinguished (i.e. distinguishing between issues such as professionalism, health and personal factors in flagging), the only significant relationships

identified were that registrars flagged for the first time in GPT2 or who received flags in multiple training terms were approximately two times more likely to fail any exam (RR_w = 2.108 and 2.489, respectively). Similarly, in RTOs where flags are not distinguished, registrars flagged in GPT3 or for the first time in GPT2 were over 2½ times more likely to fail any exam (RR_w = 2.698 and 2.449, respectively). There were no significant between-groups differences, and almost all analyses of heterogeneity for both groups showed moderate-to-large, significant heterogeneity.

Formally graded flags

In RTOs where flags are formally graded, flagged registrars were over 2½ times more likely to fail any exam (RR_w = 2.606, p = .024), whereas in RTOs where flags are *not* formally graded, flagged registrars were 40% *less* likely to fail any exam (RR = 0.616, p = .006). This between-groups difference was significant ($Q(1)$ = 9.811, p = .002).

The majority of other relationships examined produced results consistent with the above overall analysis. In RTOs where flags are graded, flagged registrars were generally significantly more likely to fail their exams (RR_w range = 1.672 to 3.595). The only two relationships where these relationships were not significant (i.e. flagging at any level, and AKT and KFP performance) were trending to significance and produced the same effect.

Notably, in RTOs where flags are not formally graded, registrars flagged in GPT1 were *less likely* to fail any exam (RR_w = 0.834, p = .038). Furthermore, registrars in these RTOs who were flagged at any level were almost two times more likely to pass their KFP as non-flagged registrars (RR_w = 0.520, p < .001). Conversely, registrars flagged for the first time in GPT2 in RTOs where flags are not formally graded were over 2½ times more likely to fail any exam (RR_w = 2.742, p = .001).

When comparing these RTOs, almost all relationships demonstrated significant between-groups differences. Interestingly, in some RTOs where flags are formally graded, there was moderate-to-high, significant heterogeneity in the relationships, suggesting variation in the way in which flags are graded may be associated with different effects. Conversely, heterogeneity in the relationships in RTOs where flags are not graded was generally minimal.

Involvement of assigned MEs in remediation

In training organisations where assigned MEs are not involved in remediation, flagged registrars were over 3½ times more likely to fail any exam than non-flagged registrars (RR_w = 3.587, p < .001). This relationship was also characterised by minimal heterogeneity. Conversely, although the same relationship in RTOs where assigned MEs *are* involved in remediation was non-significant, the confidence interval showed the effect varied from flagged registrars being up to 20% more likely to fail any exam to 30% *less* likely to fail any exam. This relationship exhibited a highly significant between-groups difference ($Q(1)$ = 25.663, p < .001).

Further analyses in RTOs without assigned MEs involved in remediation showed that registrars flagged at any, and each, stage of training or in multiple training terms were significantly more likely to fail any, and each, exam (RR_w range = 2.034 to 5.122). Conversely, in RTOs that do involve assigned MEs in remediation, flagged registrars were more likely to pass the KFP (RR_w = 0.559, p = .041). Furthermore, this relationship showed a highly significant between-groups difference ($Q(1)$ = 30.757, p < .001). Most of the other relationships examining RTOs that involve assigned MEs in remediation yielded non-significant results. However, consistent with the overall analysis, the confidence intervals suggested effects ranging from flagged registrars being moderately more likely to moderately less likely to fail any/each exam. Notably, the upper limits of the confidence

interval were still below the effects for RTOs that do not involve assigned MEs in remediation. The only exception to this was registrars flagged for the first time in GPT2, who were 2½ times more likely to fail any exam than other registrars (RR_w = 2.510, p = .001). Most relationships in these RTOs were characterised by moderate, sometimes significant, heterogeneity. Yet, the majority of relationships in these subgroup analyses exhibited significant between-groups differences.

Recording flags in an online system

Registrars flagged in GPT3 or for the first time in GPT2 in RTOs that do not use an online system to record flags were two times more likely to fail any exam as non-flagged registrars (RR_w = 2.002 and 2.484, respectively). However, comparable effects were identified for registrars in RTOs that do use an online system to record flags for registrars first flagged in GPT2 and who received flags in multiple training terms (RR_w = 2.115 and 2.183, respectively). Furthermore, none of these relationships exhibited significant between-groups differences. No other relationships were significant within this set of subgroup analyses, with high, significant heterogeneity identified in most analyses.

Benchmarking as a means for flagging

Several analyses identified that registrars flagged in GPT3, or flagged for the first time in GPT2 or GPT3 in RTOs that do not use benchmarking were approximately 2½ times more likely to fail any exam (RR_w range = 2.158 – 2.795). Comparatively, in RTOs that *do* use benchmarking, only registrars flagged for the first time in GPT2 were significantly more likely to fail any exam (RR_w = 2.292, p = .011). All relationships exhibited highly non-significant between-groups differences, while almost all relationships exhibited significant, high heterogeneity in each group.

Flagging at selection

In RTOs where flagging does not occur at selection, registrars flagged in GPT3; for the first time in GPT2; or in multiple training terms were approximately two times more likely to fail any exam (RR_w range = 1.921 to 2.207). Comparatively, in RTOs where flagging can occur at selection, the only significant relationship was that registrars first flagged in GPT2 were approximately 2½ times more likely to fail any exam (RR_w = 2.637, p = .001). However, all between-groups differences were highly non-significant and most analyses showed very high, highly significant heterogeneity.

Dedicated training coordinators (TCs)

In RTOs where there are dedicated TCs, registrars flagged at GPT3 were over 2½ times more likely to fail their exams than non-flagged registrars (RR_w = 2.616, p = .006). Registrars flagged in these RTOs were also over 1½ times more likely to fail their KFP than non-flagged registrars (RR_w = 1.629, p = .002). Furthermore, registrars flagged in GPT2 or GPT3 for the first time or in multiple training terms in these RTOs were approximately two times more likely to fail their exams (RR_w range = 1.891 to 2.066). Comparatively, in RTOs without dedicated TCs, the only significant relationship identified was that registrars flagged for the first time in GPT2 were nearly 2½ times more likely to fail any exam (RR_w = 2.476, p = .001). All other analyses were non-significant; no between-groups analyses were significant; and almost all relationships exhibited significant, high heterogeneity in each group.

Existence of a formal diagnostic process

In RTOs without a formal diagnostic process (FDP) following flagging, flagged registrars were over 2½ times more likely to fail any exam than non-flagged registrars (RR_w = 2.741, p = .029). Conversely, in RTOs that *do* have FDPs, flagged registrars were nearly two times more likely to *pass*

their exams than non-flagged registrars ($RR_w = 0.526$, $p < .001$). Furthermore, this between-RTOs difference was highly significant ($Q(1) = 11.347$, $p = .001$). Whilst there was minimal heterogeneity in this relationship in RTOs with a FDP, there was large, highly significant heterogeneity in RTOs without FDPs.

Similar patterns were observed in the analyses for registrars flagged in GPT2 and failing any exam, and flagging at any stage and AKT performance. Registrars flagged in GPT2 in RTOs without FDPs were over 1½ times more likely to fail any exam than non-flagged registrars ($RR_w = 1.630$, $p = .021$). Conversely, registrars flagged in GPT2 in RTOs with FDPs were over two times more likely to pass their exams compared to non-flagged registrars ($RR_w = 0.481$, $p = .032$), again with a significant between-groups difference ($Q(1) = 9.282$, $p = .002$). These effects were even starker for AKT performance. Registrars flagged at any stage in RTOs without FDPs were nearly 3½ times more likely to fail their AKT ($RR_w = 3.446$, $p = .011$), while flagged registrars in RTOs with FDPs were over two times more likely to *pass* their AKT than non-flagged registrars ($RR_w = 0.413$, $p = .007$), again with a significant between-groups difference ($Q(1) = 12.977$, $p < .001$).

While the remaining analyses did not show significant results in both groups *and* a significant between-groups difference, general trends suggest that in RTOs without FDPs for flagged registrars, flagged registrars were significantly more likely to fail an exam. Conversely, in RTOs with FDPs, flagged registrars were significantly more likely to pass their exams. The only exception to this trend was that registrars first flagged in GPT2 in both types of RTOs were two times more likely to fail any exam than other registrars.

Details regarding subgroup analyses are reported in Table 3 and Appendix A.

Table 3: Subgroup analyses

Subgroup	Predictor	Outcome	RR Feature Absent	RR Feature Present	Between-groups difference (*p*)
Clear flagging protocols	Flagged (any level)	Failed exam	1.280	***2.891***	.146
	GPT1 flag	Failed exam	1.158	***1.909***	.136
	GPT2 flag	Failed exam	0.913	***2.252***	***.017***
	GPT3 flag	Failed exam	1.895	***3.544***	.168
	First flag in GPT2	Failed exam	***2.132***	***2.426***	.721
	First flag in GPT3	Failed exam	1.741	***1.869***	.874
	Flagged in 2+ training terms	Failed exam	1.562	***2.823***	.102
	Flagged (any level)	Failed AKT	1.234	***4.113***	.179
	Flagged (any level)	Failed KFP	1.120	***3.230***	.111
	Flagged (any level)	Failed OSCE	1.334	2.772	.274
Distinguish between different types of flagging	Flagged (any level)	Failed exam	2.625	1.320	.494
	GPT1 flag	Failed exam	1.184	1.490	.607
	GPT2 flag	Failed exam	1.322	1.079	.694
	GPT3 flag	Failed exam	***2.698***	1.894	.576
	First flag in GPT2	Failed exam	***2.449***	***2.108***	.681
	First flag in GPT3	Failed exam	2.099	1.548	.604
	Flagged in 2+ training terms	Failed exam	1.514	***2.489***	.199
	Flagged (any level)	Failed AKT	3.337	1.199	.394
	Flagged (any level)	Failed KFP	2.676	1.247	.529
	Flagged (any level)	Failed OSCE	2.159	1.576	.647
Formally graded flags	Flagged (any level)	Failed exam	***0.616***	***2.606***	***.002***
	GPT1 flag	Failed exam	***0.834***	***1.831***	***.001***
	GPT2 flag	Failed exam	0.595	***1.672***	.051
	GPT3 flag	Failed exam	1.115	***3.595***	***.001***
	First flag in GPT2	Failed exam	***2.742***	***2.033***	.412
	First flag in GPT3	Failed exam	0.953	***2.547***	***.001***
	Flagged in 2+ training terms	Failed exam		(Insufficient data)[a]	

Subgroup	Predictor	Outcome	RR Feature Absent	RR Feature Present	Between-groups difference (*p*)
Formally graded flags (cont'd)	Flagged (any level)	Failed AKT	0.672	3.052	***.039***
	Flagged (any level)	Failed KFP	***0.520***	2.614	***.004***
	Flagged (any level)	Failed OSCE	0.913	***2.523***	.131
Assigned ME involved in remediation	Flagged (any level)	Failed exam	***3.587***	0.715	***<.001***
	GPT1 flag	Failed exam	***2.165***	0.876	***<.001***
	GPT2 flag	Failed exam	***2.154***	0.727	***<.001***
	GPT3 flag	Failed exam	***3.832***	1.464	***.017***
	First flag in GPT2	Failed exam	***2.034***	***2.510***	.524
	First flag in GPT3	Failed exam	***2.393***	1.446	.297
	Flagged in 2+ training terms	Failed exam	***2.831***	1.185	***<.001***
	Flagged (any level)	Failed AKT	***5.122***	0.687	***<.001***
	Flagged (any level)	Failed KFP	***3.828***	***0.559***	***<.001***
	Flagged (any level)	Failed OSCE	***2.831***	1.070	.132
Record flags in online system	Flagged (any level)	Failed exam	2.696	1.541	.678
	GPT1 flag	Failed exam	1.294	1.397	.885
	GPT2 flag	Failed exam	1.426	1.099	.659
	GPT3 flag	Failed exam	***2.002***	2.361	.768
	First flag in GPT2	Failed exam	***2.484***	***2.115***	.675
	First flag in GPT3	Failed exam	1.417	1.999	.435
	Flagged in 2+ training terms	Failed exam	1.722	***2.183***	.633
	Flagged (any level)	Failed AKT	2.696	1.541	.678
	Flagged (any level)	Failed KFP	1.478	1.644	.931
	Flagged (any level)	Failed OSCE	2.247	1.644	.733
Benchmarking as means of flagging	Flagged (any level)	Failed exam	1.810	1.426	.837
	GPT1 flag	Failed exam	1.337	1.373	.959
	GPT2 flag	Failed exam	1.317	0.919	.722
	GPT3 flag	Failed exam	***2.795***	1.525	.399
	First flag in GPT2	Failed exam	***2.158***	***2.292***	.871
	First flag in GPT3	Failed exam	***2.281***	1.155	.145

Subgroup	Predictor	Outcome	RR Feature Absent	RR Feature Present	Between-groups difference (*p*)
Benchmarking as means of flagging (cont'd)	Flagged in 2+ training terms	Failed exam		(Insufficient data)[a]	
	Flagged (any level)	Failed AKT	1.912	1.872	.990
	Flagged (any level)	Failed KFP	1.679	1.387	.881
	Flagged (any level)	Failed OSCE	1.851	1.893	.983
Flagging at selection	Flagged (any level)	Failed exam	1.607	1.851	.879
	GPT1 flag	Failed exam	1.510	1.154	.473
	GPT2 flag	Failed exam	1.356	0.968	.598
	GPT3 flag	Failed exam	***2.207***	2.620	.817
	First flag in GPT2	Failed exam	***2.043***	***2.637***	.472
	First flag in GPT3	Failed exam	1.755	2.046	.810
	Flagged in 2+ training terms	Failed exam	***1.921***	1.873	.965
	Flagged (any level)	Failed AKT	1.940	1.574	.853
	Flagged (any level)	Failed KFP	1.398	1.964	.754
	Flagged (any level)	Failed OSCE	2.117	1.497	.593
Dedicated TC	Flagged (any level)	Failed exam	1.265	2.298	.486
	GPT1 flag	Failed exam	1.164	1.619	.390
	GPT2 flag	Failed exam	0.891	1.615	.264
	GPT3 flag	Failed exam	1.979	***2.616***	.679
	First flag in GPT2	Failed exam	***2.476***	***2.056***	.577
	First flag in GPT3	Failed exam	1.713	***1.891***	.868
	Flagged in 2+ training terms	Failed exam	1.785	***2.066***	.740
	Flagged (any level)	Failed AKT	0.968	3.511	.208
	Flagged (any level)	Failed KFP	0.751	***1.692***	.555
	Flagged (any level)	Failed OSCE	1.440	2.371	.451
Formal Diagnostic Process	Flagged (any level)	Failed exam	***2.741***	***0.526***	***.001***
	GPT1 flag	Failed exam	1.522	0.954	.165
	GPT2 flag	Failed exam	***1.630***	***0.481***	***.002***
	GPT3 flag	Failed exam	***3.188***	0.802	***<.001***
	First flag in GPT2	Failed exam	***2.159***	***2.311***	.862

Subgroup	Predictor	Outcome	RR Feature Absent	RR Feature Present	Between-groups difference (*p*)
Formal Diagnostic Process (cont'd)	First flag in GPT3	Failed exam	***2.278***	0.804	***.001***
	Flagged in 2+ training terms	Failed exam		(Insufficient data)[a]	
	Flagged (any level)	Failed AKT	***3.446***	***0.413***	***<.001***
	Flagged (any level)	Failed KFP	2.770	***0.463***	***.003***
	Flagged (any level)	Failed OSCE	***2.272***	0.847	.177

Note: ***Bolded*** figures are statistically significant findings ($p < .05$). [a]For some features, there was insufficient data to perform subgroup analyses.

Meta-Regressions

Due to the small sample size ($N_{RTOs} = 7$), it was not possible to perform a meta-regression using all the above moderator factors simultaneously. Instead, separate regression models were run with each categorical moderator to identify which produced the greatest improvements on the intercept-only model. Regressions with these variables were only run on the relationship between overall exam performance and flagging at any stage of training.

The model including whether the assigned ME is involved in remediation explained 89% of between-groups variance and significantly improved on the intercept-only model ($Q(1) = 20.83$, $p < .01$). A goodness of fit test to assess the heterogeneity remaining in this model was approaching non-significance ($Q(1) = 11.21$, $p = .0474$), suggesting this model accounted for the majority of heterogeneity. The grading flags + intercept model explained 63% of between-groups variance and significantly improved on the intercept-only model ($Q(5) = 6.27$, $p = .01$). However a goodness of fit test to assess heterogeneity suggested this explained only a minimal amount of heterogeneity ($Q(5) = 25.17$, $p = .0001$). Additionally, the FDP + intercept model explained 32% of between-groups variance and also significantly improved on the intercept-only model ($Q(1) = 4.81$, $p = .028$). Yet, this model also explained a minimal amount of heterogeneity ($Q(5) = 39.59$, $p < .001$). Finally, having clear flagging protocols explained some between-group variance ($r^2 = .02$), although this did not significantly improve on intercept-only model ($Q(1) = 1.11$, $p = .293$), nor did it explain much heterogeneity ($Q(5) = 67.00$, $p < .001$). No other categorical variables produced significant improvements over the intercept-only model, with minimal r^2 values ($r^2 < .001$).

Given the ME-to-registrar ratio is a continuous variable, it could not be subjected to subgroup analyses which require categorical moderators. Accordingly, a series of regression models were generated exploring several relationships. Results showed that the inclusion of the ME-to-registrar ratio as a moderator did not significantly improve any model for any relationship. Although the goodness of fit test yielded a highly non-significant for the relationship between OSCE and flagging at any point ($Q(5) = 3.33$, $p = .650$) in this model, this must be interpreted in light of the consistently low heterogeneity in the relationship between flagging at any point in time and OSCE performance.

Given only three RTOs were identified as having dedicated TCs, there was an insufficient sample size for running a regression to analyse the moderating effect of the TC-to-registrar ratio on any relationships between flagging and exam performance.

Given the above results, regression models were created combining the four moderators with $r^2 > .001$ to attempt to explain further variance. It was found that almost all models containing the moderator of whether the assigned ME was involved in remediation significantly improved on the intercept-only model (all $p \leq .003$). Conversely, most models without this variable as a moderator did not produce significant improvements on the intercept-only model. The only exception to this was the model containing formal diagnostic process and grading flags as moderators, which explained 74% of between-groups variance ($Q(2) = 13.74$, $p = .001$). Notably, most models that contained assigned ME involved in remediation and a formal diagnostic process as moderators produced non-significant results for the goodness of fit test ($p > .05$), suggesting that all heterogeneity had been explained by these models.

Further details regarding the meta-regressions are reported in Appendix B.

Discussion

Key Findings

The first key finding relates to the existence of relationships between flagging and exam performance. There were mixed views amongst participants of the qualitative component as to whether there would be a relationship between flagging and exam performance. Given the analyses conducted in the quantitative arm of this study demonstrated that relationships *do* exist, this suggests that there is some overlap between the constructs focussed on in flagging processes and the constructs assessed by exams. This means flagging can act as predictor of registrar exam performance *within training* in addition to the pre-training predictors previously been identified (Patterson et al., 2016, Stewart et al., 2018).

A major theme that arose in the qualitative component of this stream, as well as in other streams, was the issue of 'failure to fail', that is, that registrars who should be flagged are not being flagged. Participants believed that much of this is due to flagging being stigmatised amongst registrars and supervisors. To help overcome issues regarding failure to fail, participants in the qualitative element reported that RTOs have implemented interventions such as providing supervisors and practices with support to help flagged registrars. Similarly, templates are useful for FLIPs for guiding supervisors and helping RTOs and MEs monitor the registrar's progress. More fundamentally though, the findings of the meta-analysis demonstrate that registrars who are flagged at any point in training are not necessarily more likely to fail any or each exam. What seems to be more important is *when* in training they are flagged.

Consistent with this notion of the timing of flagging being critical, participants in the qualitative component stressed the importance of early flagging. The rationale for this was that early flagging permits early intervention, which in turn helps to maximise patient safety and registrar success. The importance of early flagging, from an exam perspective, was supported by the findings of the meta-analysis. Registrars flagged for the first time in GPT2 or GPT3 were generally more likely to fail their exams. Similarly, registrars who were flagged in GPT3 were more likely to fail any, and each, exam than registrars not flagged in GPT3. Collectively, this evidence emphasises the importance of early interventions to maximise registrar success.

Through the qualitative arm of this project, the diversity of flagging models across RTOs became quite evident. Notably, flagging models could be distinguished by several important features. In the form of subgroup analyses, the potential effects of these features in moderating the relationship between flagging and exam performance was explored, although the risk of these results being confounded limited their practical utility. An innovative feature in some RTOs is the existence of formal diagnostic processes. These are designed to identify the specific issues for a flagged registrar and develop a plan for remediation, with a suite of WBAs tailored to address the specific issues. The results of the subgroup analyses comparing training organisations that do and do not have formal diagnostic processes provided some support for the effectiveness of formal diagnostic processes; flagged registrars in these RTOs were, generally, more likely to pass exams than flagged registrars in RTOs without formal diagnostic processes.

Another important feature identified was whether the assigned ME handles remediation, or a dedicated remediation ME becomes involved. Some participants held negative views of the latter approach, arguing that a remediation ME lacks the relationship with the registrar that the assigned ME already has, as well as not having the global context of the registrar's training. These negative views were supported by the subgroup analyses; flagged registrars in RTOs where the assigned ME is involved were generally less likely to fail (and in some instances, more likely to pass) their exams than flagged registrars in RTOs with remediation MEs. Subsequent meta-regressions also

identified that this feature, as well as formal diagnostic processes, were the critical factors in explaining between-RTO variance in the overall relationship between flagging and exam performance.

One of the few commonalities in flagging systems across RTOs was the existence of a group or committee that focuses on flagging. However, although these committees all collate data from multiple perspectives, the frequency of meetings and matters they deal with varies greatly. An interesting finding from one RTO was that having regional committees can interfere with equitable remediation and the consistency of flagging standards. Since this was a common feature to all RTOs, this could not be statistically examined in subgroup analyses.

Another universal feature is the grading of flags, but only some RTOs formally grade flags. In RTOs with a formal graded flagging system, participants reported that having an intermediate or amber flag seemed to be another helpful strategy in overcoming the 'failure to fail'. The quantitative analyses examining differences between RTOs that do and do not formally grade flags yielded mixed findings. Although the general finding was that registrars in RTOs with formally graded flagging systems were less likely to pass their exams, there was considerable heterogeneity within these groups, suggesting the variability of the specific models of grading may be implicated in moderating the effect. A further area that may also contribute to this heterogeneity is that the relationships between flagging and exam performance may vary between registrars who received 'amber' and 'red' flags.

Similar effects were noted for whether RTOs clearly articulate flagging protocols to stakeholders. The results suggested that, generally, flagged registrars in RTOs with clear flagging protocols were more likely to fail their exams, while there were non-significant effects for flagged registrars in RTOs without clear flagging protocols. Interestingly, the results for RTOs with clear flagging protocols had consistently non-significant heterogeneity, suggesting a consistent effect. These higher risks associated with formally grading flags and clear flagging protocols are not necessarily an unwelcome finding, as will be discussed shortly.

The data from the remaining subgroup analyses did not implicate the remaining features (i.e. distinguishing between different types of flags; recording flags in online systems; benchmarking flags; flagging at selection; having dedicated training coordinators; and ME to registrar ratio) in the relationship between registrar flagging and exam performance. However, these non-significant findings do not necessarily eliminate the effectiveness of these strategies, as there is a high potential for these subgroup analyses to have been confounded. Furthermore, these analyses make no comment regarding additional benefits afforded by these features outside of the context of exam performance.

When examining the features of flagging systems identified in the qualitative arm and explored in the quantitative arm of this project, several features can be divided into two groups: features designed to improve the accuracy of identifying registrars in difficulty, and features to improve the quality of remediation programs. These are two very different classes of features and require different interpretations of the effect sizes from subgroup analyses. For the former, if a feature helps to improve the accuracy of detecting registrars in difficulty (and one accepts the premise that registrars in difficulty are more likely to fail their exams), flagged registrars should be more likely to fail their exams. That is, for these features, a favourable outcome would be a higher Relative Risk (i.e. $RR > 1$). Conversely, for the latter, if a feature is beneficial in the remediation process, flagged registrars should perform *better* on their exams and so one would expect a lower Relative Risk (i.e. $RR < 1$). Within this context, consider that the purpose of clearly articulating

flagging protocols to stakeholders and formally grading flags is to improve the accuracy of flagging systems. Accordingly, that the subgroup analyses for both of these features showed flagged registrars were more likely to fail their exams in RTOs with these features (i.e. RR > 1) suggests that these features are effective strategies for improving the accuracy of flagging systems. Conversely, consider that the purpose of involving the assigned ME in remediation and having formal diagnostic processes is to improve the quality of remediation rather than the accuracy of flagging. Accordingly, given that the subgroup analyses for both of these features suggested their presence was associated with flagged registrars being more likely to pass their exams (i.e. RR < 1), this indicates that these features are beneficial for improving the quality of remediation. Although there is no capacity to disentangle these two perspectives in the present study's findings, these findings still offer useful insights into features regarding the accuracy of flagging systems and effectiveness of remediation programs.

Implications

A major implication arises from the existence of relationships between flagging and exam performance. These relationships demonstrate that, under some circumstances, flagging is an early warning sign for exam performance within training, providing an additional tool for knowing when to activate preventive measures, along with the existing pre-commencement issues (Patterson et al., 2016, Stewart et al., 2018). Given that oftentimes flags may be raised during the course of WBAs, this also suggests that, while research shows *individual* WBAs may not be effective predictors of exam performance (Patterson et al., 2016, Stewart et al., 2018), a broader overview in the form of flagging, which includes WBAs, may predict exam performance. Thus, WBAs may be seen as one, but not the only, foundational element of an early warning system in the form of flagging with regards to exam performance.

As mentioned, the present study reinforces the messages from the other streams regarding the issue of 'failure to fail'. Beyond this though, the present study offers important information about how to manage the 'failure to fail' problem. Quantitative data now exists to demonstrate to registrars and supervisors that flagging does not destine a registrar to fail their exams. Furthermore, early flagging (in association with remedial activities) is likely to be associated with the registrar being *more likely* to pass their exams. Breaking down any possible stigma against flagging may help with, but not be the sole means for, overcoming issues with failure to fail. Aside from the stigma of flagging as a contributor to the 'failure to fail' issue, practical strategies, such as providing templates for guiding supervisors and having an intermediate level flag, have been identified that may be beneficial in effectively addressing issues raised in flags.

Several areas for future research have also been identified. Results suggest that the involvement of the registrar's assigned ME in remediation and the use of formal diagnostic processes may be useful in helping flagged registrars improve their likelihood of passing exams. Additionally, there appeared to be high heterogeneity in the relationships between flagging and exam performance within RTOs that formally grade flags. Further primary research should seek to glean a richer understanding of each of these elements and to better understand the fundamental factors that contribute to their effectiveness in improving registrar exam outcomes.

Limitations

The data used in the quantitative analyses only examined a single cohort (registrars who sat at least one RACGP exam in 2018). Additionally, not all Australian GP training organisations were included, so generalisations must be made. With respect to this last issue, though, it is important to note that the Fail-safe N statistics for all analyses were highly robust, suggesting that results

may be feasibly generalised across the Australian context. Nonetheless, it would be preferable for these analyses to be repeated with future cohorts to verify if the findings can be replicated.

Similarly, not all RTOs were involved in the qualitative component of the present study. Although saturation was reached, saturation is not a useful metric with regards to the identification of different features and strategies in flagging models. This leaves room for further research examining additional models of flagging, both within the Australian and international contexts, to identify new features or strategies to further enhance flagging.

Finally, as mentioned, the subgroup analyses conducted in the meta-analysis were quite possibly confounded by other issues. For example, a potential major confounder of any subgroup analyses are the differences between GP training organisations' exam-preparation interventions. Therefore, future primary research should seek to specifically explore the major features identified from the subgroup analyses of the present study (as mentioned above). Furthermore, this research should pay specific attention to the accuracy of flagging and the effectiveness of remediation within these relationships.

Conclusion

The present study provides an overview of the important issues that exist with respect to flagging in the Australian General Practice training context. Flagging has been identified as a predictor of RACGP exam performance in registrars and a number of potential features of flagging models identified as candidates for future research. Additionally, the importance of early flagging has been reinforced and strategies for de-stigmatising flagging amongst stakeholders identified. Collectively, these findings provide useful information for maximising registrar exam success and an overview of future directions for research to further improve registrars' exam success and their capacity to become safe, independent practitioners who are lifelong learners.

References

Biostat Inc. 2014. Comprehensive Meta-Analysis. 3 ed. Englewood. NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. 2009. *Introduction to meta-analysis*, John Wiley & Sons, Ltd.

Magin, P., Stewart, R., Turnock, A., Tapley, A., Holliday, E. & Cooling, N. J. a. I. H. S. E. 2017. Early predictors of need for remediation in the Australian general practice training program: a retrospective cohort study. 22**,** 915-929.

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F. & Martin, S. J. B. M. E. 2016. The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. 16**,** 87.

Stewart, R., Cooling, N., Emblen, G., Turnock, A., Tapley, A., Holliday, E., Ball, J., Juckel, J. & Magin, P. J. M. T. 2018. Early predictors of summative assessment performance in general practice post-graduate training: A retrospective cohort study. 40**,** 1166-1174.

Appendix A: Detailed results of subgroup analyses

Table A1: Subgroup analyses of RTOs with and without clear flagging protocols

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	1.280	0.486	3.371	.617	***2.891***	***1.720***	***4.859***	***<.001***	2.112	1	.146
GPT1 flag	Failed exam	1.158	0.661	2.026	.608	***1.909***	***1.351***	***2.698***	***<.001***	2.220	1	.136
GPT2 flag	Failed exam	0.913	0.497	1.679	.770	***2.252***	***1.476***	***3.436***	***<.001***	***5.696***	***1***	***.017***
GPT3 flag	Failed exam	1.895	0.859	4.178	.113	***3.544***	***2.356***	***5.329***	***<.001***	1.901	1	.168
First flag in GPT2	Failed exam	***2.132***	***1.512***	***3.008***	***<.001***	***2.426***	***1.219***	***4.829***	***.012***	0.108	1	.742
First flag in GPT3	Failed exam	1.741	0.815	3.720	.152	***1.869***	***1.200***	***2.912***	***.006***	0.025	1	.874
Flagged in 2+ training terms	Failed exam	1.562	0.860	2.839	.143	***2.823***	***1.926***	***4.137***	***<.001***	2.673	1	.102
Flagged (any level)	Failed AKT	1.234	0.340	4.484	.749	***4.113***	***1.250***	***13.537***	***.02***	1.805	1	.179
Flagged (any level)	Failed KFP	1.120	0.367	3.419	.842	***3.230***	***1.648***	***6.332***	***.001***	2.536	1	.111
Flagged (any level)	Failed OSCE	1.334	0.601	2.960	.478	2.772	0.979	7.853	.055	1.195	1	.274

Note: ***Bolded*** figures are statistically significant findings ($p < .05$)

Table A2: Subgroup analyses of RTOs that do and do not distinguish different types of flagging

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	2.625	0.531	12.976	.237	1.320	0.418	4.173	.636	0.468	1	.494
GPT1 flag	Failed exam	1.184	0.580	2.415	.643	1.490	0.895	2.480	.125	0.265	1	.607
GPT2 flag	Failed exam	1.322	0.696	2.514	.394	1.079	0.492	2.367	0.85	0.155	1	.694
GPT3 flag	Failed exam	***2.698***	***1.246***	***5.843***	***.012***	1.894	0.717	5.004	.198	0.313	1	.576
First flag in GPT2	Failed exam	***2.449***	***1.319***	***4.547***	***.005***	***2.108***	***1.479***	***3.006***	***<.001***	0.169	1	.681
First flag in GPT3	Failed exam	2.099	0.963	4.577	.062	1.548	0.664	3.611	.312	0.269	1	.604
Flagged in 2+ training terms	Failed exam	1.514	0.772	2.972	.228	***2.489***	***1.756***	***3.528***	***<.001***	1.646	1	.199
Flagged (any level)	Failed AKT	3.337	0.578	19.284	.178	1.199	0.249	5.769	.821	0.726	1	.394
Flagged (any level)	Failed KFP	2.676	0.355	20.18	.340	1.247	0.358	4.343	.729	0.397	1	.529
Flagged (any level)	Failed OSCE	2.159	0.717	6.502	.171	1.576	0.728	3.413	.249	0.210	1	.647

Note: ***Bolded*** figures are statistically significant findings ($p < .05$)

Table A3: Subgroup analyses of RTOs that do and do not grade flags

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	***0.616***	***0.437***	***0.869***	***.006***	***2.606***	***1.131***	***6.004***	***.024***	***9.811***	***1***	***.002***
GPT1 flag	Failed exam	***0.834***	***0.589***	***1.181***	***.038***	***1.831***	***1.343***	***2.497***	***0***	***10.929***	***1***	***.001***
GPT2 flag	Failed exam	0.595	0.229	1.546	.286	***1.672***	***1.116***	***2.506***	***.013***	3.817	1	.051
GPT3 flag	Failed exam	1.115	0.620	2.004	.717	***3.595***	***2.425***	***5.329***	***<.001***	***10.558***	***1***	***.001***
First flag in GPT2	Failed exam	***2.742***	***1.473***	***5.104***	***.001***	***2.033***	***1.427***	***2.891***	***<.001***	.672	1	.412
First flag in GPT3	Failed exam	0.953	0.665	1.365	.791	***2.547***	***1.639***	***3.959***	***<.001***	***11.479***	***1***	***.001***
Flagged in 2+ training terms	Failed exam		(Insufficient Data)[a]				(Insufficient Data)[a]				(Insufficient Data)[a]	
Flagged (any level)	Failed AKT	0.672	0.303	1.491	.328	3.052	0.921	10.112	.068	***4.252***	***1***	***.039***
Flagged (any level)	Failed KFP	***0.520***	***0.367***	***0.737***	***<.001***	2.614	0.919	7.430	.072	***8.258***	***1***	***.004***
Flagged (any level)	Failed OSCE	0.913	0.319	2.617	.866	***2.523***	***1.143***	***5.570***	***.022***	2.286	1	.131

Note: ***Bolded*** figures are statistically significant findings ($p < .05$). [a]Denotes insufficient data to conduct a subgroup analysis.

Table A4: Subgroup analyses of RTOs that do and do not involve Assigned MEs in Remediation

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	***3.587***	***2.586***	***4.975***	***<.001***	0.715	0.420	1.216	.216	***25.663***	***1***	***<.001***
GPT1 flag	Failed exam	***2.165***	***1.687***	***2.779***	***<.001***	0.876	0.646	1.188	.394	***20.307***	***1***	***<.001***
GPT2 flag	Failed exam	***2.154***	***1.610***	***2.882***	***<.001***	0.727	0.461	1.147	.171	***15.477***	***1***	***<.001***
GPT3 flag	Failed exam	***3.832***	***2.893***	***5.076***	***<.001***	1.464	0.699	3.063	.312	***5.701***	***1***	***.017***
First flag in GPT2	Failed exam	***2.034***	***1.390***	***2.977***	***<.001***	***2.510***	***1.488***	***4.233***	***.001***	0.406	1	.524
First flag in GPT3	Failed exam	***2.393***	***1.652***	***3.466***	***<.001***	1.446	0.605	3.455	.407	1.088	1	.297
Flagged in 2+ training terms	Failed exam	***2.831***	***2.282***	***3.512***	***<.001***	1.185	0.837	1.677	.339	***17.447***	***1***	***<.001***
Flagged (any level)	Failed AKT	***5.122***	***2.770***	***9.473***	***<.001***	0.687	0.272	1.737	.428	***12.52***	***1***	***<.001***
Flagged (any level)	Failed KFP	***3.828***	***2.598***	***5.642***	***<.001***	***0.559***	***0.320***	***0.977***	***.041***	***30.757***	***1***	***<.001***
Flagged (any level)	Failed OSCE	***2.831***	***1.161***	***6.902***	***.022***	1.070	0.435	2.629	.883	2.271	1	.132

Note: ***Bolded*** figures are statistically significant findings ($p < .05$)

Table A5: Subgroup analyses of RTOs that do and do not record flags in an online system

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	2.696	0.294	24.692	.380	1.541	0.363	6.539	.588	0.172	1	.678
GPT1 flag	Failed exam	1.294	0.516	3.247	.583	1.397	0.873	2.237	.164	0.021	1	.885
GPT2 flag	Failed exam	1.426	0.548	3.710	.466	1.099	0.570	2.119	.779	0.195	1	.659
GPT3 flag	Failed exam	***2.002***	***1.026***	***3.904***	***.042***	2.361	0.988	5.643	.053	0.087	1	.768
First flag in GPT2	Failed exam	***2.484***	***1.273***	***4.847***	***.008***	***2.115***	***1.495***	***2.990***	***<.001***	0.176	1	.675
First flag in GPT3	Failed exam	1.471	0.955	2.103	.083	1.999	0.927	4.313	.077	0.609	1	.435
Flagged in 2+ training terms	Failed exam	1.722	0.711	4.170	.229	***2.183***	***1.448***	***3.292***	***<.001***	0.288	1	.633
Flagged (any level)	Failed AKT	2.696	0.294	24.692	.380	1.541	0.363	6.539	.588	0.172	1	.678
Flagged (any level)	Failed KFP	1.478	0.181	12.04	.715	1.644	0.498	5.425	.414	0.007	1	.931
Flagged (any level)	Failed OSCE	2.247	0.435	11.607	.334	1.644	0.790	3.421	.183	0.116	1	.733

Note: ***Bolded*** figures are statistically significant findings ($p < .05$)

Table A6: Subgroup analyses of RTOs that do and do not use benchmarking as a means of flagging

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	1.810	0.712	4.601	.213	1.426	0.179	11.386	.738	0.042	1	.837
GPT1 flag	Failed exam	1.337	0.795	2.250	.274	1.373	0.592	3.182	.460	0.003	1	.959
GPT2 flag	Failed exam	1.317	0.825	2.101	.248	0.919	0.134	6.294	.932	0.127	1	.722
GPT3 flag	Failed exam	***2.795***	***1.496***	***5.224***	***.001***	1.525	0.432	5.381	.512	0.713	1	.399
First flag in GPT2	Failed exam	***2.158***	***1.519***	***3.064***	***<.001***	***2.292***	***1.208***	***4.351***	***.011***	0.026	1	.871
First flag in GPT3	Failed exam	***2.281***	***1.286***	***4.045***	***.005***	1.155	0.556	2.357	.692	3.125	1	.145
Flagged in 2+ training terms	Failed exam		(Insufficient Data)[a]				(Insufficient Data)[a]				(Insufficient Data)[a]	
Flagged (any level)	Failed AKT	1.912	0.612	5.973	.265	1.872	0.090	38.745	.685	0	1	.990
Flagged (any level)	Failed KFP	1.679	0.535	5.267	.374	1.387	0.148	12.972	.774	0.022	1	.881
Flagged (any level)	Failed OSCE	1.851	0.865	3.959	.112	1.893	0.278	12.89	.514	0	1	.983

Note: ***Bolded*** figures are statistically significant findings ($p < .05$). [a]Denotes insufficient data to conduct a subgroup analysis.

Table A7: Subgroup analyses of RTOs that do and do not enable flagging of prospective registrars during selection

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	1.607	0.528	4.890	.404	1.851	0.436	7.864	.404	0.023	1	.879
GPT1 flag	Failed exam	1.510	0.859	2.653	.152	1.154	0.722	1.845	.550	0.516	1	.473
GPT2 flag	Failed exam	1.356	0.764	2.406	.297	0.968	0.317	2.952	.954	0.278	1	.598
GPT3 flag	Failed exam	***2.207***	***1.123***	***4.335***	***.022***	2.620	0.720	9.531	.144	0.053	1	.817
First flag in GPT2	Failed exam	***2.043***	***1.426***	***2.928***	***<.001***	***2.637***	***1.456***	***4.774***	***.001***	0.518	1	.472
First flag in GPT3	Failed exam	1.755	0.979	3.146	.059	2.046	0.676	6.196	.205	0.058	1	.810
Flagged in 2+ training terms	Failed exam	***1.921***	***1.161***	***3.180***	***.011***	1.873	0.674	5.203	.228	0.002	1	.965
Flagged (any level)	Failed AKT	1.940	0.444	8.476	.378	1.574	0.306	8.103	.587	0.035	1	.853
Flagged (any level)	Failed KFP	1.398	0.375	5.214	.618	1.964	0.368	10.477	.430	0.098	1	.754
Flagged (any level)	Failed OSCE	2.117	0.822	5.453	0.12	1.497	0.639	3.506	.353	0.285	1	.593

Note: ***Bolded*** figures are statistically significant findings ($p < .05$)

Table A8: Subgroup analyses of RTOs that do and do not have dedicated TCs

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	1.265	0.447	3.578	.657	2.298	0.615	8.596	.216	0.485	1	.486
GPT1 flag	Failed exam	1.164	0.833	1.628	.374	1.619	0.827	3.171	.160	0.740	1	.390
GPT2 flag	Failed exam	0.891	0.382	2.077	.789	1.615	0.875	2.979	.125	1.246	1	.264
GPT3 flag	Failed exam	1.979	0.638	6.138	.237	***2.616***	***1.323***	***5.176***	***.006***	0.171	1	.679
First flag in GPT2	Failed exam	***2.476***	***1.455***	***4.214***	***.001***	***2.056***	***1.410***	***2.998***	***<.001***	0.312	1	.577
First flag in GPT3	Failed exam	1.713	0.621	4.723	.299	***1.891***	***1.053***	***3.396***	***.033***	0.028	1	.868
Flagged in 2+ training terms	Failed exam	1.785	0.966	3.298	.064	***2.066***	***1.125***	***3.793***	***.019***	0.110	1	.740
Flagged (any level)	Failed AKT	0.968	0.241	3.883	.964	3.511	0.826	14.913	.089	1.585	1	.208
Flagged (any level)	Failed KFP	0.751	0.518	1.088	.131	***1.692***	***1.214***	***2.359***	***.002***	0.348	1	.555
Flagged (any level)	Failed OSCE	1.440	0.641	3.235	.377	2.371	0.859	6.542	.095	0.567	1	.451

Note: ***Bolded*** figures are statistically significant findings ($p < .05$)

Table A9: Subgroup analyses of RTOs that do and do not have formal diagnostic processes for flagged registrars

		Absent				Present				Difference		
			95% CI				95% CI					
Predictor	Outcome	RR	Lower	Upper	p Value	RR	Lower	Upper	p Value	Q	df	p Value
Flagged (any level)	Failed exam	***2.741***	***1.108***	***6.779***	***.029***	***0.526***	***0.382***	***0.724***	***<.001***	***11.347***	***1***	***.001***
GPT1 flag	Failed exam	1.522	0.949	2.440	.082	0.954	0.603	1.511	.841	1.926	1	.165
GPT2 flag	Failed exam	***1.630***	***1.078***	***2.466***	***.021***	***0.481***	***0.247***	***0.938***	***.032***	***9.282***	***1***	***.002***
GPT3 flag	Failed exam	***3.188***	***1.927***	***5.277***	***<.001***	0.802	0.523	1.230	.311	***16.765***	***1***	***<.001***
First flag in GPT2	Failed exam	***2.159***	***1.531***	***3.043***	***<.001***	***2.311***	***1.158***	***4.615***	***.018***	0.030	1	.862
First flag in GPT3	Failed exam	***2.278***	***1.437***	***3.614***	***<.001***	0.804	0.520	1.243	.327	***10.347***	***1***	***.001***
Flagged in 2+ training terms	Failed exam		(Insufficient Data)[a]				(Insufficient Data)[a]			(Insufficient Data)[a]		
Flagged (any level)	Failed AKT	***3.446***	***1.321***	***8.898***	***.011***	***0.413***	***0.217***	***0.785***	***.007***	***12.977***	***1***	***<.001***
Flagged (any level)	Failed KFP	2.770	0.905	8.481	.074	***0.463***	***0.302***	***0.710***	***<.001***	***8.570***	***1***	***.003***
Flagged (any level)	Failed OSCE	***2.272***	***1.086***	***4.756***	***.029***	0.847	0.248	2.890	.791	1.824	1	.177

Note: ***Bolded*** figures are statistically significant findings ($p < .05$). [a]Denotes insufficient data to conduct a subgroup analysis.

Appendix B: Detailed results of meta-regressions

Table B1: Results of meta-regressions with single categorical moderator variables

			Test of Model			Goodness of Fit		
Moderator	*B*	R^2	*Q*	Df	*p*	*Q*	Df	*p*
Intercept-only	0.646	0	0	0	1	80.15	6	<.001
Clear flagging protocols	0.9401	.02	1.11	1	.293	67.00	5	<.001
Distinguishing different types of flags	-0.6603	0	0.47	1	.492	79.17	5	<.001
Grading flags	***1.4012***	***.63***	***6.27***	***1***	***.010***	25.17	5	.0001
Assigned ME involved in remediation	***1.6259***	***.89***	***20.83***	***1***	***<.01***	11.21	5	.0474
Recording flags in online system	0.018	0	0	1	.986	77.97	5	<.001
Benchmarking as means of flag	-0.297	0	0.1	1	.751	60.62	5	<.001
Selection and selection interview can flag	0.1247	0	0.02	1	.892	70.58	5	<.001
Dedicated TCs	0.5697	0	0.45	1	.503	63.13	5	<.001
Formal Diagnostic Process	***-1.6212***	***.32***	***4.81***	***1***	***.028***	39.59	5	<.001

Note: Relationship examined in all regressions was between being flagged at any stage and failing any exam. ***Bolded*** figures denote moderators that significantly improve on the intercept-only model.

Table B2: Results of meta-regressions with multiple categorical moderator variables

		Test of Model			Goodness of Fit		
Moderator variables	R^2	*Q*	Df	*p*	*Q*	Df	*p*
Assigned ME involved in remediation+ Grading flags	***.85***	***18.00***	***2***	***.0001***	10.77	4	<.001
Assigned ME involved in remediation + Clear flagging procedures	***.85***	***17.39***	***2***	***.0002***	10.11	4	.0307
Assigned ME involved in remediation + Formal Diagnostic Process (FDP)	***.87***	***20.43***	***2***	***<.00001***	***8.15***	***4***	***.0862***
Grading flags + Clear flagging procedures	.50	5.26	2	.0721	25.12	4	<.001
Grading Flags + FDP	***.74***	***13.74***	***2***	***.001***	12.57	4	.0136
Clear Flagging Procedures + FDP	.16	4.19	2	.1233	36.88	4	<.001
Assigned ME involved in remediation + Grading flags + Clear flagging procedures	***.80***	***14.24***	***3***	***.0026***	9.68	3	.0141
Assigned ME involved in remediation + Grading Flags + FDP	***.81***	***17.49***	***3***	***.0006***	7.10	3	***.0689***
Assigned ME involved in remediation + Clear Flagging Procedures + FDP	***.77***	***13.91***	***3***	***.003***	***7.05***	***3***	***.0703***
Assigned ME involved in remediation + Grading Flags + Clear Flagging Procedures + FDP	.52	8.93	4	.0628	6.00	2	.0499

Note: Relationship examined in all regressions is between being flagged at any stage and failing any exam. ***Bolded*** figures denote moderators that significantly improve on the intercept-only model. ***Bolded and italicised*** figures for the goodness of fit data reflect models which support the null hypothesis that all heterogeneity has been explained.

Table B3: Results of meta-regressions examining the effect of Medical Educator to Registrar ratio

					Test of Model			Goodness of Fit		
Predictor	Outcome	Model	*B*	R^2	*Q*	Df	*p*	*Q*	Df	*p*
Flagging (at any point)	Pass/fail any exam	Intercept-only	.646	0						
		ME:Registrar Ratio	-0.004	0	0.01	1	.920	76.88	5	<.001
GPT1 Flagging	Pass/fail any exam	Intercept-only	0.165	0						
		ME:Registrar Ratio	0.004	0	0.04	1	.845	21.97	5	<.001
GPT2 Flagging	Pass/fail any exam	Intercept-only	0.107	0						
		ME:Registrar Ratio	0.002	0	0.01	1	.928	24.91	5	<.001
GPT3 Flagging	Pass/fail any exam	Intercept-only	0.737	0						
		ME:Registrar Ratio	0.003	0	0.01	1	.934	44.40	5	<.001
First flag in GPT2	Pass/fail any exam	Intercept-only	.7776	0						
		ME:Registrar Ratio	.0001	0	0	1	.993	1.42	5	.923
First flag in GPT3	Pass/fail any exam	Intercept-only	.709	0						
		ME:Registrar Ratio	-.0035	0	.02	1	.902	31.22	5	<.001
Flagging in multiple training terms	Pass/fail any exam	Intercept-only	.2031	0						
		ME:Registrar Ratio	.014	0	0.58	1	.445	14.43	4	.006
Flagging (any point)	AKT pass/fail	Intercept-only	0.375	0						
		ME:Registrar Ratio	0.006	0	0.01	1	.909	34.85	5	<.001
Flagging (any point)	KFP pass/fail	Intercept-only	0.402	0						
		ME:Registrar Ratio	0.001	0	0	1	.980	59.02	5	<.001

Predictor	Outcome	Model	*B*	R^2	Test of Model			Goodness of Fit		
					Q	Df	*p*	*Q*	Df	*p*
Flagging (any point)	OSCE performance	Intercept-only	1.244	0						
		ME:Registrar Ratio	-0.019	0	0.31	1	.578	3.33	5	.650

Note: ***Bolded*** figures denote moderators that significantly improve on the intercept-only model.

Designing a Workplace-based Assessment Framework for Australian General Practice

Dr Taryn Elliott (GPEx)
Assoc Prof Jill Benson AM (GPEx)
Dr Emily Kirkpatrick (GPEx)
Prof Lambert Schuwirth (Prideaux Centre, Flinders University)

In collaboration with

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Acknowledgements

GPEx would like to acknowledge that this research project has been supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

GPEx would also like to acknowledge:

- The Workplace-based Assessment Framework Working Group: Dr Taryn Elliott (GPEx), Assoc Prof Jill Benson AM (GPEx), Dr Emily Kirkpatrick (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Mr Shaun Prentice (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Nina Grujic (GPEx).
- Steering Group members: Assoc Prof Jill Benson AM (GPEx), Prof Lambert Schuwirth (Prideaux Centre, Flinders University), Dr James Brown (EV), Dr Cat Kirby (EV), Dr Michael Bentley (GPTT), Mr Tim Clement (MCCC), Dr Lina Zbaidi (NTGPE), Dr Colleen Bradford (WAGPET), Dr Karin Jodlowski-Tan (RVTS), Prof Parker Magin (GP Synergy), Dr Graham Emblen (GPTQ), Dr Taryn Elliott (GPEx), Ms Kiara Beens (GPEx), Ms Michelle Pitot (GPEx) and Prof Cees Van der Vleuten (Maastricht University) for their guidance and input into this research.

GPEx In collaboration with

Flinders UNIVERSITY

This research project is supported by the Royal Australian College of General Practitioners with funding from the Australian Government under the Australian General Practice Training program.

Introduction

Project Context

GPEx was awarded a special Education Research Grant by the Royal Australian College of General Practitioners (RACGP) to design a Workplace-based Assessment (WBA) Framework for use within Australian General Practice Training.

Figure 1 below from "*The RACGP Education Research Grants 2018-2019 Special Grant Round application guide for RTOs*" shows the requirements for inclusion within this project.

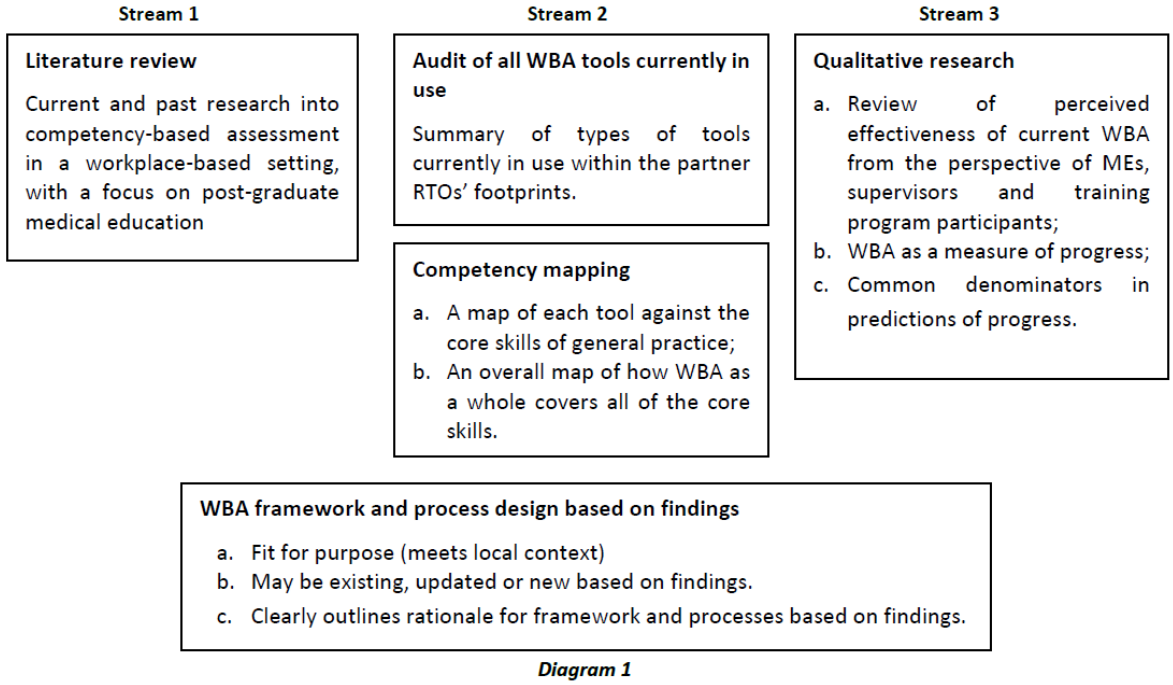


Figure 1. Project requirements set by the RACGP for the Special ERG

To meet the above-mentioned RACGP requirements, GPEx completed six project streams:

- Stream 1: Literature review
- Stream 2: Australian Regional Training Organisation (RTO) and Remote Vocational Training Scheme (RVTS) WBA audit
- Stream 3a: General practice registrar, supervisor and Medical Educator (ME) use of consultation observation as an educational and assessment tool
- Stream 3b: A qualitative investigation of acceptability of WBAs in AGPT
- Stream 3c: Evaluation of EPAsStream 3d: WBAs as a predictor for RACGP exam performance.

The findings from each of these streams have been reported in the previous chapters.

As per the initial project requirements, the findings from each of the project streams needed to be used to inform the development of an evidence-based, practical and contextualised WBA framework which:

- Is fit for purpose (meets local contextual needs).
- May be existing, updated or new (based on findings).
- Clearly outlines the rationale for the framework and processes, based on the findings.

This chapter describes the process used to develop the WBA Framework for Australian General Practice Training (the Framework), which takes into consideration the results from the various project streams, an environmental scan and expert feedback and input.

Method

Overall process for Framework design

The following steps were taken in designing the Framework:

- Completion of an environmental scan.
- Creation of a triangulation table.
- Development of a draft Framework.
- Final publication of the Framework.

Expert feedback and input was gathered throughout the design process through the WBA Framework Working Group and the Project Steering Group. Professor Cees Van der Vleuten reviewed and endorsed the final Framework.

The Framework Working Group

A Framework Working Group was established to ensure that the Framework was developed according to the outcomes of the project streams and embedded within the standards and policy context of Australian General Practice Training.

The Working Group consisted of the GPEx Lead Research Medical Educator (Assoc. Prof Jill Benson AM), the Lead Researcher for the Prideaux Institute (Prof Lambert Schuwirth), the GPEx Junior Medical Educator (Dr Emily Kirkpatrick), GPEx Research Officer (Mr Shaun Prentice), GPEx Project Manager (Ms Michelle Pitot), GPEx Research Manager (Dr Taryn Elliott) and GPEx Project Officer (Ms Kiara Beens).

The Working Group was responsible for each of the stages of the Framework design including:

- Conducting an environmental scan to determine the standard and policy context within which the Framework would need to operate and to identify other WBA frameworks and models from Australia and internationally.
- Reading and understanding the reports from each project stream in order to extract the important themes that would inform the Framework.
- Creating a triangulation table to synthesise the results from each project stream and the environmental scan.
- Developing a draft Framework, integrating results from the triangulation table.
- Consulting with and seeking expert feedback and input from the Project Steering Group and Professor Cees Van der Vleuten.
- Final publication of the Framework.

Steering Group

The project Steering Group consisted of representatives from: Eastern Victoria GP Training (EV), General Practice Training Tasmania (GPTT), Murray City Country Coast GP Training (MCCC), Northern Territory General Practice Education (NTGPE), Western Australian General Practice Education and Training (WAGPET), Remote Vocational Training Scheme (RVTS), GP Synergy and General Practice Training Queensland (GPTQ).

In regards to Framework design, the Steering Group were responsible for:

- Providing feedback on the identification of key findings from each project stream for inclusion in the triangulation table.
- Providing input and feedback on the draft Framework.

Environmental scan

An environmental scan was completed to ensure:

- The relevant standards and policy context within Australian General Practice Training were identified and considered.
- Other national and international WBA frameworks and models were considered in the design.

The environmental scan was completed using the following steps:

1. Identification of documents
   Documents were identified by:
   - Drawing on the expertise of both the Steering Group and Working Group to identify relevant documents for review and summation.
   - Completing an internet search to identify documents that may be relevant.

2. Review of documents
   All identified documents were reviewed by the Working Group and sections relevant to the Framework design were extracted and discussed.

3. Identification of relevant information
   The Working Group determined relevant information from the document review which should be integrated into or referenced within the WBA Framework.

Triangulation of research streams

Within a research project as complex and interrelated as this one, it is important to understand the results from each project stream within the overall context of the results from all other project streams. To assist with this process a triangulation table was created. This table was based on the theory of triangulation, as used within the social sciences. Triangulation is often used to address the internal validity of a project because it uses more than one method to answer a research question (Barbour, 2001). More importantly within this project, triangulation was useful to build a more comprehensive understanding of the key findings emerging from each project stream and how they are inter-related.

Figure 2 below shows how the themes from each stream of research documented in the triangulation table, in combination with results from the environmental scan, were used to develop the Framework.

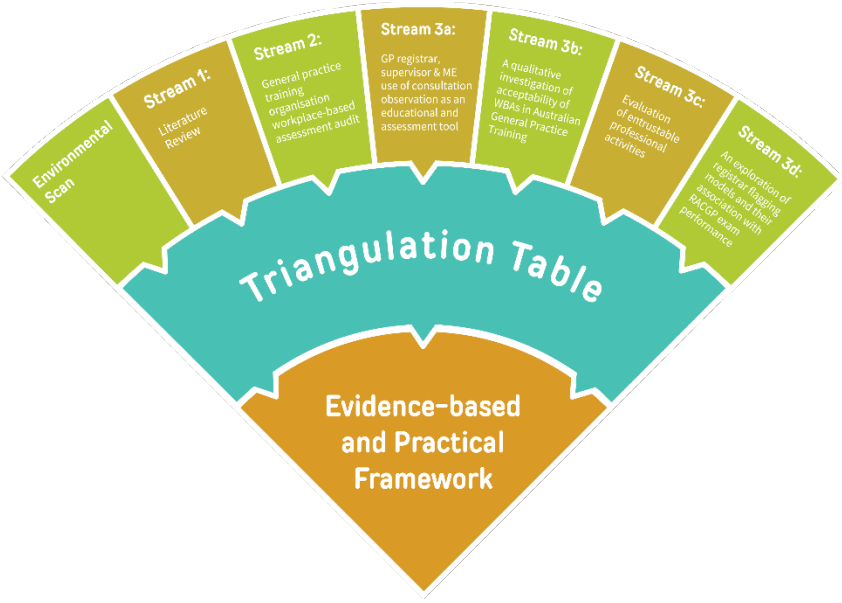


Figure 2. Evidence informing the development of the Framework

In order to create the triangulation table, as each project stream and the environmental scan were completed, key findings were themed and extracted by the Lead Researchers. Each theme was documented within the triangulation table as a separate row. The Framework Working Group then liaised with the Lead Researchers from each project stream to ensure that the triangulation table indicated which project streams supported this theme, and what key information was learned in regard to this theme.

To refine the triangulation table, all themes were reviewed and refined by the Working Group using an iterative process to ensure that similar themes were merged where appropriate. The Working Group used a consensus technique to guide this process.

Developing the Framework structure and content

To inform the Framework development, initially the themes captured within the triangulation table were grouped into key elements for inclusion in the Framework. This informed a basic structure for the Framework visual model. The notes from the table were then turned into a series of recommendations and a narrative summary which informed the WBA Framework content. Each recommendation within the narrative was cross-referenced to the project streams recorded against this theme within the triangulation table.

The overall Framework structure was informed by the key elements and themes within the triangulation table and a review of other frameworks identified through the environmental scan.

The final Framework was reviewed against the RACGP standards to confirm alignment, after which the Framework was reviewed by the Working Group and the Steering Group, and feedback was incorporated.

Results and Discussion

Environmental scan

Workplace-based assessment Frameworks environmental scan

Table 1 shows the documents identified through the environmental scan to determine other national and international work-place based assessment frameworks and models that were considered in the design.

Table 1. Documents identified through WBA Framework environmental scan

Doc No.	Document name	Organisation Responsible	Country	Publication year
1	Work-based assessment: a practical guide	Royal College of Physicians and Surgeons of Canada / Royal Australasian College of Physicians, Royal Australasian College of Surgeons	Australia/ Canada	2014
2	Framework for Workplace Based Assessment	Royal College of General Practitioners	UK	2019
3	Royal College of General Practitioners Curriculum Blueprint	Royal College of General Practitioners	UK	2015
4	Workplace based assessment: A guide for facilitators	London Deanery	UK	2012
5	Professional development Framework for educators	London Deanery	UK	2018
6	Fellowship assessment: Standards and guidelines for candidates	The Royal New Zealand College of General Practitioners	New Zealand	2017
7	Fellowship Pathway Regulations	The Royal New Zealand College of General Practitioners	New Zealand	2017

Documents discovered from the environmental scan were all from within the postgraduate medical training environment from Australasia, Canada, New Zealand and the United Kingdom and were published between 2014 and 2017.

The first three documents listed in Table 1 provided examples of content that might be relevant for inclusion within a Workplace-based Assessment Framework for Australian General Practice Training (document number 1 and 2). These WBA rameworks included:

- Theory and rationale relating to WBA
- A description of the competencies assessed
- Indicators of potential underperformance
- A description of WBA tools used (how the competencies were assessed)
- Examples of mapping WBAs against competencies
- Information about trainee portfolios and documentation of assessment feedback
- Description of how to establish an appropriate context for implementing WBAs (e.g. programmatic assessment, assessor support and training)
- Description of existing barriers to WBA and how to overcome these barriers.

These topics have informed the development of our WBA Framework. The learning log and personal development plans used within the Royal College of General Practitioners Framework provide a promising model for adaptation to an Australian context, translating to tools for planning and monitoring reflection, development of action plans and learning. Procedural skills are also assessed within this model from the UK. Assessment of procedural skills is not currently a formal requirement in RACGP training. .

Document four and five, *Workplace based assessment: A guide for facilitators* and *Professional development Framework for educators* provide useful information on underpinning theories, practical application of WBAs, and a framework for assessor training. These documents are relevant references for assessor training and support.

Documents six and seven provide information about how Fellowship in general practice is assessed and managed within New Zealand. These documents provide useful insights relevant to the Australian context. Firstly, they provide an example of how WBAs have been mapped against the competencies required for Fellowship for the Royal New Zealand College of General Practitioners. This model has been reviewed and used to inform mapping of WBAs against the RACGP domains of general practice and core competencies within the current Framework development.

Secondly, the documents detail mandatory WBAs required for Fellowship by the Royal New Zealand College of General Practitioners. Formative assessment requirements are more intensive in the initial general practice community training terms within this model and mandatory WBAs include:

- Video Consultations.
- In-practice visits from a ME.
- Patient feedback survey.
- Colleague multi-source feedback.
- Professional development plan.

This provides a similar range of WBAs to those identified through the audit of RTO WBAs, which was completed within stream 2 of this project. The recommendation of scheduling more frequent WBAs within earlier training terms, also reflects the findings from the project streams and was integrated into the Framework.

Finally there are specific details about completion of a Fellowship Assessment Visit. Candidates for Fellowship of the Royal New Zealand College of General Practitioners are assessed using this method. Document 5 states that:

> *"The purpose of the Fellowship assessment visit is to examine the candidate's actual practice to ensure that it is safe, competent and meets the standards for Fellowship of the College. In the Fellowship assessment process, candidates are assessed by a visiting assessor against the standards that are set out in this booklet. All indicators listed are essential, and must be met in order for the candidate to gain Fellowship of the College. In addition to evidence collected during the visit, the assessor has available to them other education programme materials, activities and assessments that relate to the candidate. This includes medical educator reports, patient and colleague feedback survey results, audits, and examination results."*

It is noted that this method takes into account the work-based context and a programmatic assessment philosophy. Our recommendations within the Framework include a similar final assessment. The recommended 'STAR' assessment (Statement of Awarded Responsibility) can occur once two different assessors (supervisor, ME, External Clinical Teacher) assess the registrar as 'safe to practise unsupervised' in all areas, and the registrar is in GPT3. This would involve a direct observation of 4-6 patients followed by random case analysis by a trained external clinical teacher (eg senior Medical Educator) using entrustment scales mapped to the curriculum. In order to meet final Fellowship assessment requirements the registrar must be deemed 'safe to practise unsupervised' in the STAR assessment's global EPA, and successfully complete the College's written exam requirements.

Context for the Framework

Table 2 shows the documents identified through the environmental scan exploring the relevant standard and policy context that needed to be considered in the Framework design.

All documents identified were Australian, from either the RACGP or AMC. Documents were current policy and standard documents, and were published between 2011 and 2019.

Table 2. Documents identified through WBA Framework context environmental scan

Doc No.	*Document name*	*Organisation Responsible*	*Publication year*
7	RACGP Standards for Training Providers (2nd edition)	RACGP	2017
8	RACGP Curriculum (2011)	RACGP	2011
9	RACGP Curriculum (2016)	RACGP	2016
10	RACGP Competency Profile	RACGP	2015
11	Vocational Training Pathway – Examination Eligibility Policy	RACGP	2016
12	Fellowship of the Royal Australian College of General Practitioners Policy	RACGP	2016
13	Fellowship Exams Candidate Handbook	RACGP	2019
14	Standards for assessment and accreditation of specialist medical programs and professional development programs by the Australian Medical Council	Australian Medical Council (AMC)	2015
15	AMC Workplace Based Assessment Accreditation Guidelines and Procedures	AMC	2019

The RACGP Standards for Training Providers (2nd edition) was comprehensively reviewed to determine the standards relevant to the development of a WBA framework. These standards were populated into a table and notes regarding implications for the Framework added, in order to ensure that the resultant Framework was in line with the existing RACGP standards (see Appendix A).

The RACGP Curriculum (2011 and 2016) and the RACGP Competency Profile were reviewed and used to map the WBAs detailed within the Framework. It was also identified that these documents were important reference points for the Framework.

Relevant RACGP policies and AMC standards were also identified and reviewed and it was found that the resultant Framework did not breach any of these established guidelines.

Finally the ACRRM curriculum, standards, examination handbook and procedural log were reviewed for context. It was noted by the Working Group that while these documents are aligned with the WBA Framework, they do not directly inform it.

It was recommended by the Working Group that the ACRRM procedural log and the procedural skills list from the 2011 RACGP curriculum filled a gap identified within the projects streams. Therefore, a procedural skills log has been included within the Framework.

Triangulation of research streams

The triangulation table contained 82 key 'themes', which emerged from each of the project streams and the environmental scan. These themes fell under the broad elements of: WBA tools, assessors, trainees and context. The triangulation table can be found in Appendix B.

Within the triangulation table 'themes' are mapped against each project stream where evidence to support this theme was discovered. The majority of themes were mapped against multiple project streams (n=80). A notes column was included where information was recorded to assist to describe the theme.

Developing the Framework structure and content

The Framework was designed by drawing together the results from the project streams and the environmental scan, in combination with Australia-wide consultation and expert input and feedback.

An overview of the key messages for inclusion within the Framework is reported below including: key elements, the visual model, assumptions, recommendations and structure. These provide the skeleton for the Framework. The resulting Framework has been included as the final chapter of this report and is written as a stand-alone document for ease of use and accessibility.

Key elements of the Framework

The themes emerging from the triangulation table fell under the broad elements of: WBA tools, assessors, trainees and context. These four elements form the backbone for the construction of the Framework. Each element must be considered to ensure an effective and contextual framework can be developed (see Figure 3).

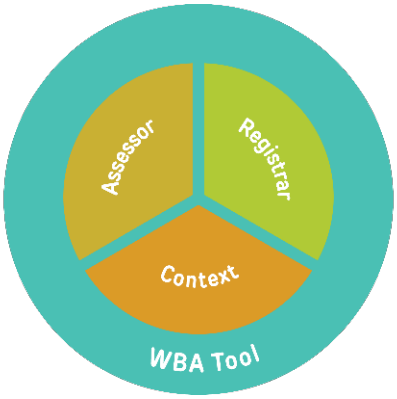


Figure 3. Essential elements informing the Framework

Visual model

The evidence for the use of programmatic assessment to inform decisions about a registrar's progress is clear from the literature and project streams. Therefore, the Framework is based on a programmatic assessment approach. Figure 4 shows how each of the four key Framework elements (WBA tools, the assessor, the trainee and the context) operate within the context of programmatic assessment to make decisions about a registrar's progress (see Figure 4).

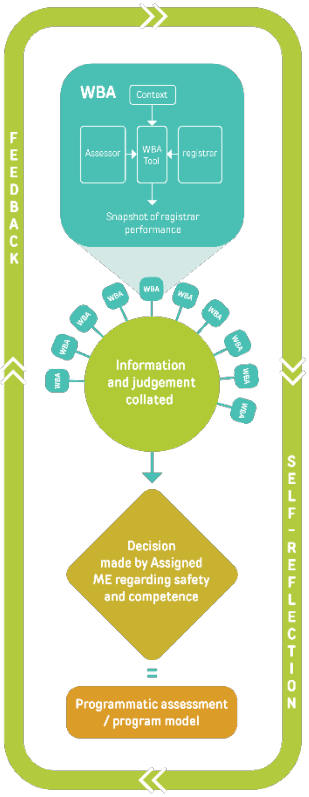


Figure 4. Visual model of the Framework

Structure for the Framework design

A structure for development of the Framework was established by the Working Group based on a review of existing WBA frameworks (identified through the environmental scan) moderated against the outcomes from this project. Table 3 outlines the structure for Framework development (see Table 3).

Table 3. Structure for Framework design

Section	*Broad content headers*
Introduction	• Overview of workplace-based assessment • Why was this Framework designed? • How was this Framework designed?
The Framework Overview	• Visual model of Framework • Assumptions which underpin the Framework • Theoretical underpinning
WBA Tools	• Recommended workplace-based assessment tools • Recommended implementation plan for workplace-based assessments • Mapping of workplace-based assessments to the RACGP domains of general practice • WBA tool delivery • WBA purpose • Mapping registrar progression • WBA tool features
Assessors: Supervisors, Medical Educators and External Clinical Teachers	• Assessor roles • Multiple assessors • Feedback • Assessor Training
High-stakes assessments	• Flagging and remediation • Consequences for non-completion of WBAs • Programmatic assessment for high-stakes decisions
Registrars	• Training required prior to undertaking WBAs • Self-reflection and training in accepting feedback • Tools for self-reflection • Empowerment
Context	• Training organisation • Practice context
Conclusion	

Conclusion

This chapter has described how each of the project streams and environmental scan have been synthesised to inform the Framework creation. An overview of the key messages for inclusion within the Framework has been reported, including: key elements, the visual model, assumptions, recommendations and structure.

The resulting Framework has been developed based on extensive national stakeholder consultation. Input has been given from eight of the nine Regional Training Organisations and the RVTS at both a project governance and operational level. The Framework incorporates feedback from registrars, supervisors, MEs and RTOs across urban, rural and remote areas within every state of Australia.

The Framework has been informed by internationally recognised experts in the field of medical education. The project was guided by Professor Lambert Schuwirth (Prideaux Centre, Flinders University). Steering Group member and international expert in WBA, Professor Cees Van der Vleuten also provided project guidance and endorsement of the final Framework. Professor Cees Van der Vleuten writes in the Framework foreword:

"I am delighted to have been of assistance with this formidable research project. The resultant workplace-based assessment framework and implementation plan is comprehensive and clear. This will help to guide the future of general practice training in Australia."

The robust process used to facilitate the Framework development has ensured that it is evidence-based, practical and contextualised. The Framework endorses many of the current practices throughout Australia but also provides additional recommendations based on the evidence collated. The environmental scan has confirmed that the Framework is aligned with the current RACGP curriculum and standards.

This project has culminated in the development of the first Workplace-based Assessment Framework for Australian General Practice Training. The Framework will be an important tool for RACGP and training organisations to use to inform the design of regionally appropriate WBA systems. This Framework will ensure that workplace-based assessment can be administered in an acceptable and appropriate manner. It will provide the building blocks to assist the RACGP to ensure that the quality of WBA can be standardised across Australia rather than the method in which assessments are performed.

References

Australian Medical Council (2015). *Standards for assessment and accreditation of specialist medical programs and professional development programs by the Australian Medical Council*, https://www.amc.org.au/wp-content/uploads/accreditation_recognition/specialist_edu_and_training/assessment/standards_for_assessment.pdf, (accessed May 2019).

Australian Medical Council (2019). *AMC workplace based assessment accreditation guidelines and procedures*, http://www.amc.org.au/wp-content/uploads/2019/05/Accreditation-of-WBA-Providers-Standards-and-Procedures-2019-secured.pdf, (accessed May 2019).

Barbour, R.S. (2001). Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ*; 322(7294): 1115–1117.

London Deanery (2012). *Workplace based assessment: A guide for facilitators*, https://faculty.londondeanery.ac.uk/guide-for-facilitators/files/WPBA%20facilitators%20notes.pdf (accessed February 2019).

London Deanery (2018). *Professional development Framework for educators*, https://faculty.londondeanery.ac.uk/new-multiprofessional-framework-for-educators/professional-development-framework-for-educators-2018, (accessed February 2019).

Royal Australian College of General Practitioners (2011).*The RACGP curriculum for Australian general practice,* https://www.racgp.org.au/FSDEDEV/media/documents/Education/Curriculum/2011-Curriculum.pdf, (accessed February 2019).

Royal Australian College of General Practitioners (2015). *Competency profile of the Australian general practitioner at the point of Fellowship,* https://www.racgp.org.au/download/Documents/VocationalTrain/Competency-Profile.pdf, (accessed February 2019).

Royal Australian College of General Practitioners (2016). *Curriculum for Australian general practice*, https://www.racgp.org.au/download/Documents/Curriculum/2016/core-skills-unit_v10.pdf, (accessed November 2018).

Royal Australian College of General Practitioners (2016). *Fellowship of the Royal Australian College of General Practitioners policy,* https://www.racgp.org.au/education/registrars/fellowship-pathways/policy-framework/policies/fellowship, (accessed May 2019).

Royal Australian College of General Practitioners (2016). *Vocational training pathway – examination eligibility policy*, https://www.racgp.org.au/download/Documents/Policies/Assessment/2015/VTP-Examination-Eligibility-Policy.pdf, (accessed May 2019).

Royal Australian College of General Practitioners (2017). *RACGP standards for training providers (2nd edition*), https://www.racgp.org.au/education/education-providers/regional-training/standards-for-general-practice/standards-2nd-edition, (accessed February 2019).

Royal Australian College of General Practitioners (2019). *Fellowship exams candidate handbook* (2019). https://www.racgp.org.au/education/registrars/fellowship-pathways/policy-framework/policies/fellowship-exams-candidate-handbook, (accessed May 2019).
Royal College of Physicians and Surgeons of Canada, Royal Australasian College of Physicians, Royal Australasian College of Surgeons (2014). *Work-based assessment: a practical guide*, https://www.surgeons.org/media/20786937/bkt_tripartite_wba__march_7__2_.pdf, (accessed February 2019).

Royal College of General Practitioners (2015). *Royal College of General Practitioners curriculum blueprin*t, https://www.rcgp.org.uk/training-exams/training/gp-curriculum-overview.aspx, (accessed February 2019).

Royal College of General Practitioners, (2019). *Framework for workplace based assessment.* https://www.rcgp.org.uk/training-exams/training/mrcgp-workplace-based-assessment-wpba.aspx, (accessed February 2019).

The Royal New Zealand College of General Practitioners (2017). *Fellowship assessment: standards and guidelines for candidates*, https://oldgp16.rnzcgp.org.nz/assets/New-website/Become_a_GP/FellowshipAssessmentStandards2017FINAL4APR.pdf, (accessed February 2019).

The Royal New Zealand College of General Practitioners (2017). *Fellowship pathway regulations*, https://oldgp16.rnzcgp.org.nz/assets/New-website/Become_a_GP/2017-Fellowship-Regulations-version-1-002.pdf, (accessed February 2019).

Appendix A: Review of standards for Workplace-based Assessment Framework design

The *RACGP standards for training providers* were reviewed to determine relevant standards related to the development of a Workplace-based Assessment Framework. The Table over the following pages documents the standards identified as highly relevant, and what implications have been identified and embedded within the Framework. This Table was used as a checklist for Framework review.

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*How this is addressed in the Framework*
1.1 Supervision is matched to the individual registrar's level of competence and learning needs in the context of their training post.	1.1.1 Competence is matched by appropriate supervision.	1.1.1.1 The registrar's competence is assessed prior to placement in a post and monitored throughout the training term	• Regular assessment during training • A reporting process between RTO and supervisor	• Regular assessment during training is required • Clear communication between supervisor and medical educator is required • EPAs used to assess level of supervision needed at initial safety assessment and then all mid and end-term assessments
		1.1.1.3 Appropriate supervision and training is matched to the registrar's learning needs and rate of progression.	• Registrars' needs are identified and documented in formal planned learning at every training placement. The training provider can determine the format of the planned learning to suit their context. • the medical education and supervision teams are involved in identifying and addressing the learning needs of the registrar	• Learning needs are identified and documented at least once per training placement as part of the learning log • MEs and supervisors are involved in identifying needs and addressing the needs • Flagging identifies those registrars who need more input • Number and type of WBA is determined by level of supervision required • EPAs are used because they measure the level of supervision required

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
1.1 Supervision is matched to the individual registrar's level of competence and learning needs in the context of their training post.	1.1.2 Feedback mechanisms are in place and the feedback is used to improve the quality of training and supervision	1.1.2.1 The registrar participates in timely, constructive feedback with the supervision team.	There are documented processes in place to ensure the registrar understands what is expected and how they are performing against expectations. Feedback is documented. The training provider is able to provide evidence of: how registrar progress is monitored, the processes used to give timely and appropriate feedback, how feedback is used.	• Expectations are made clear to the registrar • Feedback on registrar's performance against the expectations are made clear to the registrar • Feedback is documented • Registrars are trained in how to ask for and accept feedback • Supervisors and MEs are trained in how to give constructive feedback
	1.1.2 Feedback mechanisms are in place and the feedback is used to improve the quality of training and supervision.	1.1.2.3 Training posts are evaluated on a timely basis and the information is used to improve the quality of the post	The following data are collected – number of patients seen each week by the registrar – number of patients booked per hour for the registrar – diversity of patients seen by the registrar, (eg age, gender, Aboriginal and Torres Strait Islander status and reasons for the encounter).	• PETAL assessments allow data to be collected to demonstrate this standard being met, in addition to being useful for reflection and learning.
1.2 A model of supervision is developed in the context of the general practice training post to ensure quality training for the registrar and safety for patients.	1.2.1 The supervision model ensures that all elements of supervision can be addressed within the context of the post.	1.2.1.2 The training post has an RACGP approved model of supervision that meets or exceeds all supervision requirements.	Monitor registrar's competence (Criterion 1.1.1.1, Criterion 1.3.2.2, Criterion 2.3.1.1) – assist registrar with planning their learning (Criterion 1.1.1.3 and Criterion 2.2.1.1) – provide feedback to registrar (Criterion 1.1.2.1)	• Supervisors must monitor registrar competence, assist registrars with their learning planning and provide feedback to the registrar • Supervisors complete DOVs with the registrar • Supervisors and MEs can discuss the registrar's learning log at any time

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
	1.2.2 The supervision team is skilled and able to deliver quality training and patient safety.	1.2.2.2 Supervisors and the supervision team are skilled, and participate in regular quality improvement and professional development activities relevant to their supervisory role.	Formal feedback processes to monitor, improve and remediate the performance of supervisors.	• Supervisors must be trained and provided with professional development in WBAs • Training coordinators are trained in the flagging process and have a clear escalation process for registrars in difficulty • MEs discuss training with supervisors at the time of the DOV and assess the practice context
1.3 The practice environment is safe and supports training.	1.3.2 Learning opportunities and clinical experiences for the registrar meet patient safety requirements.	1.3.2.2 When working independently, registrars only undertake procedures and management of high-risk situations that they are competent to perform.	Patient safety is paramount during registrar training. Assessing a registrar's competence before allowing them to act without direct supervision not only reduces risk of litigation for the supervisor, it ensures patient safety. Registrars must be able to practice safely at the registrar level of training, that is, without direct supervision of every clinical encounter. Identified areas that pose high risk for patients and GPs are: diagnosis of malignancies, diagnosis of serious medical and/or life threatening problems, diagnosis of serious surgical problems, assessment of trauma, diagnosis and assessment of children, medication misadventure – prescribing error, inappropriate medication, drug administration error, adverse drug reaction, privacy procedures, procedures including intramuscular injections, venepuncture, ear syringing, minor	• Safety assessment undertaken within weeks 1 to 8 of training including DOV, RCA, MSF and MCQ. • Supervisor assesses level of supervision needed using entrustment scales twice a term for a range of EPAs mapped to the curriculum. • Procedure log should be signed off by the finalisation of training • Registrar keeps learning log to discuss with supervisor and ME and writes reflective learning plan each semester • Requirement for all DOVs includes assessment of level of supervision needed for patient safety

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			surgery, cryotherapy, implants and intrauterine device (IUD) insertion. The supervisor conducts a risk assessment of the registrar's ability to manage these high-risk situations within the context of the training post, level of supervision and current stage of training. The supervisor can assess the registrar's ability through consideration of training and experience, or through direct observation. The results of this assessment should form part of the registrar's planned learning.	
2.1 The registrar is selected and commences training.	2.1.2 The RACGP's Curriculum for Australian general practice is delivered	2.1.2.2 Training provider educational programs are clearly defined, consistent with the curriculum, appropriate to the learning needs of the registrar and the local context.	The training provider maps the RACGP's Curriculum for Australian general practice to the documented educational program. The educational program is published and easily accessed. Current priority areas are addressed. These include: Aboriginal and Torres Strait Islander health, training in early management of trauma and advance life support, paediatric competencies, critical thinking and research.	• WBAs are mapped to the curriculum • WBA requirements should be published and easily accessible • Paediatric competencies are assessed using PETAL, learning log, DOV, RCA and EPAs
		2.1.2.3 The educational program is planned, delivered, monitored and evaluated by an education team that is suitably skilled, experienced and adequately supported.	Medical educators and registrars play a major role in the planning and delivery of educational activities. The medical educator is fundamental in the development of the learning program ensuring consistency with the RACGP's Curriculum for Australian general practice. Direct or videotaped	• ECTVs should have both verbal and written feedback • Direct or videotaped observation can be utilised to give feedback • ECTVs need to be delivered by clinical teachers who are not the supervisor of the registrar •

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			observation of registrar consultations with verbal and written feedback to the registrar delivered by clinical teachers who are not the supervisor of the registrar is another essential element of the training program. These observations are known as external clinical teaching (ECT) visits. It is expected that each registrar has a minimum of five ECT visits throughout training. The training provider has adequate number of staff with appropriate qualifications and expertise available to conduct and administer training. There is a core group of medical educators with a high level of general practice educational expertise who are actively working in general practice. There is an adequate level of support staff for effective financial and general administration of the program. The training provider is responsible for providing relevant training and professional development opportunities for medical educators. Medical educators are expected to maintain and improve their knowledge and skills through continuing professional development. The training provider may encourage and will support medical educators who wish to undertake a higher degree in general practice or medical	• Each registrar has a minimum of five ECT visits throughout training • MEs require professional development in WBAs, giving feedback, how to give pastoral support, how to deal with difficult registrars and in remediation skills MEs should be will be supported to achieve formal higher qualifications in Medical Education

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			education. This support may include the provision of: information, advice and career counselling, leave to attend medical education courses, financial subsidies and support for medical education courses.	
2.1 The registrar is selected and commences training.	2.1.2 The RACGP's Curriculum for Australian general practice is delivered	2.1.2.4 A broad range of teaching, learning and assessment methods are used in a variety of settings and contexts using a variety of techniques, tools and technologies.	Teaching, learning and assessment strategies are adapted to accommodate the context, learning needs of the registrar and education setting. Feedback is delivered in a structured way to provide the registrar with an accurate assessment of progress	• WBAs should be adaptable to meet the needs of the registrar and the context • Feedback is structured to provide an accurate assessment of progress
		2.1.2.4 A broad range of teaching, learning and assessment methods are used in a variety of settings and contexts using a variety of techniques, tools and technologies.	*Education program principles* The education program delivered by the training provider is based on principles and standards of postgraduate medical education that ensures registrars develop the professional standing of a GP. Training providers will ensure that the education programs meet the following: The aims and goals of the program are clearly documented and readily accessible by all participants, The program will have a stated goal to train GPs who are competent for unsupervised general practice anywhere in Australia, and are able to provide high-quality primary healthcare services relevant to individual and community health needs, The program addresses the	• Assessment of needs must be across the five domains • Registrars who are trusted to practise unsupervised in all areas, can aim for a higher level of entrustment where they are trained and trusted to supervise junior learners. This could be medical students, interns or more junior registrars. • Experiential learning can be measured and assessed using PETAL and the learning log • The culture and WBA programme needs to encourage registrars to commit to life-long learning

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			learning needs of participants prescribed in all five domains described in the RACGP's Curriculum for Australian general practice, The program is underpinned by educational concepts and principles appropriate to professional postgraduate vocational training including – emphasising the integration of vocational training with undergraduate, postgraduate and continuing professional development – emphasising experiential learning through practical clinical experience – this involves consulting with patients who present with common and significant conditions that exemplify general practice – encouraging commitment by registrars to continuous improvement of their knowledge and skills throughout their professional careers.	
2.2 Registrars learn in a structured way in posts that are accredited and engaged in the teaching and learning process.	2.2.1 Post-based learning activities are planned, structured and referenced to curriculum, learning needs of the registrar and context of the post.	2.2.1.1 Registrar learning activities and the teaching strategies used are customised to the registrar's needs and training context.	The RACGP curriculum provides the framework for the education of registrars. The registrar and supervisor need to use the curriculum and learning programs to plan training post-based learning strategies. Teaching will be based on the registrar's planned learning and other perceived needs that may arise during training. Teaching within the training post should include a range of methods such as: direct observation,	• WBAs should map directly against the RACGP curriculum • DOVs, RCAs, and procedural skills demonstration are all included in the Framework

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			discussions on clinical problems and interesting cases, joint consultations, formal teaching on specific topics, review of consultations – recorded or observed, demonstrations and participation in clinical procedures, selected or random case analysis, small group discussions with members of the supervision team.	
			Planned learning -The registrar, in consultation with the supervisor and, where appropriate, the medical educator, develops a plan for their learning that is practical and relevant, to ensure the adequate planning of training post-based learning activities. The registrars need to discuss their experience and learning needs with their supervisors as early as possible to enable individualised planned learning to be developed. It is recommended that the formal planning of learning between the supervisor (and medical educator where possible) and the registrar has commenced by week four of each six months of training. This may take the form of a formal written plan for learning lodged with the training provider. The supervisor and registrar will regularly engage to review the learning, and if needed, modify the planned learning to ensure that the training post-based teaching and	• The registrar's learning log is regularly discussed with the supervisor and ME • Each term the registrar submits a formal learning plan signed off by the ME • All WBAs have self-reflection and direct links to action plans and learning associated with them • Registrars have a formal safety assessment at week 4-8 of their first placement • Registrars who are flagged will have a formal intervention plan developed and monitored • ECTVs must be supported by the practice • Procedure log is ongoing throughout training and will be discussed with the supervisor in each training post • PETAL monitors whether registrars are seeing a range of patients

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			learning activities match the needs of the registrar and training context. The training post supports access for a clinical teacher to undertake direct observation sessions (which could be by video review) as prescribed by the training provider.	
2.2 Registrars learn in a structured way in posts that are accredited and engaged in the teaching and learning process.	2.2.2 The registrar's learning and development is well supported.	2.2.2.1 The registrar is adequately prepared to participate fully in the operations and scope of practice in the training post.	The registrar has a structured induction to the practice that includes information about systems, resources, support and context. Training in how to use systems is included where appropriate. The registrar has an available and appropriately equipped area for conducting consultations.	• Orientation to each practice should include discussions about systems, resources, support in relation to the particular context of the practice (e.g rural, Aboriginal health) including an orientation to WBAs • At the beginning of the first placement, and optionally at the beginning of other placements, the registrar should sit in with the supervisor as an observer • Registrars should receive information at orientation about the timeline and expectations of WBAs throughout their training
		2.2.2.2 The registrar is provided with quality, safe and well supported learning opportunities.	The training provider is able to provide evidence of the registrar being exposed to an appropriate number and variety of patients for their stage of training and the context of training. For example, Aboriginal and Torres Strait Islander health training posts may differ in the number and variety of patients (longer consultations with patients with more complex and chronic conditions)	• PETAL assists in monitoring this standard while also providing information to assist in monitoring registrar progress and learning needs

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
2.3 The development of each registrar is optimised.	2.3.1 The progress of the registrar throughout training is monitored and addressed.	2.3.1.1 The registrar's progress is documented and readily available to the registrar, training post, training provider and RACGP.	Any problems are identified early and addressed. Each registrar receives at least two external clinical teaching visits from clinical teachers or general practice supervisors not working in the same training post in both GPT1 and GPT2, and one in GPT3/4. The training provider is able to provide evidence of: monitoring and addressing the registrar's progress in the training post, how the registrar's progress is monitored and addressed throughout training, including the involvement of cultural educators/mentors for registrars in Aboriginal and Torres Strait Islander health posts, how the registrar is able to view the documentation of their progress, regular training review meetings.	• The safety assessment in GPT1 ensures that any problems are identified early and addressed • This standard supports the WBA Framework recommendation for a documented portfolio of WBAs which is able to be viewed by registrar and signed off by the ME, and of a learning log and of sequential mid and end-term assessments that can map progress • Those registrars who are in Aboriginal and Torres Strait Islander posts will have additional cultural training and mentors and the MSF in this post will include Aboriginal Health Workers
2.3 The development of each registrar is optimised.	2.3.1 The progress of the registrar throughout training is monitored and addressed.	2.3.1.1 The registrar's progress is documented and readily available to the registrar, training post, training provider and RACGP.	*Training review meetings* The training provider has a documented process for the regular review of the registrar's training progress throughout training. Training review meetings between the registrar and a medical educator will occur prior to and after every training placement. Frank discussion about the registrar's progress to date will be informed by: feedback from previous supervisors and supervision teams, feedback from clinical teachers conducting external clinical visits, formative assessment outcomes,	• This emphasises the ME role in collation and interpretation of programmatic assessment and also when additional support or remediation is required • At a Training Review meeting, the ME will bring together information from the training portfolio from WBAs and stakeholders • Training review meetings are facilitated by the training coordinator. • Regular review of flags and intervention plans occur by the ME and registrar

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			registrar's self-assessment of competencies referenced against the curriculum, registrar's planned learning, registrar's FRACGP training portfolio (Criterion 3.3.1.3), registrar's log of educational events attended. In collaboration with the registrar, the medical educator consolidates and, if needed, modifies the registrar's training plan. Training provider staff will assist the medical educator and registrar to coordinate and implement the registrar's training plans. Training review meetings provide an opportunity to identify if the registrar is encountering any difficulties and whether extra assistance or remediation is required before the registrar can progress through training. Early identification of problems is encouraged to ensure that adequate support can be given to the registrar to aid progression through training.	
2.3 The development of each registrar is optimised.	2.3.3 At-risk registrars are identified and appropriate remediation implemented.	2.3.3.1 Learning intervention and remediation opportunities are identified and addressed.	The training provider has a clearly documented process for the early identification of registrars who need additional support. The registrar is informed of concerns as soon as they are identified. Additional support is readily available and tailored to the needs of the registrar. There is a fair and transparent process for managing	• This emphasises that flagging registrars should be a transparent process • Early safety analysis and ongoing monitoring of WBAs by MEs • Process for WBA, flagging and additional support needs to be documented and transparent • This supports more assessments early in training which would reduce as competence increases

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			registrars who do not progress after they have been through remediation. Peer-to-peer support is available to registrars through a position that is funded and managed by the training provider. The position provides pastoral care, confidential information/advice, and a registrar voice in planning and evaluating the training program. The training provider is able to provide evidence of: formal feedback processes in place to monitor and improve the performance of registrars, feedback documentation maintained in the registrar's training file, learning intervention, and remediation plans and outcomes, registrars having access to formal training provider-based peer advocacy and support, registrars being actively involved in the intervention and remediation process from the time of identification, cultural educators and mentors are involved in the remediation process for registrars in Aboriginal and Torres Strait Islander health training issues.	• All WBA feedback should be documented and accessible
2.4 The training provider delivers quality education and training.	2.4.3 Systems and processes support the education program and the registrars.	2.4.3.1 The systems and processes used to keep records, deliver training and monitor the progress of the registrar are up-to-date and secure.	The data and learning management system of the organisation is up-to-date, fit for purpose and overseen by a suitably qualified person or team of people. The system is regularly upgraded and maintained with minimal disruption to the training	• This standard supports that an online learning management system is used which is accessible to supervisors, the supervision team, training organisation staff and registrars • The system needs to be relevant to the context, reliable and user-friendly

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			registrars who do not progress after they have been through remediation. Peer-to-peer support is available to registrars through a position that is funded and managed by the training provider. The position provides pastoral care, confidential information/advice, and a registrar voice in planning and evaluating the training program. The training provider is able to provide evidence of: formal feedback processes in place to monitor and improve the performance of registrars, feedback documentation maintained in the registrar's training file, learning intervention, and remediation plans and outcomes, registrars having access to formal training provider-based peer advocacy and support, registrars being actively involved in the intervention and remediation process from the time of identification, cultural educators and mentors are involved in the remediation process for registrars in Aboriginal and Torres Strait Islander health training issues.	• All WBA feedback should be documented and accessible
3.1 The registrar is competent to commence training.	Outcome 3.1.2 The registrar is able to demonstrate competence to work under supervision as a GP in Australia.	Criterion 3.1.2.2 The registrar demonstrates the professional attributes expected of a GP.	Registrars are aware of and progressively establish general practice knowledge, skills and attitudes consistent with the RACGP's Competency profile of the Australian general practitioner. The registrar is expected to take active responsibility	• The RACGP's Competency profile of the Australian general practitioner provides a breakdown of core skills across each of the domains. This is referred to in the Framework, and WBAs are mapped against RACGP domains of general practice and core skills

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			for their own learning. The training provider is able to provide evidence of: how it demonstrates ongoing registrar progression in the development of their clinical practice and cultural competence, how it determines the registrar's readiness to accept responsibility for the clinical decisions that they make, how it assesses the professional attributes of the registrar against the RACGP's Competency profile of the Australian general practitioner, actions that promote shared learning responsibility between the registrar and training provider, strategies for managing/supporting registrars who fail to commit to their training.	• Cultural competence is discussed in the Framework • During Aboriginal and Torres Strait Islander Training Post placements, feedback from cultural mentors and Aboriginal Health Workers is encouraged (ie. through MSF)
3.2 The competence of the registrar is articulated and benchmarked to inform progress throughout training.	Outcome 3.2.1 There is a robust process of assessment.	Criterion 3.2.1.1 The competencies that the registrar must attain for successful completion of each training term and whole of training are identified.	It is the responsibility of the training provider, in collaboration with the medical education team, supervisor and registrar to assess the competence of the registrar prior to, during and at the end of each training term. The early assessment will inform the formulation of the planned learning that should include: the competencies to be achieved during the training placement, measures by which the competencies are defined, steps that the registrar needs to take along the way to develop the competencies, progress and achievements signed off by the	• Registrar should be informed about competencies to be achieved each term and how they will be measured. This is achieved through clear communication about WBAs, benchmarking and WBA orientation at practice • There is a documented process to assess and record the achievement of the competencies

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			supervisor at completion of training placement. The training provider has a documented process to assess and record the achievement of the competencies of registrars (Criterion 1.1.1.1 and Criterion 2.4.3.1). The training provider is able to provide evidence of: the registrar being informed about the competencies to be achieved, the competencies defined in measurable terms, defined outcomes the registrar needs to reach along the way to develop the competencies. It is the responsibility of the training provider to ensure that registrars are aware of the competencies expected to be achieved by the completion of training. These are clearly articulated in the RACGP's Curriculum for Australian general practice.	
3.2 The competence of the registrar is articulated and benchmarked to inform progress throughout training.	Outcome 3.2.1 There is a robust process of assessment	Criterion 3.2.1.2 The assessment methods ensure that the registrar's level of competence is measured against the competencies required for the stage of training and training post context.	Assessment methods use multiple sources and processes that are flexible and relevant to the registrar's stage and context of training. Assessments are benchmarked against the required competencies, and are comprehensive and integrated with the educational program. The training provider is able to provide evidence of: assessment methods, assessment results being used to improve its educational	• This standard is highly relevant to the project and aligns well with the Framework • Multiple assessment sources and processes should be used • Assessment should be flexible and relevant to the registrar's stage of training • Assessments need to be relevant to the context of training (eg. rural, Aboriginal Health)

Standard	Outcome	Criterion	Relevant notes from standards	Framework
			program. Guidance: An integral and critical part of the education and training in the program will be high-quality, regular formative assessment with constructive feedback to registrars on their performance. It is important that the program assesses the registrar's progress to provide information to guide the registrar's future learning activities. This is in addition to the monitoring undertaken by supervisors. To be effective, formative assessment activities should assess the key knowledge, skills, attitudes and behaviours required for independent general practice, be iterative, and provide information to guide registrar's future learning activities. Formative assessment principles The training program's formative assessment activities may: be designed to adequately assess across the range of knowledge, skills, attitudes and behaviours outlined in the RACGP's Curriculum for Australian general practice, be relevant to the training post, be used early enough, and with sufficient frequency, to provide the opportunity for registrars to regularly update their plans for learning, incorporate specific, timely and regular feedback to registrars about their performance, provide specific information about what needs	• The Framework references the RACGP curriculum and maps WBAs against domains and core skills. It is noted that WBA content should be based on the RACGP curriculum • The ME collates the WBAs, including the mid and end-term assessments to provide an overview of progress in order to give feedback to the registrar and the RTO

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
			to be improved and an agreed plan for how to go about making the desired changes, include the registrar's own assessments of their performance and learning needs, be documented in order to guide future learning and demonstrate progress, include an agreed plan for meeting significant learning needs and reassessing performance in identified areas, invoke remediation or 'registrar in difficulty' processes if appropriate. Assessment sources Assessments should occur across multiple sources and may include: external clinical teaching visit reports, supervisor and supervision team feedback – verbal and written training review meetings, registrar self-assessments, remediation assessment tools, patient surveys, adverse reports/critical incident reports. Results from individual and group assessments are important sources of data that should be analysed by the training provider to inform its quality improvement process.	
3.2 The competence of the registrar is articulated and benchmarked to	Outcome 3.2.2 Assessment results are used to monitor and	Criterion 3.2.2.1 The assessment methods ensure that the registrar's competence is known to the		• The assessment results are readily available to the registrar, supervisor and training organisation • Early identification of problems leads to support and remediation

Standard	*Outcome*	*Criterion*	*Relevant notes from standards*	*Framework*
inform progress throughout training.	improve performance.	registrar, supervision team and training provider, and are used to plan the registrar's learning.	Progressive assessment results track the current performance of the registrar with the performance expected for the level of training. Early identification of problems leads to support and remediation (Criterion 2.3.3.1). The assessment results are readily available to the registrar, supervisor and training provider while complying with privacy legislation. (Criterion 2.3.1.1). The training provider is able to provide evidence of: progressive assessment results, remediation that follows adverse outcomes of the assessment, how it determines what registrar information is shared and with whom. Guidance An integral and critical part of the education and training in the program will be high-quality, regular formative assessment with constructive feedback to the registrars on their performance. The training provider has processes to conduct quality formative assessment that are valid, flexible and fair to registrars in meeting their learning needs. Characteristics of high-quality assessment include validity, reliability, educational impact, acceptability, feasibility and efficiency. The assessment methods should form part of a supportive quality improvement process aimed at improving clinical	• High-quality, regular formative assessment with constructive feedback to the registrars on their performance • Quality formative assessments that are valid, flexible and fair to registrars in meeting their learning needs • Characteristics of high-quality assessment include validity, reliability, educational impact, acceptability, feasibility and efficiency. These are discussed in the Framework assumptions and theoretical underpinning • Trends of progressive assessments should be used to track a registrar's performance and allow for early identification of struggling registrars (Framework is based on programmatic assessment) • Trends of progressive assessments should be used as part of ongoing learning with the registrar so that all concerned have clear information on the areas that need to be addressed

Standard	Outcome	Criterion	Relevant notes from standards	Framework
			practice. Any significant deficiencies will ideally be discovered and corrected during training rather than at the FRACGP examination. Trends of progressive assessments should be used to track a registrar's performance and allow for early identification of struggling registrars. They should also be used as part of ongoing learning plan with the registrar so that all concerned have clear information on the areas that need to be addressed.	

Appendix B: Triangulation table

This Table contains the themes extracted from each research stream and the environmental scan. The key elements within this table are:

- Tools
- Assessors
- High-stakes assessment
- Registrars
- Context

The themes are mapped across the project streams and environmental scan to show where evidence for each theme was discovered.

Tools

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
Tool selection							
Initial assessment to determine safety and readiness for practice and appropriate level of supervision required. Standardised and led by the training organisation.		✓	✓	✓		✓	
DOVs should be done by the supervisor, by assigned ME and/or an external clinical teacher	✓	✓	✓	✓		✓	
Direct observation of the supervisor by the registrar should occur as well as by the supervisor of the registrar.				✓			
MSF (staff) should be online and include a question to the staff about patient satisfaction.		✓		✓		✓	
Take into account not only the supervisor's opinion but that of the other doctors and staff and patients in their practice (eg MSF).		✓	✓	✓		✓	
Random case analysis should be included both formally and informally.				✓		✓	
Patient Encounter Tracking and Learning (PETAL) tools will assist registrars and practices to identify gaps.		✓		✓			
Professionalism and communication skills should be assessed regularly using Term assessments, DOV and MSF.			✓	✓	✓	✓	
Every registrar should be able to perform a core list of procedures	✓			✓			✓

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
before completion of training-procedural log.							
Statement of awarded responsibility (STAR) assessment should be used.	✓			✓	✓		✓
A learning log should be kept.		✓		✓		✓	✓
Video reviews for self-reflection purposes with an option to discuss with supervisor and ME and for remediation.				✓		✓	
Mapping							
Each WBA should have a clear description of how it can be mapped against competencies and how it is assessed.	✓			✓	✓	✓	
Tool delivery							
Registrars should be clearly informed about each WBA and its purpose.	✓			✓	✓	✓	
WBAs should be defined as low, medium or high-stakes.	✓			✓		✓	
Participants for DOVs, video reviews, MSF and random case analysis should not be chosen by the registrar.	✓	✓		✓			
The training organisation should set up processes to ensure that specific, practical and timely feedback is given for each WBA.	✓	✓	✓	✓			
'Action or learning plan' from each assessment.	✓	✓		✓			
WBAs should be delivered using a user friendly, practical, integrated, electronic system.	✓	✓	✓	✓	✓	✓	

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
WBAs should leave opportunity for self-reflection by the registrar as to how they are tracking in comparison to their peers.	✓			✓	✓	✓	
Learning opportunities from WBAs should lead to dynamic learning plans that are actionable.	✓	✓		✓		✓	
Tool purpose							
Formative assessment purpose of WBA a priority.		✓	✓	✓	✓		
Professionalism and communication skills can be measured.			✓	✓	✓	✓	
WBAs monitor registrar safety.	✓	✓	✓	✓	✓	✓	
DOVs develop consulting skills, clinical skills and knowledge.			✓	✓	✓	✓	
Mapping registrar progression							
Ensure that there are regular longitudinal assessments from supervisors that are 'intuitive' as well as more formal assessments.			✓	✓	✓		
A programmatic assessment approach with collation of outcomes and opinions maximises the effectiveness of WBA as tools for mapping progression.	✓		✓	✓	✓	✓	
Tool features							
Narrative is essential in WBAs and if binary tick-boxes or Likert scales are used they should be accompanied by narrative.	✓	✓	✓	✓	✓	✓	

Theme	Stream 1	Stream 2	Stream 3a	Stream 3b	Stream 3c	Stream 3d	Environmental scan
All scales should have clear and meaningful entrustment-based definitions.	✓			✓	✓	✓	
The minimum expectation for each assessment should be defined and clear.	✓	✓		✓	✓	✓	
Standardised assessment criteria based on whether the registrar is safe to practise independently.	✓		✓	✓	✓	✓	
When a registrar is undertaking a placement in a culturally specific environment, the WBAs utilised should aim to gather opinions from those knowledgeable in cultural safety and competency.			✓				✓

Assessors

Theme	Stream 1	Stream 2	Stream 3a	Stream 3b	Stream 3c	Stream 3d	Environmental scan
Assessor roles							
Assessors may provide psychological support.		✓	✓	✓		✓	
Assessors have multiple roles.		✓	✓	✓		✓	
Multiple assessors							
Different assessors should be used for the formative assessment than those used for summative assessment.	✓			✓		✓	
The same group of assessors should conduct formative WBAs for each registrar to facilitate mapping of progression.	✓		✓	✓	✓	✓	

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
Multiple assessors should be used to complete assessments to ensure a variety of viewpoints are incorporated and to decrease bias.	✓	✓	✓	✓		✓	
Coordination between assessors is essential.			✓	✓			
Feedback							
Assessors should give timely feedback.	✓		✓	✓			
Debriefing discussions are valuable to better understand clinical reasoning.			✓	✓			
Feedback may need planned follow-up.	✓	✓	✓	✓			
Giving effective feedback is an art.			✓	✓			
Feedback to registrars should consider the needs of the registrar and the context in which they are working.	✓		✓	✓		✓	
Feedback should be outcome-based, specific, honest, observation-based, timely, and actionable.	✓	✓	✓	✓			
Assessor training							
Supervisors and MEs require training on: • culture of excellence expected of all registrars • purpose of formative assessment • how to complete formative assessment • self-reflection on their own thinking processes (eg unconscious biases etc)	✓		✓	✓	✓	✓	

Theme	Stream 1	Stream 2	Stream 3a	Stream 3b	Stream 3c	Stream 3d	Environmental scan
• the importance of narrative and appropriate vocabulary • how to give feedback • examples of how to give both verbal and written feedback • feedback on their completion of a formative assessment and undertaking the feedback process. • developing an understanding of and expectations for particular tools (eg RCA, EPAs).							
Regular 'network' meetings of supervisors and also of MEs who discuss WBAs, what works and doesn't.			✓	✓			
Regular meetings of supervisors who are involved with a registrar in a practice or in a town to discuss how the registrars are going.			✓	✓			

High-stakes assessment

Theme	Stream 1	Stream 2	Stream 3a	Stream 3b	Stream 3c	Stream 3d	Environmental scan
Flagging and remediation							
WBAs are useful tools to flag registrars and then useful for remediation and monitoring progress.	✓	✓	✓	✓	✓	✓	
There should be a multi-factorial assessment which feeds into the flagging and ongoing monitoring process.	✓			✓		✓	

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
Flagging should have a co-ordinated and timely monitoring, review and escalation process.	✓			✓		✓	
The flagging process should lead to a bespoke remediation plan tailored to the registrar.				✓		✓	
Flagging should be reviewed regularly and if it is a high-risk flag then by a panel.	✓	✓		✓		✓	
When making high-stakes decisions a variety of WBAs should be used and a panel should review the WBA portfolio.	✓					✓	
Flagging should be online and trackable.				✓		✓	
Registrars should know that they have been flagged.				✓		✓	
Consequences for non-completion							
Mandatory requirements should be followed up and have consequences if they are not completed.		✓		✓		✓	
Programmatic assessment for high-stakes decisions							
Summative assessment should be programmatic with a variety of different WBAs used and a variety of different assessors. Final assessment should include exams plus the programmatic portfolio from WBAs.	✓			✓		✓	
The final summative programmatic assessment should be by a panel.	✓	✓		✓		✓	
Summative assessments should not be done by the registrar's supervisor.	✓			✓		✓	

Theme	Stream 1	Stream 2	Stream 3a	Stream 3b	Stream 3c	Stream 3d	Environmental scan
Don't do away with the exam but have a 'blended' final assessment that includes information from the supervisor and MSF but also from External clinical teaching visits that include DOVs and random case analysis.				✓			✓
Registrars should not choose or personally know their high-stakes assessors.	✓			✓			

Registrars

Theme	Stream 1	Stream 2	Stream 3a	Stream 3b	Stream 3c	Stream 3d	Environmental scan
Registrars should also be trained in how to ask for and accept feedback and should understand the concept of formative assessment for learning (rather than of learning).	✓		✓	✓	✓		
The value and skills of self-reflection should be emphasised to registrars.	✓		✓	✓			
Registrars value feedback as important in progressing in their training.	✓		✓	✓			
An orientation at the beginning of the first semester should include understanding the assessments in their training, sitting in with the supervisor, learning about the structure and business of general practice etc.	✓			✓			

Context

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
Training organisation							
Excellence should be encouraged and not just a minimum standard. This should be part of the culture of the training organisation and the training of supervisors and MEs.	✓		✓	✓	✓		
Acknowledgement for registrars who are high performers, encouraging excellence	✓			✓	✓		
A culture of reflection and life-long learning should be fostered.	✓		✓	✓			
A collaborative culture between training organisations and assessors should be established to improve assessor buy-in.	✓	✓		✓			
Minimum standards should be enforced by the training organisation, but flexibility permitted to maximise the use of WBAs.	✓			✓		✓	
The training organisation should develop and communicate clear guidelines about WBA expectations.	✓	✓		✓			
Training organisations should develop and communicate clear guidelines for registrars, supervisors, practice managers and MEs to create a shared understanding of the WBA system.	✓	✓		✓			
Medical educators provide an essential backbone to any WBA system.	✓	✓		✓			

Theme	*Stream 1*	*Stream 2*	*Stream 3a*	*Stream 3b*	*Stream 3c*	*Stream 3d*	*Environmental scan*
There should be designated people within the organisation whose job it is to ensure that the admin runs smoothly, that registrars, supervisors and MEs are accountable and that the flagging process is followed up.		✓		✓			
Training organisation support for WBA assessors is critical.		✓		✓			
Training plans allow registrars to envisage their future training timeline and assessors to better understand future requirements		✓		✓			✓
The training organisation should support supervisors and registrars with enough time and remuneration to adequately perform WBAs.	✓			✓			
Practice Context							
An unfamiliar assessment space can impact performance.			✓	✓			
Room size may impact direct observation performance.			✓	✓			
Placement of registrar, observer and patient in the room impact performance.			✓	✓			
The practice environment can impact on case-load and case presentations.				✓			
The relationship between the registrar and the supervisor/ME impacts on the integration of feedback into learning.	✓		✓	✓			

Appendix C: List of recommendations for the Framework

WBA Tools

Recommended workplace-based assessment tools

- Safety assessment.
- Supervisor direct observation of the registrar.
- Direct observation of registrar by ME or External Clinical Teacher (ECT).
- Direct observation video reviews.
- Mid and end-term assessments.
- MSF (staff).
- A learning log.
- A procedural skills log.
- Random case analysis (RCA).
- PETAL (Patient Encounter Tracking and Learning).
- 'STAR' assessment (Statement of Awarded Responsibility).

WBA tool delivery

- Registrars should be clearly informed about each WBA and its purpose.
- It must be clearly defined whether the WBA is used for low, medium or high-stakes purposes.
- WBAs should be available through a user-friendly, efficient online learning system.
- Registrars should not choose patients/cases or assessors for WBAs.
- The training organisation should set up processes to ensure that specific, practical and timely feedback is given for each WBA.
- WBA feedback should be followed by an action plan.

WBA purpose

- Registrar safety is assessed using WBAs.
- Professionalism and communication skills can be taught and assessed using WBAs.

Mapping registrar progression

- WBAs should be spaced at regular intervals throughout training to map registrar progression.
- Entrustment scales are one of the best ways to assess the level of supervision required in various different areas and then to map the registrar's improvement and decreased need for supervision.
- A programmatic assessment approach with collation of outcomes and opinions maximises the effectiveness of WBA as tools for mapping progression.

WBA tool feature:

- EPAs should be used within WBAs.
- Benchmarking should be integrated into WBAs.
- Narrative is essential in WBAs.
- When a registrar is undertaking a placement in a culturally specific environment, the WBAs utilised should aim to gather opinions from those knowledgeable in cultural safety and competency.

Assessors: Supervisors, Medical Educators and External Clinical Teachers

Assessor roles

- Assessors must be aware of their various roles and responsibilities.
- Assessors may provide psychological support to the registrar and must be aware of available support and processes.
- Assessors are well equipped to support and flag registrars for support.

Multiple assessors

- Multiple assessors should be used to reduce bias.
- The same group of assessors should reassess the registrar at regular intervals.
- Co-ordination between assessors is essential.

Feedback

- Feedback is an essential component within WBA and appropriate systems and training must be in place to ensure quality feedback is provided.

Assessor Training

- The success of WBAs hinges on assessors having quality training and support.

High-stakes assessments

Flagging and remediation

- The training organisation should have a documented transparent process outlining how registrar flagging occurs.
- Consequences of non-compliance with WBAs, lack of insight, or unsafe practice should be clearly documented.
- Registrar flagging should take into consideration a collation of WBA feedback using a programmatic assessment approach.
- Flagged registrars who are deemed by an ME to require additional support require a tailored intervention plan that addresses the developmental gaps identified.
- Flagged registrars should be monitored and reviewed regularly and managed by a panel of senior MEs.
- Flags should be recorded in an online portfolio accessible to the registrar, assessors and training organisation.
- All stakeholders should be trained in the flagging process.

Consequences for non-completion of WBAs

- Non-compliance with flagging or remediation requirements should also have clear accountability pathways and consequences documented.

Programmatic assessment for high stakes decisions

- The final high-stakes programmatic assessment decision should be made by a panel of senior MEs.
- Ensuring that several assessors have affirmed that the registrar is ‘safe to practise unsupervised’ in all areas should be a prerequisite for completion of training.
- Supervisors should not be responsible for high-stakes assessments.
- High-stakes assessments for remediation should be collated in a programmatic manner with a variety of different WBAs used and a variety of different assessors.

Registrars

Training required prior to undertaking WBAs

- It is imperative that registrars have adequate in-practice orientation before undertaking WBAs.

Self-reflection and training in accepting feedback

- Registrars should be trained in self-reflection and accepting feedback.

Tools for self-reflection

- Tools such as PETAL and EPAs support registrar reflection and learning.
- A learning portfolio can guide reflection and learning.

Empowerment

- A procedural skills log will help to empower the registrar to seek more active supervision when it comes to undertaking procedures.

Context

Training Organisation

- A culture of excellence should be fostered within the training organisation.
- Medical educators provide an essential backbone to any WBA system.
- A medical educator should be assigned to each registrar for the whole of training to ensure assessment and training continuity.
- Training organisations need to ensure that registrars are aware of the mandatory expectations of training and that for independent and safe practice additional WBAs may be required.
- There is a need for training organisations to establish early on that training in general practice is not simply about passing exams, it's about life-long, reflective learning.
- Training organisations should develop and communicate a clear guideline for registrars, supervisors, practice managers and MEs to create a shared understanding of the WBA system.
- Training organisation support for WBA assessors is critical. Training coordinators are essential to provide support to assessors.
- Recognition/payment for assessor investment into WBAs is important.
- Training plans allow registrars to envisage their future training timeline and assessors to better understand future requirements.

Practice Context

- A positive relationship between the supervisor and registrar leads to more effective WBA outcomes.
- Guidelines for minimising the impact of context on the DOV should be adopted.
- Medical educators and other assessors should be aware that the context of the assessment can impact on the outcomes.