Judging Competency

A study of in-training evaluation of veterinary students

A thesis presented in partial fulfilment of the requirements

for the degree of

Doctor of Education

at Massey University, Manawatū, New Zealand

Elizabeth J Norman

2016

Abstract

In-training evaluations are a common but highly criticised method of assessing the competency of veterinary students completing training. They involve assessment of on-going performance in the workplace, performed by the supervisor. They are highly feasible and one of the few ways that a student's performance in an authentic context can be evaluated. Psychometric research has suggested, however, that in-training evaluations are unreliable, do not discriminate aspects of performance, and do not predict performance on other assessments, casting doubt on the credibility of scores. Research on rater judgement processes suggests, in contrast, that multiple aspects are discriminated and that accounting for context and inferred reasons for behaviour contributes to rater variability. Very little research has considered in-training evaluation in a veterinary context.

In a mixed method study this research investigated how well the in-training evaluation used during clinical placements in one veterinary school captured the aspects of student performance it was designed to capture. It explored the supervisor's view of student performance, and how that related to the dimensions being assessed in in-training evaluation, and to the constructs of competency articulated in frameworks. Complementary research strands involved analysis of semi-structured interviews with supervisors, common factor analysis of in-training evaluation scores, ordinal logistic regression relating factors to overall judgement, and thematic comparisons of findings with competency frameworks.

Together, the nature of what supervisors considered, the dimensional structure of scores, and the relationship of dimensions with the overall judgement suggested that the in-training evaluation is both holistic and discriminating, and that important aspects of performance are student engagement and trustworthiness. The aspects captured by the evaluation aligned well with the design of the instrument, and generally well with the veterinary competency frameworks. However, some areas were highlighted where concepts of veterinary competency and the competencies required in different subdisciplines need further consideration by the profession. The findings give insights into the process of judgement of competency by veterinary supervisors that will inform further research. They support some aspects of a validity argument in relation to scoring processes, and inform the design of evaluation instruments by underscoring the construct-relevance of interrelated dimensions.

Acknowledgements

Firstly, I would like to thank my supervisors, Dr Peter Rawlins and Dr Linda Leach. You have been a tremendous support and source of encouragement from the beginning. Thank you for being so available, for listening, and for your wise and gentle guidance. I have always enjoyed discussing things with you, and hope that may continue.

I would also like to thank those who have encouraged me on this research journey in one way or another. In particular, Professor Mark Brown, who first suggested I should do an EdD. You were so right! Thank you for your mentorship and support. Thanks also to Professor Tim Parkinson, Dr Marg Gilling, and Dr Dianne Gardner who encouraged me to research in education, and Dr Jenny Poskitt, who believed in me enough to let me move from science to the EdD programme.

A big note of thanks goes to all the interview participants who enthusiastically gave up their time and shared their thoughts with me. Without you, the research could not have happened, and I am tremendously grateful.

Thanks also to all those who helped me with the research in various ways: Andrew Rowatt who anonymised data for me so carefully; Simon Verschaffelt who set up computers for me; Sue Leathwick and Georgie Cowley, who gathered up data for me and answered my hundreds of questions; Professor Cord Heuer who helped me with SAS code while I was still getting my head around GLIMMIX. Thanks also to all those workmates who also helped with this by taking on work themselves or waiting patiently for things to be done. You all helped me have time to do this research.

Thank you also to Massey University for the financial support that accompanied the Vice Chancellor's Award for Sustained Commitment to Teaching Excellence that I received in 2012. This has covered all the costs of the research and enabled me share it at conferences.

To my walking buddies, breakfast buddies, and other friends: Naomi, Wendi, Jenny, Eloise, Kirsty, Mandy, Bridget, Johanna, Rose, Elizabeth, and Susan. You have all, at various times, listened with interest while I told you about my research. It has been so helpful to have people to tell, and I thank you for your friendship and support.

To my husband Richard, thank you for your unwavering love and support, and for listening and discussing things with me. To my children, Catie and Rob, thank you for your patience and understanding with a mother welded to a computer, and for still managing to growing up into such fine people. And especial thanks to Catie for your help with grammar when I was stuck.

Finally, I would like to dedicate this thesis to my late mothers, Marianne and Verity; I know you would both be proud.

Contents

Appendices

List of tables

List of tables in appendices

List of figures

List of boxes

Chapter 1:
Introduction

How do we know that a veterinary student has become a competent veterinarian? Upon graduation, young veterinarians transition from the wholly supervised environment of the university and clinical placements to a working environment that may be more or less supportive. There they will carry responsibility for at least some aspects of veterinary work from day one. Monitoring a student's progress towards achieving the necessary knowledge, skills, and attitudes for day one competency, and then certifying their achievement of it, is important to the student, so their learning is appropriately directed, but also important to ensure the safety of the public, the welfare of animals, the reputation of the university, and the standing of the profession.

However, veterinary competency is not easy to assess. It is complex, context dependent, and partly involves skills and attitudes that are not observable. It is also not well defined. Frequently definitions are locally created, and are supported more often by expert and stakeholder opinion than empirical research (Cake et al., 2016). Veterinary competency encompasses cognitive aspects of formal and tacit discipline knowledge, problem solving abilities, and clinical judgement; functional aspects of veterinary-specific technical skills and generic skills such as record keeping; and social aspects such as communication, interpersonal skills, and personal and professional behaviours and attitudes. Overarching these are metacompetencies that enable students to acquire and grow in competency and to mobilise and combine different dimensions of competency in a way appropriate for the context (Fernandez et al., 2012; Le Deist & Winterton, 2005).

To consolidate the competency of veterinary students and prepare them for independent practice, Massey University, like many veterinary schools worldwide, uses workplace-based training. Students spend their entire final year rotating through placements in different types of veterinary practices within and outwith the University. On placements, veterinary students work alongside veterinarians and within veterinary teams. Learning is informal, opportunistic, and authentic. Students see real veterinary work with real clients in real situations. They

encounter all the complexity of practice and thus have opportunity to develop their skill in applying and integrating aspects of competency as the demands of the situation require.

Utilising the workplace environment for assessment provides opportunities to assess what a student actually does in an authentic situation, which is thought a better indicator of future performance than methods that assess what a student knows, knows how to do, or shows how to do (Miller, 1990). Workplace-based assessments are therefore an important supplement to written, oral, and practical examinations, providing a view of competence-in-action not captured in many other types of assessment. Although there are several types of workplace-based assessment, the most commonly used in veterinary medicine in the United Kingdom, United States, Canada, and the Caribbean (Hardie, 2008) and in Australia (van Gelderen, 2015) is in-training evaluation.

In-training evaluation involves assessment of on-going performance in the workplace, performed by the supervisor, and allows evaluation of what a student does in real practice situations. They involve the supervisor rating a number of domains of performance using Likert-type items. In-training evaluation instruments are often locally developed and differ in detail, including the number of domains encompassed, the number of items for each domain, and the number of levels spanned by Likert-type items. It is common for there to be a separate rating given for overall performance and for this to form the student's recorded grade. Specific written feedback to accompany each domain is also encouraged on most instruments and increases the formative value. Evaluations are made at various points, such as at the end of each placement, as is the practice at Massey University, and therefore the ratings refer to performance involving a number of cases over a number of days or weeks.

In-training evaluations have advantages in simplicity of implementation. They utilise resources already committed for training, and do not involve an additional ethical cost of using animals only for assessment, as some more structured assessments do. In addition, extrapolation of performance from the assessment to future practice is more defensible because of the realistic nature of the evaluation setting and tasks. In contrast to standardised assessments such as simulations or practical examinations, students can demonstrate how they are able to integrate aspects of competency as demanded by the contextual complexity. Students can also demonstrate aspects of competency such as professionalism that might be hard to

demonstrate in standardised environments (Prescott, Norcini, McKinlay, & Rennie, 2002), and may not be possible to assess without qualitative judgement (van der Vleuten & Schuwirth, 2005). Govaerts, van der Vleuten, Schuwirth, and Muijtjens (2007) conclude that in-training evaluations are a valuable tool for assessment of clinical competency. However, despite their advantages and wide use, in-training evaluations are heavily criticised and they are thought not suitable for certification decisions (Miller, 1990; Swing, 2002; Turnbull & van Barneveld, 2002). Critics point to their poor score reliability, subjectivity, limited sampling, and the lack of discrimination between dimensions assessed and between students. Thus, the validity of inferences based on in-training evaluation is questioned (Lurie, Mooney, & Lyness, 2009).

The rater is seen as the source of these problems, leading to distrust of this and other workplace-based assessments. While much research in medical education has focussed on psychometric issues and rater errors, some have argued for a fundamental shift in thinking. Instead of a psychometric, quantitative standpoint that views discrepancies as measurement error, they advocate that we should value unique perspectives on a qualitative performance that it is simply not appropriate to quantify (Govaerts et al., 2007) and look for ways to capture this richness (G. Regehr et al., 2012). There have been continuing calls for qualitative research that furthers our understanding of the process of judgement in complex situations and how an overall evaluation is made (Bogo, Regehr, Hughes, Power, & Gioberman, 2002; G. Regehr, Eva, Ginsburg, Halwani, & Sidhu, 2011; G. Regehr et al., 2012; van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010; Wood, 2014).

Much progress on rater judgement has been made by researchers in medical education and other disciplines, but research on assessment in veterinary medicine is scarce (Rhind, Baillie, Brown, Hammick, & Dozier, 2008). The issues, outlined above, of score unreliability and lack of discrimination of in-training evaluations are suspected to be present, but have had limited investigation. The different context of veterinary medicine from human medicine may contribute to these issues in unanticipated ways. This means that the same problems cannot be assumed, but also means that studying in-training evaluations in another context may further illuminate aspects of rater judgement and contribute to what is already known. Although research in the veterinary context is so limited that there are many unanswered questions, of particular interest is the relationship between the scores awarded by supervisors and the construct of competency we are trying to assess. If competency does not explain the scores awarded, then any inferences we draw from scores about the student's competency are

invalid. We need to know what supervisors pay attention to in forming their judgement, how this corresponds with what they are asked to assess, and how well their judgment is captured by the scores awarded for a particular student. The aim of this research then was to investigate these aspects of the supervisor's judgement of the competency of students in a veterinary context. The overarching question guiding the research was: How well do the scores on in-training evaluation capture the aspects of student performance we intend to assess?

In particular, the research addresses the following questions:

1. What qualities of performance do supervisors value when making judgements about the competency of veterinary students on placements?
2. What is the nature of the dimensions captured by the items on in-training evaluation?
3. How do the dimensions captured relate to the overall grade?
4. What is the relationship between what supervisors value, the dimensions captured by the evaluation, and the competency frameworks?

A convergent mixed-method approach, with four phases, enabled me to examine layers of complexity and to develop an in-depth understanding from a critical realism perspective. In interviews, I gathered in-depth information about the aspects of student performance that are important to supervisors. Using exploratory common factor analysis, I examined the dimensions considered by supervisors during evaluation. I used ordinal logistic regression with generalised linear mixed modelling to examine the contribution of the factors and other variables to the overall grade supervisors awarded. A thematic analysis then enabled me to compare the factors assessed in the in-training evaluation with the aspects of student performance that supervisors value and with the social conceptualisation of veterinary competency as articulated in competency frameworks.

Structure of the thesis

The thesis is made up of nine chapters including this one. In Chapter 2, I review the veterinary competency frameworks in use in New Zealand and the research on veterinary competency, to provide background on what the in-training evaluation is aiming to assess. I then discuss the

research on the problems with in-training evaluation, drawing from the literature in medical education where much more work has been done than in veterinary education. A conceptual framework of the process of rater judgement is used to organise the literature. The chapter ends with discussion of the limited research on in-training evaluation in veterinary education.

Chapter 3 concerns the research design. I present the setting for the research, the research stance, and discuss threats to the validity of the research conclusions and how I have minimised these. The details of methods used have not been presented in the research design chapter, but instead are presented together with the findings in subsequent chapters, because they are integral to understanding the findings.

Chapters 4, 5, 6, and 7 present the methods and findings of each of the four phases of research. The first phase (Chapter 4) focusses on the interviews with supervisors and the themes I interpreted about the aspects of student performance that are important to them. Chapter 5 presents the exploratory common factor analysis and the dimensional structure apparent in the in-training evaluation. Its findings are also interpreted in this chapter as they inform the methods of the next quantitative phase. Chapter 6 presents the ordinal logistic regression and the findings of the relationship of factors and other variables to the overall grade. Its findings are also interpreted. Chapter 7 presents Phase 4 of the research that links the findings from the first three phases together with the competency frameworks using thematic analysis, and presents my interpretation.

In Chapter 8, I provide a general discussion that links all the phases of research. Chapter 9 draws conclusions, reflects on the value of the research, and discusses their implications, making suggestions for further research.

Terminology

Before moving on, it may be helpful to provide some clarification of terminology in use in the research literature and throughout the thesis.

In-training evaluations

In-training evaluations are called by a variety of names throughout the research literature. While in-training evaluation report, abbreviated to ITER, is reasonably common, other terms abound including in-training assessment, global rating form, clinical evaluation form, supervisor report form, and ward evaluation. Common usage at Massey University names the evaluation after the software used to record the result. I settled on *in-training evaluation* as close to a common term, and to focus on the evaluation as the object of study rather than the form itself. Although the term *evaluation* is often used to describe programme evaluation rather than individual student assessment, in this thesis I use the terms *evaluation* and *assessment* synonymously.

Student, supervisor, and rater

Medical education involves various levels of undergraduate and postgraduate training and each of these levels is associated with different terminology for the trainee doctors, such as clerks and residents. To avoid confusion I have chosen to refer to all of them as *students*, even though some may indeed be quite senior, well-qualified, postgraduate students. No disrespect is intended. Likewise, supervisors of students on placements may, in medical education, be more senior students themselves. For example, residents may be supervisors of clerks but supervised themselves by attendings. Because this terminology is confusing for those like me who do not work in the system, I refer to those performing the evaluation in a clinical situation as *supervisors*. I also use another term, *rater*, somewhat interchangeably when discussing the rating process, especially when the situation is experimental, and the evaluation is not made in a supervisory capacity.

Competency

The words *competency* and *competence* are often used interchangeably (ten Cate & Scheele, 2007), and are defined the same way in general usage ("competence", 2012), however some authors distinguish them. Khan and Ramachandran (2012) have proposed that the word competence be used to describe the attribute and ability of the person and competency as the skill itself. It is difficult to see the difference as ability involves skills, and skills cannot be

divorced from the person and their capability. Winterton (2009) notes that competence is most commonly used to indicate "what a person needs to know and be able to do" (p. 684) and competency to indicate the "characteristics of an individual that are associated with superior performance" (p. 684). Sultana (2009) noted that some usages distinguish the two along the lines of behaviour (competency) and outcome (competence) and that others see competencies as components of competence. Moore, Cheng, and Dainty (2002) give further examples. I find these distinctions subtle and confusing and therefore use the words interchangeably throughout this thesis.

Discipline-specific and non-discipline-specific aspects of competency

As the next chapter will discuss, competency is a concept that has a number of aspects including knowledge, skills, and attitudes. Some of these are particular to a discipline such as veterinary medicine, for example diagnosis of animal disease, or surgical procedures on animals. Other aspects of competency are not specific to veterinary medicine and are important in many or all disciplines, for example, communication skills, honesty, and problem-solving skills. These two groups of aspects often need to be distinguished in discussions of curricula and assessment and various terms have been used, in particular for the non-discipline-specific aspects. One term in common use for non-discipline-specific aspects is *nontechnical* but this has been criticised as implying skills and attributes that are of less value and requiring less training than discipline-specific technical skills (Nestel, Walker, Simon, Aggarwal, & Andreatta, 2011). Another term in common use for non-discipline-specific aspects is *professional* (Cake et al., 2016; Hodgson, Pelzer, & Inzana, 2013; North American Veterinary Medical Education Consortium, 2011) but use in this context is confusing given that it is the discipline-specific aspects that distinguish one profession from another, not the non-discipline-specific aspects. The term *professional* can also be used to indicate the holistic interplay of both discipline-specific and non-discipline-specific aspects of competency, and sometimes is used in that sense in the same document as it being used in a sense confined to non-discipline-specific aspects (North American Veterinary Medical Education Consortium, 2011). This means that a reader could only distinguish whether discipline-specific, non-discipline-specific, or both aspects of competency were being referred to by the context. Other terms that have been used to indicate non-discipline-specific aspects include *soft skills* (Walker, Roberts, & Mehlhorn, 2015) which seems derogatory, and *noncognitive skills* (Lane & Bogue, 2010) which is inaccurate. Nestel et al. (2011) propose the term *human factors* be used, however this

seems equally problematic as it implies some sort of non-human aspect to discipline-specific knowledge and skills. The dictionary definition of nontechnical is straightforward—"not requiring or assuming specialized or technical knowledge" ("non-technical", 2012)—and conveys the main sense of the word and is a little less unwieldy than the other accurate alternative of non-discipline-specific. For this reason, I will use the term *nontechnical* to indicate non-discipline-specific aspects of competency throughout this thesis.

Workplace-based assessment

Workplace-based assessment is a term used to describe the assessment of performance in the authentic context of the workplace (Hamdy, 2009). It is often used as a general term describing a variety of methods of assessment that use ratings, including in-training evaluation (Campbell, Crebbin, Hickey, Stokes, & Watters, 2014; Eva et al., 2015; Govaerts & van der Vleuten, 2013; McGill, van der Vleuten, & Clarke, 2011; Swanwick & Chana, 2005 ; van der Vleuten et al., 2010; Weijs, Coe, & Hecker, 2016). However, some authors do not include in-training evaluation as a type of workplace-based assessment (for example, Gingerich, Kogan, Yeates, Govaerts, & Holmboe, 2014; Hecker, Norris, & Coe, 2012; Massie & Ali, 2016). These authors focus instead on more recently introduced methods of assessing performance in the workplace such as the mini-clinical evaluation exercise (miniCEX), direct observation of procedural skills (DOPS), and multisource feedback. This reflects a change in practice, with newer methods now being more frequently used. They have replaced in-training evaluation in some medical and veterinary training programmes, but in other programmes, newer methods are incorporated alongside in-training evaluation. In this thesis, I use the term workplace-based assessment as a general term to include all forms of assessment performed in the workplace, including in-training evaluation.

Summary

This chapter has introduced the importance of assessing veterinary competency well and highlighted the need for veterinary-specific research on in-training evaluations to improve our understanding of rater judgement in this context. Four research questions were identified focussing on what supervisors value, how that is captured in the scores, and how it corresponds to what the evaluation is intended to assess. Some key terminology has also been

discussed. In the next chapter, veterinary competency and the use of in-training evaluations for its assessment will be reviewed in more detail.

Chapter 2:
Literature review

Reviewing veterinary competency and in-training evaluation

As a starting point for investigating the process of judging the competency of veterinary students in the workplace, this review first considers what veterinary competency is and what the implications of that are for assessment. Next, I review what is known about the accuracy of the in-training evaluation and the process of evaluation. There is a large literature on in-training evaluation and clinical judgement in medical education that I draw on, as well as some of the literature on rater-based judgement in job performance assessment. Lastly, I review the limited literature on in-training evaluation in veterinary education.

Veterinary competency

The importance of veterinary competency

The veterinary profession has seen a changed emphasis on competency that is in keeping with global workforce changes. While there are multiple drivers for the change, a key one for the health professions globally has been accountability. This emphasis followed significant changes in public attitude because of some high profile medical negligence cases (Cruess & Cruess, 2005). No longer was attaining entry to a profession sufficient to ensure public trust in the ongoing competency of a professional for life. Continuing competency needed to be demonstrated. In New Zealand, revisions to the Medical Practitioners Act (1995) included requirements for practitioners to demonstrate on-going competency, and a mandate for colleagues to report incompetence. New legislation for other health professionals followed in 2003, and in 2005, the Veterinarians Act was also revised to incorporate competency provisions. The new Veterinarians Act (2005) reflected the need to ensure veterinarians were competent, not just qualified.

Contemporaneously there were increasing international concerns about veterinary education. A series of commissioned reports in the United States highlighted the fact that veterinary education, though it provided excellent technical knowledge and skills, was not delivering important outcomes in other areas such as communication and business management (J. P. Brown & Silverman, 1999; Cron, Slocum, Goodnight, & Volk, 2000; Pritchard, 1989; Volk et al., 2005). The findings of veterinary employer consultation in other nations, including New Zealand, indicated that this was a worldwide issue (Allan & Parkinson, 2010; Coleman, Salter, & Thornton, 2000; Faculty of Veterinary Science University of Sydney, 2004; Heath & Mills, 1999). Universities and professional groups were galvanised to reconsider the competencies required for success in veterinary practice and develop competency frameworks spanning both technical and nontechnical areas.

A competency framework is an agreed statement articulating what it is to be competent in the discipline. Two competency frameworks are important for New Zealand veterinarians and veterinary students. Minimum practicing standards are prescribed by the Veterinary Council of New Zealand (hereafter referred to as the VCNZ standards). In defining the required competencies for veterinarians in New Zealand, the Veterinary Council has drawn from the current graduating competencies of students from New Zealand's only veterinary school at Massey University (Veterinary Council of New Zealand, 2012). These graduating competencies form the learning outcomes for Bachelor of Veterinary Science students and are hereafter referred to as the BVSc learning outcomes. In turn, Massey University's graduating competencies draw heavily from the veterinary competency statements prescribed in jurisdictions of Australia, United Kingdom, United States, and Canada (Massey University, 2012). A series of overarching accreditation frameworks and inspection processes ensures a common minimum standard of teaching in veterinary schools across New Zealand, Australia, United Kingdom, United States, Canada, and some elsewhere (Craven, 2004).

Competency is thus of central importance to both the education of veterinarians in New Zealand and their initial and continued certification to practice. Knowing if a veterinarian is competent depends entirely on knowing what competency is and being able to assess it, however, as this review will show, competency is a debated concept and is not easy to assess.

Defining competency

Competency is a multidimensional concept that includes the cognitive, functional, and social skills required for success in an occupation (Le Deist & Winterton, 2005). There is a sense of sufficiency and capacity ("competence", 2012) and etymologically, not only capacity, but also licence or permission (Mulder, Weigel, & Collins, 2007). An important overarching metacompetency is the ability to mobilise and combine different component dimensions of competency in a way appropriate for the context (Fernandez et al., 2012). This means that the aspects of competency that are brought to bear are different for different tasks and situations. Fernandez et al. (2012) noted differences in the way the combining of these elements is perceived. Some authors have focused on the selection of appropriate competency components to apply to a given complex situation, whereas others have emphasised the synergistic combination of components to form a competence that is more than the sum of its parts. Le Deist and Winterton (2005) emphasise that, although it is multidimensional, competency is a holistic and unitary concept.

The Veterinary Council of New Zealand defines a competent veterinarian as one who “applies knowledge, skills, attitudes, communication and judgement to the delivery of appropriate veterinary services in accordance with their field of veterinary practice” (Veterinary Council of New Zealand, n.d., p. 1). Consideration is given to performance of tasks to an acceptable standard on a consistent basis. The VCNZ standards contain 8 standards which refer to (1) veterinary knowledge and its application; (2) investigation and record keeping; (3) diagnosis and treatment planning; (4) knowledge of when and how to refer; (5) treating animals and performing veterinary procedures; (6) using skills to promote animal welfare, health and productivity, and to protect public health and biosecurity; (7) effective communication; and (8) professional, ethical and legal practice. Inherent in these standards is a complex mixture of different aspects of performance and it is recognised that the VCNZ standards do not necessarily encompass all that competency means (Veterinary Council of New Zealand, n.d.).

Competency frameworks for veterinary science in mutually accredited schools tend to be broadly similar but often differ in detail. Recently Cake et al. (2016) systematically reviewed published frameworks and organised the nontechnical components into a taxonomy based on the taxonomic framework produced by Englander et al. (2013) for health sciences. An adapted version that is expanded to include discipline-specific competencies (not included by Cake et

al. (2016) as their focus was on nontechnical aspects) is presented in Table 2.1. The taxonomy provides an aggregate of all the aspects considered in all the frameworks, but not every aspect is mentioned in each individual framework, with variation tending to be in the nontechnical aspects rather than the discipline-specific knowledge and technical aspects. Cake et al. (2016) found that communication skills and professional behaviour were included in all frameworks and that written communication, collaboration and teamwork, and business and practice management were commonly included. Emotional intelligence and self-awareness, and self-efficacy and confidence were infrequently included. The VCNZ standards make mention of all aspects of the taxonomy except financial awareness, business and practice management, and also do not mention any aspects of the personal development domain shown in Table 2.1.

This variation in what is included in frameworks is not surprising since a diversity of opinion on the components of competency has also been noted in medicine (Fernandez et al., 2012). Partly, this might reflect the difficulty of expressing, in words, the abstract concepts that competency comprises. It also likely indicates differences in how veterinary competency is conceived by different individuals and groups, even though the global networks of mutual recognition in veterinary medicine impose a high degree of similarity. The Veterinary Schools Accreditation Advisory Committee (VSAAC), for example, sees the need for individual institutions to incorporate their own aspirations for distinct graduate attributes and therefore has avoided overly prescriptive outcomes (Craven, 2004). Bok et al. (2014) compared international perceptions of the importance of various domains of competency and found consensus about the broad domains but regional differences in their relative importance, in particular those of veterinary expertise, entrepreneurship, and scholarship. There is much potential for differences to arise because of differences in context, differences in opinions on the abstract aspects of competencies, and impreciseness of language which can lead to varied descriptions of the same thing. To some extent, also, the differences reflect an on-going conversation about the changing roles of veterinarians in today's world and the dilemmas of society's expectations, such as whether savvy business practice is compatible with altruistic care of animals (Theis, 2003), and whether the primary obligation is to animals or people (Willis et al., 2007). The way competencies are derived will also influence the descriptions produced because the results of empirical investigations tend to produce different descriptions of competency than expert analysis of the components of job expectations (Norris, 1991).

Table 2.1: Taxonomy of veterinary competencies.

(1) Patient care
- Veterinary-specific skills to diagnose and treat patients and prevent disease
- Workflow management
- Effective communication with clients
- Relationship-centred care

(2) Knowledge for practice
- Veterinary-specific knowledge
- Critical thinking and problem-solving
- Research skills and practice

(3) Practice-based learning and improvement
- Awareness of limitations
- Lifelong learning
- Information literacy and evidence-based approach
- Information technology
- Reflection and goal-setting
- Educating others

(4) Interpersonal and communication skills
- Verbal communication
- Empathy and bond recognition
- Emotional intelligence and self-awareness
- Written communication and records

(5) Professionalism
- Professional values
- Professional behaviour
- Commitment to animal welfare
- Cultural sensitivity and diversity

(6) Systems-based practice
- Health and welfare advocacy
- Financial awareness
- Business and practice management

(7) Interprofessional collaboration
- Effective communication with colleagues
- Collaboration and teamwork

(8) Personal development[a]
- Resilience
- Work-life balance
- Adaptability
- Self-efficacy and confidence
- Leadership

Note. Derived from Englander et al. (2013) and Cake et al. (2016).

[a]In the original (Englander et al., 2013), domain 8 is named *personal and professional development*, but has been altered for clarity as, in the sense the term is usually used, *professional development* is present in domain 3 under lifelong learning.

Lurie, Mooney, and Lyness (2011) have been critical that competency frameworks are socially derived statements that, in attempting to articulate shared values, must necessarily be a compromise of individual views, and for which relationships with observable behaviours have

not been empirically established. Certainly the methods reported for deriving competency frameworks in veterinary medicine are commonly opinion-based and include surveys or interviews of practitioners, faculty, students, and recent graduates; consensus groups of experts; practitioner focus groups; and large scale stakeholder group dialogue followed by expert consensus (Cake et al., 2016). A Delphi expert consensus procedure has also been used (Bok, Jaarsma, Teunissen, van der Vleuten, & van Beukelen, 2011). In their systematic review Cake et al. (2016) found only a small number of studies that provided empirical support for relationships of aspects of competency with outcomes such as client satisfaction, client compliance, employer satisfaction of veterinarian wellbeing, and not all provided high quality evidence. Other empirical methods to derive competency frameworks such as patient outcomes, task analysis using on-the-job observations, critical-incident surveys, or other investigations of the differences between expert and non-expert performers have not been reported.

The importance of empirical research to establish competencies required was demonstrated by Cake, Rhind, and Baillie (2014) who examined the evidence for inclusion of business skills in competency frameworks. They found that empirical evidence of associations with income, employability, and employer satisfaction, provided ample evidence supporting its inclusion in competency frameworks. Yet business skills were perceived to be relatively less important than other competencies when Likert scale survey methods were used to survey students, new graduates, more experienced veterinarians, and clients. Other nontechnical aspects for which there was empirical support based on outcome measures were communication skills, empathy, relationship-centred care, self-efficacy, confidence, optimism, and reflective practice (Cake et al., 2016). Amongst these are competencies that are infrequently included in competency frameworks. Based on their meta-analysis of survey data (Cake et al., 2016) found that awareness of limitations, professional values, critical thinking, collaboration, and resilience were other nontechnical aspects of perceived importance, and that information technology, leadership, health and welfare advocacy, cultural competency, and research had little perceived importance for veterinary graduates. Bok et al. (2011) found that two aspects of competency did not reach consensus for inclusion using the expert Delphi procedure. These were ability to conduct scientific research and ability to educate and teach colleagues, co-workers, and students using sound pedagogical principles. However, whether these perceptions of importance reflect the actual importance in practice has not been investigated.

An important study that was not examined by Cake et al. (2016) applied a constructivist grounded theory methodology to formulate a definition of veterinary professionalism (Mossop, 2012). Semi-structured in-depth individual and focus group interviews with theoretical sampling were used to collect data from veterinarians, veterinary nurses, and client owners of a range of species, as well as representatives from the governing professional association in the UK, where the study was conducted. Participants were asked to describe aspects of a "good" and "bad" veterinarian and what they understood veterinary professionalism to mean. From her analysis, Mossop (2012) derived a model for understanding veterinary professionalism in which a central component was the concept of balance, facilitated by a series of attributes. The attributes arising from the data were altruism, general attitude and manners, caring and being empathetic, honesty and trust, core personal values, personal efficiency, maintaining technical competency, communication skills, decision making and problem solving, autonomy and self-regulation, and confidence and knowing limits. Mossop (2012) noted the absence of reflective practice arising as an attribute from the data, but considered that this was an underlying and unrecognised enabling attribute of several of the other attributes. She also commented on the lack of leadership as an identified attribute and proposed it may be an underdeveloped, and therefore little recognised attribute. The key linking theme of the framework was that vets must use these attributes to balance the interests of the animal, the client, the practice (or business), and the wider society. Thus achieving some sort of balance was the core art of professionalism and the attributes were the facilitators of that. Inherent in the concept of balance is that the aspects being balanced are connected and therefore management of one requires attention to how the others are impacted in order to best maintain balance. Thus, this conceptual model is consistent with the idea of competency as a multidimensional, holistic, and unitary concept.

A point to note about the taxonomic framework (Table 2.1) is that there is great deal of overlap and interconnectedness between aspects in each domain. For example communication, which has its own domain, also appears under patient care and interprofessional collaboration. Furthermore, it may be argued that health and welfare advocacy, which is part of the systems based practice domain, also requires effective communication in what may be challenging circumstances, and so does educating others, which is part of the practice-based learning and improvement domain, and cultural sensitivity, which is part of the professionalism domain. This demonstrates the point that competency

requires integration of components and that component competencies cannot be thought of as separable and independent.

Nor can the importance of any aspect of competency be considered in the absence of context. As reviewed by Gonczi and Hager (2010), the context can influence the attributes required to such an extent that the same occupation can require different competencies in different workplaces. Variable opinions of the competencies required in an occupation are related to job complexity, contextual factors such as the interdependence with others in a team, and the specific activities undertaken (Lievens, Sanchez, Bartram, & Brown, 2010). There can also be so much overlap in the nontechnical competencies required that occupations cannot be distinguished by looking at a list of competencies out of context (Norris, 1991). In addition, while the legal definition of competency is that held by an individual, there is recognition that competency is also held in the work of teams, and that an individual performance may be affected positively or negatively by the performance of the team (Lingard, 2012). As argued by ten Cate, Snell, and Carraccio (2010), competency varies with changes in context and should be described in terms of the interplay between the learner and the environment. It is the view of Govaerts (2008) that the key importance of context means that competencies are not generalisable knowledge, skills, and attitudes and must be understood and assessed in connection with their context.

Another reason that competency frameworks may differ from each other is a variation in the level of performance or career stage described. Competency usually is taken to mean the minimum standard and does not define the excellence that should be aimed for or the expertise that can be developed (Brooks, 2009). Carraccio, Benson, Nixon, and Derstine (2008) placed competency on a developmental continuum between advanced beginner and proficient, with expert and master above that. Yet some competency statements are phrased in terms that suggest a later career stage, for example “veterinarians are proactive leaders in the profession and are recognized voices of authority in important areas, such as animal welfare and One Health medicine” (North American Veterinary Medical Education Consortium, 2011, p. 6). In other competency frameworks, such as the BVSc learning outcomes at Massey University, a subset of competencies that are expected at day one after graduation are specified, divided into those for which it is expected graduates can work unsupervised and those that it is expected they would need support and supervision as they further develop expertise. These were developed in consultation with the profession about expectations at

graduation (Massey University, 2012) and draw on previous draft versions of the day one competencies developed by the Royal College of Veterinary Surgeons, UK (Royal College of Veterinary Surgeons, 2014), which have been widely adopted worldwide.

Thus, it can be seen from this discussion that there is no one defining competency framework for veterinary medicine. Frameworks differ in what they contain and the relative importance placed on aspects, as well as the way aspects are linked together. Because there is little empirical support for what to include or not to include, competency frameworks are mainly based on opinion. As such, they are socially constructed and interpreted but this is entirely appropriate since they represent the profession's contract with society, which requires negotiation from both sides (T. J. Wilkinson, Moore, & Flynn, 2012). They should not be seen as static and fixed, but as reflecting the social contexts in which they are created and need to be adapted as social contexts change (Whitehead, Selleger, van de Kreeke, & Hodges, 2014). In addition, there is a potential advantage in having multiple descriptions, metaphors, and images of competency to help capture its complex nature (Whitehead et al., 2014) and different formats may be useful for different purposes (Vandeweerd et al., 2014).

Evaluating competency in the context of the workplace

As discussed, a focus on competency as the outcome of education is currently the norm in medical and veterinary education. Workplace-based learning is an integral part of competency-based curricula. It is mandated as part of veterinary school accreditation requirements and is managed similarly in many schools. At Massey University, the entire final year of the degree is spent in clinical training through placement in veterinary practices within and outwith the University. On placements, students work alongside veterinarians and within veterinary teams. This consolidates the formal didactic and practical training of the earlier years, with practice in real clinical contexts, and allows the development of holistic applied skills. Learning is informal and opportunistic (Magnier et al., 2011). Ongoing formative feedback is essential for supporting development of day one competency in this final year of training. In addition, since upon graduation veterinarians in this country are immediately able to register and practice with no further formal assessment, the University has an important role in certifying competency of its graduates. Workplace-based assessments are an important part of the assessment of overall competency because they assess students in-situ as they

work in authentic environments. At Massey University, therefore, a combination of assessments is used in the final year of study including formal written, oral, computer simulated, and practical examinations, assignments, and in-training evaluation.

In-training evaluation

In-training evaluation is a traditional and commonly used workplace-based assessment in both medical and veterinary training. A survey of assessment methods used in veterinary schools in the United States, United Kingdom, Canada, and the Caribbean found that in-training evaluation was the most commonly used, although all schools used multiple methods (Hardie, 2008). Van Gelderen (2015) recently reported its use in every Australian veterinary school, and recent publications would suggest that it is still in common use today in other parts of the world.

There is a wide variety of practice in the use of in-training evaluations by medical and veterinary colleges within and between subdisciplines, with procedures and instruments often being created or adapted to suit local conditions. While all in-training evaluations involve the supervisor rating student performance in the workplace over a period of time, there is variation in the instruments used, the frequency of evaluation, and the contribution of the evaluation to summative assessment. Ratings are performed with Likert-type items that specify achievement levels within domains of performance, and additional descriptive formative feedback is encouraged. In-training evaluations may reflect an individual supervisor's rating, but are frequently a summary of the input of several supervising staff (Walsh, Zeck, Wall, Smith, & Wilson, 2012; Weijs et al., 2016).

In-training evaluation instruments differ in the number of domains assessed as well as the number of items for each domain. Some instruments contain several items for each domain of performance, such as the veterinary instruments reported by Fuentealba and Hecker (2008), Root Kustritz, Molgaard, and Rendahl (2011), and Weijs et al. (2016) which contained 21-29 items designed to assess 3-5 domains of performance. Others contain only one item for each of 3-9 domains of performance, in which case the item is usually conceptually more complex (Bateman et al., 2008; Matthew, Taylor, & Ellis, 2010; Roush, Rush, White, & Wilkerson, 2014; Walsh et al., 2012). Instruments also vary in the number of levels of performance represented by the Likert-type items, and in how much description of domains, items, and levels is provided

to help supervisors. Performance levels are specified in different ways in reported veterinary instruments, for example in terms of excellence (outstanding, good, unsatisfactory) (Walsh et al., 2012), frequency of behaviour (most of the time, rarely) (Fuentealba & Hecker, 2008), or in terms of meeting expectations (below expectations, exceeds expectations) (Roush et al., 2014). Instruments used in medical education have sometimes used normative levels of performance (below peer level, above peer level) or those related to independence (requires frequent assistance, performs independently) (Baker, 2011). At some veterinary schools, separate in-training evaluation instruments have been developed for use by different veterinary subdisciplines (for example, Walsh et al., 2012) and others use a common instrument (for example, Fuentealba & Hecker, 2008; Roush et al., 2014).

While it is common in veterinary education to assess students approximately every 2 weeks (Fuentealba & Hecker, 2008; Roush et al., 2014; Walsh et al., 2012), in-training evaluations are sometimes performed far less frequently in medical education, where they may be used as an annual or 6-monthly summary of performance (Borman, Augustine, Leibrandt, Pezzi, & Kukora, 2013; Durning et al., 2010; Woloschuk, McLaughlin, & Wright, 2013).

Practice also differs in whether and how the in-training evaluation contributes to final grades. In veterinary education, a range of practice is reported, with some in-training evaluations used only formatively (Weijs et al., 2016), some contributing to grades with other assessments (as at Massey University), and some used as the only assessment of students in clinical years (Fuentealba, Mason, & Johnston, 2008). An assessment grade is derived from the evaluation in different ways. Some in-training evaluation instruments require supervisors to make an overall rating that is separate to domain items (Matthew et al., 2010; Walsh et al., 2012). Other practice involves averaging items for which levels have been assigned a prespecified numerical value (Fuentealba & Hecker, 2008). There may also be a requirement for at least adequate performance, with students required to repeat placements if they receive an unsatisfactory grade (for example, Walsh et al., 2012).

Evaluations are frequently captured using web-based systems (for example, Fuentealba & Hecker, 2008; Roush et al., 2014; Walsh et al., 2012). Some reports describe rater-training provided for supervisors (Fuentealba & Hecker, 2008), but in other cases raters are likely to be

untrained, as students can choose a placement location and supervisor training is not a pre-requisite for placements to be considered suitable for students (Fuentealba et al., 2008).

Other workplace-based assessment instruments

There is increasing use of other workplace-based assessment instruments such as the mini-clinical evaluation exercise (miniCEX), direct observation of procedural skills (DOPS), and multisource feedback (Bok et al., 2013; Hecker et al., 2012). A brief explanation of these tools will facilitate discussion as they share some features and problems with in-training evaluations in that all are rater-based assessments, and therefore much research on rater-based judgement is applicable to all.

The miniCEX involves the supervisor observing the student during an encounter with a real patient, usually a consultation, during which the student explains their diagnosis and treatment plan to the supervisor. The supervisor gives immediate verbal and then written feedback including a performance rating which may comprise several items involving different domains of competency (Norcini, 1995). These are short assessments (approximately 20 minutes) and designed to be undertaken repeatedly during training in different workplace contexts. An advantage of the miniCEX over the in-training evaluation is that it ensures direct observation by the supervisor and recording of the evaluation is immediate, so it can be used formatively. Also, because it is short, it is more feasible to conduct the multiple assessments needed to provide sufficient score reliability.

The miniCEX is the most well studied direct observational tool in medical education (Kogan, Holmboe, & Hauer, 2009), however there are few studies on response processes (the cognitive processes of judges), internal structure (the dimensions being assessed), or how well performance relates to future performance (Hawkins, Margolis, Durning, & Norcini, 2010; Sandilands & Zumbo, 2014). Problems with rater leniency, high item intercorrelation, and low dimensionality have been highlighted (Hawkins et al., 2010). These are also areas of concern with in-training evaluations. Direct observation of procedural skills (DOPS) is performed similarly to the miniCEX but the student performs a procedure of some sort, rather than a consultation. It has been even less well studied than the miniCEX (Jelovsek, Kow, & Diwadkar, 2013). The miniCEX and DOPS are two examples of a host of direct observational assessment instruments that have been developed (Kogan et al., 2009).

Multisource feedback involves a series of evaluations from different types of raters, including patients, nursing staff, colleagues, and supervisors, and may also include self-assessment. Each rater assesses the aspects of performance that they see, and together this builds a picture of competency based on a variety of perspectives that may be more comprehensive than that of one rater. Multisource feedback has been widely used and extensively researched in business contexts (Campion, Campion, & Campion, 2015) and also been widely researched in medical education, where it is often used to assess practicing physicians for documentation of ongoing competency. Multisource feedback involves the use of similar rating scales to in-training evaluations and suffers from similar problems with interrater variability, especially between different groups of raters. However, with sufficiently large sampling good score reliability is obtained (for example eight physicians, eight co-workers, and 25 patients) (Donnon, Al Ansari, Al Alawi, & Violato, 2014) and even these large samples may be reasonably feasible. The number of dimensions captured by multisource feedback instruments varies widely between reports, and between the rater source group, with reports ranging from one dimension to seven in a series of studies reviewed by Donnon et al. (2014).

This brief review of other workplace-based assessments shows that the in-training evaluation is not alone in having problems with score reliability and dimensionality. Both multiple sampling and incorporation of a range of tools are therefore recommended for assessing competency (Epstein, 2007; Hecker et al., 2012; Hodgson et al., 2013; Magnier, Dale, & Pead, 2012; van der Vleuten & Schuwirth, 2005). Multiple assessments can be combined and evaluated in a portfolio (Bok et al., 2013) with achievement of good composite score reliability (Moonen-van Loon, Overeem, Donkers, van der Vleuten, & Driessen, 2013).

Rating scales for workplace-based assessment instruments

Competency goals tend to be “by definition lofty, vague, and far-reaching” (Carraccio, Wolfsthal, Englander, Ferentz, & Martin, 2002, p. 362) and therefore difficult to assess. In the behaviourist tradition they are often broken down into smaller discrete steps in an effort to allow achievement to be demonstrated and assessed “objectively” (Huddle & Heudebert, 2007). Such atomisation has been strongly criticised for its “exhaustive detail, leading to bulky, fragmented documents that lose practical value for education as they become less and less connected with the real world” (ten Cate et al., 2010, p. 671). The specification of competencies as lists of discrete tasks means that relational and complex aspects are lost (Govaerts, 2008). The holistic nature which integrates attributes such as cognitive skills

(knowledge, critical thinking, and problem-solving strategies), interpersonal skills, affective attributes, and technical/psychomotor skills required for competent performance is not captured (Gonczi & Hager, 2010). Furthermore, it does not take into account the fact that experts perform things in a different way to the newly competent. By skipping steps, experts may appear incompetent if a narrow stepwise approach is taken (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). A more integrated approach, as competencies are now more frequently conceptualised, can help avoid this, but only if assessments are also approached in an integrated manner and performance considered holistically.

Two ways to operationalise assessment of competency that enable an integrated, holistic approach to be taken involve the use of entrustable professional activities and developmental milestones. Both systems are now in widespread use in medical education.

Entrustable professional activities are descriptions of significant tasks or units of professional activity for which the student may be entrusted once they have reached a required level of competency (ten Cate et al., 2015). The adjectives *entrustable* and *professional* are both important in distinguishing these activities from others that learners may undertake. They are activities that are within the remit of the particular profession and not more generally, and are activities that a student would not be entrusted to undertake without being sufficiently competent. For example, conducting a literature review would not be an entrustable activity because it is not an activity limited to a profession and it is one that even those not yet capable could be assigned. Discharging a patient after surgery, however, would be. Entrustable professional activities therefore operationalise the conception of competency as an activity for which one is given permission or licence because of having reached a certain level of ability (ten Cate, 2005).

Entrustable professional activities are not in themselves as assessment instrument and are used with whatever instrument is suitable for the activity (ten Cate et al., 2015). Rather, they provide a new type of scale for assessing performance that focuses on levels of trust and independence. They harness key questions that supervisors make decisions about every day, of “would I trust this student to do this activity in this situation?”, and “can I leave them unattended or should I stay and supervise?” They involve consideration of the complex interrelated dimensions of competency to be brought to bear in a given context, including the

unobservable aspects, and thus enable a holistic and integrated approach to be taken. Each entrustable professional activity spans multiple aspects of competency and matrices can be used to blueprint entrustable professional activities against components of competency to document development of overall competency. Ten Cate et al. (2015) summarised ten qualities of students that enable supervisors to trust them, based on the research to date. These were (1) competency and clinical reasoning; (2) conscientiousness and reliability; (3) truthfulness or honesty; (4) discernment of limitations and inclination to ask for help if truly needed; (5) empathy, openness, and receptiveness toward patients; (6) skill in collegial and interprofessional communication and collaboration; (7) self-confidence and feeling safe to act; (8) habits of on-going self-evaluation, reflection, and development; (9) sense of responsibility; and (10) adequately dealing with mistakes of self and others. These aspects are at the core of many aspects of competency, demonstrating the alignment between the concept of assessing trustworthiness and assessing competency. As Gingerich (2015) explains, inferences about trustworthiness are an important part of all social judgments, and entrustable professional activities enable us to take advantage of this for assessment. The affective component to trust may require us, however, to take a different approach than that we are used to. How a supervisor feels, for example “I felt as though I had to take over”, becomes part of the assessment and is naturally subjective.

Milestones are descriptors of levels of achievement along a continuum of development from novice to advanced beginner, competent, proficient, expert, and then master, using the frameworks developed by Dreyfus (2004) and Carraccio et al. (2008). They have been introduced to medical education to facilitate assessment and reporting progress of students through levels of training, for the purposes of accreditation and accountability (Swing et al., 2013). They provide a way of describing different levels of a component of competency in terms of observable behaviours that are more specific and more easily assessed (Swing et al., 2013). Milestones may be accompanied by an expected timeframe for reaching the specified stage of development (Green et al., 2009). Like entrustable professional activities, milestones are not an assessment tool in themselves, but another type of scale descriptor that may facilitate the operationalisation of competency. One way of incorporating them into an assessment system is their use by a competency committee to judge the level of performance reached by a student based on data from a number of separate assessments (Holmboe et al., 2015). An assessment system may thus incorporate both entrustable professional activities and milestones (Royal College of Physicians and Surgeons of Canada, 2015).

Summary of the background to veterinary competency

The focus on competency in veterinary medicine in New Zealand is driven by legislative changes that have their roots in an increased need for accountability, and accompany a global increase in the use of a competency approach in other health professions. Contemporaneously problems in the veterinary industry have been attributed to a lack of business-related and other nontechnical skills, such as communication, amongst veterinarians. A number of veterinary competency frameworks have been developed, but show differences that reflect the fact that competency is a complex and abstract concept which combines aspects that are not easily described, and about which there are different perspectives and little empirical foundation. The concept encompasses a holistic interplay of components that is context-dependent. As well as discipline-specific knowledge and skill, a variety of nontechnical skills including communication skills, empathy, relationship-centred care, self-efficacy, confidence, optimism, reflective practice, and business skills make up aspects of competency. Some of these aspects are difficult to assess because they are abstract and unobservable.

Assessment of competency may be difficult to operationalise without either reducing competency to a series of observable aspects or assessing it globally. Considering developing independence and trustworthiness, and developmental milestones can help in establishing meaningful benchmarks for making decisions about performance.

Assessment of competency is workplace-based and should utilise a variety of assessment tools. The holistic and contextual nature of competency assessment requires that these be rater-based and therefore they suffer some similar problems. The next sections will consider the problems with the in-training evaluation and what is known about the process of judgement in in-training evaluation.

Criticisms of the in-training evaluation

The in-training evaluation has many advantages as discussed in the introduction, including authenticity and feasibility. However, it has been criticised for being subjective, based on impressions and not providing a meaningful basis for either summative or formative assessment (Lurie et al., 2009; McGill et al., 2011; Miller, 1990; Turnbull & van Barneveld,

2002). The criticisms centre around three aspects: the reliability of scores, the dimensions captured, and how in-training evaluation scores relate to other measures of performance. Variation between raters, and variation between cases and contexts, are thought to be at the heart of the issues.

This section will examine and discuss these criticisms. Because of the scarcity of veterinary specific research, much of the discussion in this section is drawn from the medical education literature. It takes a psychometric perspective because this has been the dominant frame for development and validation of assessment instruments in medical education until recently, and it is the frame from which the criticisms arise. It is a quantitative, realist perspective that approaches assessment in terms of true scores and error. There is a prominence of the concept of reliability and a focus on standardisation of contexts and raters, detailed specification of criteria, and of multiple sampling. As a perspective, it is criticised for making assumptions that competency is quantifiable, objectivity is desirable, and that subjectivity is unfair (Hodges, 2013). Other research perspectives that involve theoretical and qualitative approaches are increasingly present in the medical education literature. These enable new types of questions to be answered regarding the complexities of rater and student behaviour and its relationship with the social context of judgement. After reviewing the criticisms of the in-training evaluation in this section, the subsequent section will draw more from these other research perspectives in discussing the cognitive processes of evaluation.

The reliability of in-training evaluation scores

Reliability reflects the confidence we have that the score we award is a good representation of the student's actual performance (Furr & Bacharach, 2014). Demonstrating a reasonable reliability is thus an important part of making a defensible argument for the validity of scores. Reliability can be conceptualised in terms of the consistency of scores over replications (Haertel, 2006). In in-training evaluations, it is replications over placements or raters—thus interrater reliability—which is important for generalising performance to other placements or ratings.

Reported interrater reliabilities for scores on in-training evaluations are generally poor although there is great variation between studies. Some of this variation arises from

differences in the number of ratings per student, because reliability increases with the number of ratings. Studies involving only one to three ratings per student tend to report very low interrater reliability coefficients of 0.29-0.38, but much higher reliabilities (0.53-0.87) are seen in studies with at least seven ratings per student (Auewarakul, Downing, Jaturatamrong, & Praditsuwan, 2005; Ginsburg, Eva, & Regehr, 2013; Kwolek et al., 1997; Maxim & Dielman, 1987; Thomas, Beckman, Mauck, Cha, & Thomas, 2011; R. G. Williams, Verhulst, Colliver, & Dunnington, 2005). To put these values into perspective, for high stakes assessments a value of at least 0.80 is recommended and at least 0.90 is preferred (R. G. Williams et al., 2005). Very few studies have reported interrater reliability coefficients that approach 0.80 for in-training evaluations.

Generalisability studies, which separate the component sources of unreliability, provide a similarly gloomy perspective on the reliability of in-training evaluation scores. The variance attributable to student performance, which is the component of overall variance we are trying to measure, ranges from 4% to 28.8% in various reports (Kreiter, Ferguson, Lee, Brennan, & Densen, 1998; Kreiter & Ferguson, 2001; McGill et al., 2011; van Barneveld, 2005). This means that in these studies, between 71% and 96% of the variation in ratings was attributed to error. The usual assumption is that the error is all, or mostly, rater error.

Separating out components of this error further is made difficult in in-training evaluations because students usually receive evaluations from a unique set of raters in unique contexts and on unique tasks. Therefore, the effects of different raters and different cases or contexts are confounded and cannot be separately examined. However, studies of other clinical performance assessments, in which rater and context are not confounded, indicate that the error component likely comprises rater leniency, rater subjectivity, case or context specificity, and interactions between rater, student, and case (Alves de Lima, 2013; Margolis et al., 2006; J. R. Wilkinson et al., 2008). Rater leniency indicates that raters differ in the leniency or stringency of scores they award across all students they evaluate. It does not affect student rank, and just shifts the scores up or down compared to other raters. Rater subjectivity indicates that raters score some students differently to others. It contrasts with leniency-stringency in affecting the ranking of students relative to each other. Case specificity indicates that individual students differ in which cases and contexts they find more difficult. Interactions between rater, student, and case indicate differences in the way raters evaluate the student

performance *given the case or context*. This may indicate that raters make allowances for case complexity in reaching a judgement (J. R. Wilkinson et al., 2008).

There may also be other factors not investigated that contribute to the size of the error component, such as occasion. In-training evaluations are often conducted serially over a long period, such as a year, during which substantial growth in performance is expected. If not accounted for, variance over occasions could reduce the variance attributed to student (and hence reduce reliability) and increase the variance attributed to other components or the residual variance (Brennan, 2005). Van Lohuizen et al. (2010) demonstrated a statistically significant growth in medical student performance on the in-training evaluation over just a 14-week period. Accounting for it increased reliability such that the number of evaluations needed to produce a reliability coefficient of 0.8 decreased substantially from 17 to 11.

The results of these studies show that the interrater reliability of in-training evaluation scores is poor and rarely meets acceptable levels of reliability for high stakes assessment, although it can be improved by increasing the number of evaluations. Both rater errors and case specificity are likely to contribute to the unreliability of in-training evaluation scores. Rater and context factors may also interact so that rater-student subjectivity largely depends on the case or context, and may also depend on the occasion, which has generally been unaccounted for in studies of in-training evaluation to date.

Leniency-stringency

Leniency-stringency error refers to a general tendency for a rater to award scores that are lower or higher than other raters. Leniency is common in in-training evaluations, and findings of most scores being higher than the level benchmarked as *average* are frequently reported (Baker, 2011; Durning et al., 2010; Durning et al., 2005; McGill et al., 2011; McLaughlin, Vitale, Coderre, Violato, & Wright, 2009; Paget et al., 2013; Ryan, Mandel, Sama, & Ward, 1996). Some evidence suggests that this arises because supervisors interpret the word *average* more negatively than the literal meaning would suggest (Albanese, 2001; Kan Ma, Min, Neville, & Eva, 2013). High levels of performance are expected in courses in which there are high entrance standards and Govaerts et al. (2007) propose this may contribute to rater leniency.

Impaired detection of students who are not performing adequately is a problem thought related to leniency (Albanese, 2001). As R. G. Williams, Klamen, and McGaghie (2003) review, it has been commonly noted that students who fail are given scores that indicate adequate or even good performance on in-training evaluation. For example, Schwind, Williams, Boehler, and Dunnington (2004) studied the ability of in-training evaluations to detect deficiencies in clinical performance of senior medical students and found that, while deficient students tended to receive lower scores than others, 98.2% of their evaluations were *adequate* or better and almost 20% were *outstanding*.

Recent qualitative studies involving interviews with supervisors in medical education have documented the reasons given for lenient ratings. Stroud, Bryden, Kurabi, and Ginsburg (2015) reported that supervisors often gave students the benefit of the doubt when they had been too busy to observe them frequently or as a way of compensating for negative effects of the busy and stressful workplace environment. A variety of other reasons have been reported in studies by Cleland, Knight, Rees, Tracey, and Bond (2008), Kogan, Conforti, Bernabeo, Iobst, and Holmboe (2011), and McQueen et al. (2016). Personal and relational reasons included concern for the impact on the student and a sense of conflict between a supervisor's role as mentor and teacher and their role as evaluator and gatekeeper. Some supervisors were concerned that low ratings would demotivate students and were more lenient when students were aware of their problems, trying hard, or very committed to medicine. Supervisors felt they should be more stringent if they liked a student and more lenient if they did not. Confirmation from colleagues was important in alleviating uncertainty about the grade to award. Supervisors also mentioned feelings of being unskilled at giving negative feedback and anticipating an unpleasant task. Feelings that their own failings as a teacher were partly to blame for poor performance or that they had not had sufficient time with the student to observe them were also reported. Some supervisors expressed concerns about their own reputation. A major reason for leniency being extended to failing students was the anticipation of challenges, appeals, and potential litigation (Berendonk, Stalmeijer, & Schuwirth, 2013; Cleland et al., 2008; Dudek, Marks, & Regehr, 2005; Kogan et al., 2011; McQueen et al., 2016). Govaerts et al. (2007) note that the supervisor's trust that their assessment contributes to fair decisions and consequences is important in reducing leniency.

Leniency-stringency is a problem in in-training evaluation that means that ratings are generally high. Although high standards of performance are expected in medical and veterinary students

and may account for this tendency, differences in leniency between individual raters contributes substantially to the unreliability of in-training evaluation scores. In addition, leniency appears to contribute to difficulties in identifying failing students. Reasons for supervisor leniency may be personal, relational, administrative, and political.

Halo error and dimensionality of in-training evaluations

Halo error was a term coined by Thorndike (1920) to describe his suggestion that raters formed a general opinion or impression of the performance of a candidate that would then (unduly) influence the scores they awarded when assessing specific aspects of performance. The term halo misleadingly implies a tendency to award high scores, but this is not the intent. The general opinion held could be positive, negative, or intermediate, but the idea is that the scores on individual items have a tendency to be similar, and thus correlated with each other, to a greater degree than is warranted. The presence of halo error is important as it suggests that supervisors are unable to differentiate the multiple aspects of competency (Holmboe & Hawkins, 1998; Lurie et al., 2009; Turnbull & van Barneveld, 2002). This then calls into question the validity of scores awarded. Observations that have been used to conclude the presence of halo error include low variance (or low standard deviation) between items, high interitem correlation, and few factors found on factor analysis (Balzer & Sulsky, 1992).

Halo error is thought to be ubiquitous (Cooper, 1981). There are many reasons to expect it to be common, however the evidence cited for its presence is often based on findings that could represent a construct-relevant finding rather than error, or could arise from the unit of analysis chosen, the way the data was analysed, or characteristics of the data. Therefore, while some authors claim its prevalence in performance assessment is underestimated (Fisicaro & Lance, 1990), others dispute this (Murphy, Jako, & Anhalt, 1993). Because of these issues, it is difficult to determine whether halo error is an important problem in in-training evaluation, though it is often said to be so. Further research is needed. Here I will briefly review the reasons we expect halo error to be common, and then consider some of the methodological issues that make it difficult to identify halo.

Halo error is thought to be common because of two lines of evidence. One is that the mechanism by which it is thought to arise is common. Structural equation modelling indicates

that initial formation of a general impression best explains how halo might arise compared to other models (Lance, LaPointe, & Fisicaro, 1994). Impression formation is common (Wood, 2014), and conditions that predispose to forming impressions increase findings associated with halo error. These conditions include delays in rating, inability to observe some aspects of performance, and the structure and content of mark schemes (Cooper, 1981; Delandshere & Petrosky, 1998; Humphry & Heldsinger, 2014; Murphy et al., 1993). Familiarity with the ratee seems to decrease the presence of halo, presumably because it enables more details to be recalled (Murphy et al., 1993). The formation of general impressions is thought to be the result of Type 1 cognitive processing (Wood, 2014). This is the automatic way of thinking by which we make rapid and effortless decisions, without necessarily knowing what made us decide what we decided. Type 1 cognitive processing contrasts with Type 2 processing, which is more deliberative and involves working memory (Evans & Stanovich, 2013). A common perception is that Type 1 processing is error-prone, further solidifying the concept of halo as error. However, as Norman (2009) has reviewed, there is much evidence that it is not necessarily more error prone than Type 2 processing and is an important aspect of expert thinking. Therefore, we should not conclude that because a judgement was based on impressions it is necessarily wrong. However, even if the overall impression were accurate, halo error is still important because it reduces discrimination between dimensions assessed. This reduces the formative value of the evaluation and also makes it uncertain what the overall impression is based on.

The second line of evidence for halo error being common is that low dimensionality is a common finding in factor analysis studies, which are the usual means of determining the number of dimensions (called factors) present. In-training evaluations are frequently found to have few dimensions underlying scores and high interitem correlation. Table 2.2 collates the findings from several studies in medical and veterinary education and shows that the finding of one or two dimensions has been most frequently reported. When two dimensions are present they consistently encompass technical knowledge and skills as one dimension, and interpersonal skills and or professionalism as the other. This is much less detail for identifying strengths and weaknesses than the multiple different aspects of performance the instruments were designed to assess.

Table 2.2: Details of studies investigating the internal structure of in-training evaluations.

Aspect of analysis and findings	Studies
Number of dimensions (factors) found	
1	9, 10, 11, 15, 18, 21, 22
2	2, 3, 5, 12, 14, 16, 19, 23, 24, 25
3	4, 6, 7, 13, 20
4 or 5	1, 8, 17
Analysis method used	
Principal components analysis	1, 4, 8, 9, 13, 14, 16, 17, 18, 19, 20, 22, 23, 24, 25
Common factor analysis	2, 3, 6, 7, 10, 11, 12, 13
Not specified	5, 15, 21
Rotation method used	
Orthogonal	1, 3, 4, 5, 6, 7, 8, 10, 13, 14, 16, 17, 19, 20, 21, 23, 24, 25
Oblique	
Both orthogonal and oblique	2, 12
Not rotated (only one factor)	9, 11, 15, 18, 22
Methods for determining number of factors	
Eigenvalue > 1 only	2, 7, 9, 11, 14, 16, 18, 19, 24, 25
Eigenvalue >1 plus other criteria	3, 4, 8, 12, 13, 17, 21
Scree test included	8, 10, 12, 13, 17, 20, 21
Parallel analysis included	13, 15
Variance accounted for included	3, 5, 21, 23
Dependency (repeated measures) management	
Management not mentioned	1, 4, 5, 8, 9, 10, 12, 13, 18, 19, 21, 22
Managed by using mean	3, 6, 11
Not present	2, 7, 14, 15, 16, 17, 20, 23, 24, 25
Missingness management	
Management not mentioned	1, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 25
Pairwise deletion	13
Listwise deletion	2, 4, 5, 11, 17, 18
Mean replacement	3
Balanced design	9
Overall score inclusion in factor structure	
Included in factor structure	5, 11, 10, 12, 13, 14, 22
Present but not included*	7, 17, 21, 23, 24
Not present in the instrument	1, 2, 3, 4, 6, 8, 9, 15, 16, 18, 19, 20, 25

Key to studies: **1**: Brasel, Bragg, Simpson, and Weigelt (2004); **2**: Dielman, Hull, and Davis (1980); **3**: Forsythe, Mcgaghie, and Friedman (1985); **4**: Fuentealba and Hecker (2008); **5**: Ginsburg et al. (2013); **6**: Gough, Hall, and Harris (1964); **7**: Hojat, Veloski, and Borenstein (1986); **8**: Kassam, Donnon, and Rigby (2014); **9**: Kreiter and Ferguson (2001); **10**: Kwolek et al. (1997); **11**: Lee and Wimmers (2010); **12**: Maxim and Dielman (1987); **13**: McGill, van der Vleuten, and Clarke (2013); **14**: McLaughlin et al. (2009); **15**: Metheny (1991); **16**: Nasca et al. (2002); **17**: Paolo and Bonaminio (2003); **18**: Pulito, Donnelly, and Plymale (2007); **19**: Silber et al. (2004); **20**: Skakun, Wilson, Taylor, and Langley (1975); **21**: Teman, Minter, and Kasten (2016); **22**: W. G. Thompson, Lipkin, Gilbert, Guzzo, and Roberson (1990); **23**: Verhulst, Colliver, Paiva, and Williams (1986); **24**: Woloschuk, McLaughlin, and Wright (2010); **25**: Woloschuk, Myhre, Jackson, McLaughlin, and Wright (2014).

Note. *indicates the overall score was present in the evaluation instrument but was not included in the factor analysis in determining the factor structure.

Concluding that low dimensionality and high item intercorrelation are the result of halo error, however, neglects to allow for other reasons for these findings, and there are several of these that are important to consider when studying in-training evaluations. Firstly, intercorrelation between aspects of performance might be an expected part of what we are assessing (Cooper, 1981). Many aspects of veterinary competency are likely to be correlated, even for those that seem distinct, such as discipline-specific knowledge and communication skills. It is not unreasonable to imagine that those with better knowledge would be able to communicate about it more clearly and confidently, use technical language correctly, and to simplify appropriately for clients, therefore scoring well on these aspects of communication. Empirical support for high correlations between dimensions of performance even after halo error has been controlled for has been found in other types of evaluation (Lai, Wolfe, & Vickers, 2014; Lance, Hoffman, Gentry, & Baranik, 2008; Viswesvaran, Schmidt, & Ones, 2005).

Secondly, low dimensionality and high intercorrelation may be a result of characteristics of the data or the way it is analysed. A common issue is that halo error applies to the ratings made by a single rater on a single ratee. It is the lack of variance or high correlation between items within each evaluation that is the correct unit of analysis. Yet in factor analysis, the data from all evaluations is usually pooled for analysis, which can result in the appearance of halo error where there is none (Balzer & Sulsky, 1992; Murphy, 1982). Another common issue is that the analysis method may affect the number of dimensions concluded are present. Factor analysis studies have been criticised for using methods inappropriate for the assumptions of the study (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Watson & Thompson, 2006; Wetzel, 2012). The criticisms are well founded because Fabrigar et al. (1999) found that for a number of published reports in psychological research, repeating the analysis using more appropriate factor analysis methods revealed different conclusions, including a different number of dimensions. Areas particularly criticised include using principal components analysis rather than common factor analysis, rotating factors using orthogonal methods rather than obliquely, and determining the number of factors using the eigenvalue greater than one rule. Table 2.2 shows that these criticised methods have been commonly used and therefore it is possible that reanalysis using different methods would reveal different conclusions about the number of dimensions present.

Additional issues that may cause intercorrelation and low dimensionality to be found if not accounted for during analysis are the presence of missing data and dependency from nested or

repeated measures data. Both violate assumptions of statistical modelling and are serious issues that can affect the validity of the research conclusions by introducing bias or increasing the correlations found (Bliese & Hanges, 2004; Glass, Peckham, & Sanders, 1972; White & Carlin, 2010). Cook, Beckman, Mandrekar, and Pankratz (2010) illustrated the problems for factor analysis resulting from unaccounted for dependency. Item intercorrelations were substantially inflated from a range of 0.14-0.46 when repeated measures were accounted for to 0.38-0.63 when they were not. The inflation produced by dependency could be misinterpreted as halo error. Table 2.2 shows that many studies of in-training evaluation do not discuss how missingness or dependency has been managed in the analysis. An additional problem is that some studies have included the overall score as if it were another item when factoring (Table 2.2). Since the overall score is expected to be strongly related to at least some of the items in the instrument (if not all), its inclusion would increase correlations and could alter the factor solution (Gorsuch, 1983). The overall score item would be expected to appear as a strong common factor related to all or most items, which may be misinterpreted as indicating the presence of halo error (Murphy, 1982).

Given the discussion to this point, it seems that we cannot reach firm conclusions about the degree of item intercorrelation present in in-training evaluations, and therefore how much halo error, as technically defined, may be present. Nor can we conclude that what is technically defined as halo error is all construct-irrelevant and therefore rightly categorised as error. Research that helps clarify the dimensions actually being assessed by in-training evaluations and how these relate to each other is likely to be particularly helpful in distinguishing correlation we expect, from correlation we don't expect (halo error).

Other research that gives us insight into halo effects is that which investigates the cognitive aspects of the rating process. There is much evidence of cognitive processes that are associated with halo effects, but also evidence that suggests supervisors do distinguish a range of dimensions, and therefore that correlation between items may be part of the construct we are assessing. This will be discussed after discussing how in-training evaluations relate to other measures of performance.

How in-training evaluation scores relate to other measures of performance

In-training evaluations are also criticised for their low correlations with other measures of performance (Turnbull & van Barneveld, 2002). A strong relationship with other measures is expected if both instruments reflect the same underlying constructs and therefore a lack of such a relationship undermines arguments for the validity of scores in in-training evaluations. However, although very low correlations (0.22-0.28) have been frequently reported, as Hamdy et al. (2006) found in their systematic review, most early reports had not corrected for the attenuation that results from scores being less than perfectly reliable. More recent studies in which unreliability was accounted for have produced much more substantial correlations 0.48-0.71 (Auewarakul et al., 2005; Daelmans et al., 2005; Ferguson & Kreiter, 2004).

Aside from unreliability, other characteristics of the data may also reduce correlations between measures of performance including the skewed distribution of ratings, lack of a linear relationship between the two measures, heteroscedasticity (variation in the degree of variance over different values of a variable), and homogeneity of performances which limits the variability of scores (Gonnella, Hojat, Erdmann, & Veloski, 1993). Each of these aspects is common in in-training evaluation data and limits the amount of correlation with other measures that is possible; therefore, perfect correlation should not be expected in any case. In addition, in making such comparisons it is often assumed that the validity of scores on the comparison assessments is well established and this may not always be the case. T. J. Wilkinson and Frampton (2004) dramatically demonstrated the effect of improvements to the written examination on the correlation of its scores with in-training evaluations, which rose from 0.17 (which was not significantly different to zero) to a statistically significant moderate correlation of 0.54.

In summary then, while low correlations between in-training evaluations and other performance assessments have frequently been found, more recent studies suggest that methodological issues may well have been obscuring moderate relationships. Low correlations with other measures may cast doubt on validity if they imply the in-training evaluation is not assessing what it is designed to assess, but equally, methodological issues aside, could indicate that the in-training evaluation is capturing something the other assessments are not. Whether that something is valuable or not is an important question which requires research into what

the in-training evaluation is capturing. Without knowing this, the significance of relationships, or lack of them, between in-training evaluation scores and other measures of performance are difficult to determine.

Summary of problems with the in-training evaluation

Major criticisms of the in-training evaluation centre on score reliability, its inability to distinguish the dimensions of performance we intend to assess, and a lack of relationship with other measures of performance. The presence of these issues has made it difficult to sustain an argument for the validity of scores on the in-training evaluation and as a result its use for high stakes assessment has been questioned.

Reviewing the literature on the reliability of in-training evaluation scores confirms that interrater reliability is poor, although it can be improved by performing multiple evaluations. Unreliability appears to be contributed to by a combination of rater leniency, rater subjectivity, case or contextual effects, and their interaction. Occasion may also contribute although there has been little research on this to date. In-training evaluations have been generally found to be uni- or two-dimensional, with high interitem correlations leading to a presumption of halo error. However, because other causes of low dimensionality have infrequently been considered, whether this indicates halo error needs further investigation. The research on the relationship of in-training evaluation scores with other measures of performance suggests that moderate relationships are present. Characteristics of the data may prevent very high correlations, but equally, the in-training evaluation may be assessing aspects not assessed by other performance measures.

Further research that investigates what the in-training evaluation is assessing in terms of the number and types of constructs, how these are related to each other, how they are influenced by case or context, and how they vary from supervisor to supervisor would be of value. Developing an understanding of the rating process, and how a supervisor's impressions are formed and translated to scores, is needed. The next section will review the research on this to date.

Elucidating the process of evaluation

Within the field of medical education the process of evaluation has been an important area of study, especially in recent years. Related to this is research on the process of clinical judgement, as diagnostic errors (and avoiding them) are of great importance. Medical education researchers have drawn on the wider literature in social and cognitive psychology as a basis for developing theory. The studies are often qualitative in nature. Many involve the use of recorded clinical consultations, usually performed by actors, but, in some studies, real patients and doctors or students are recorded. Study participants view the recordings and are either asked to verbalise their thoughts as they evaluate the performance of the recorded student doctor, or are interviewed about aspects of it afterwards. Other studies have analysed the narrative feedback provided on in-training evaluations.

For the purposes of review, it is useful to break down the process of evaluation into a series of several steps. Although artificial and implying a simple linear sequence for a process that is far more complex and, as yet, not fully understood, it forms a useful framework from which to study the process of evaluation in in-training evaluation. The framework is illustrated in Figure 2.1. It is a distillation of ideas derived from the work of many researchers, particularly Govaerts et al. (2007), Kogan et al. (2011), Mazor, Zanetti, et al. (2007), G. Regehr et al. (2011), St-Onge, Chamberland, Levesque, and Varpio (2016), R. G. Williams et al. (2003), and Yeates, O'Neill, Mann, and Eva (2013).

Within this framework, the evaluation process is conceptualised as involving the bringing together, for the purposes of comparison, the "picture" the rater forms of the student's performance through observation, and the rater's expectations of performance for competent students, in a process of judgement that results in the decision to award a particular score. That process involves an initial evaluation that is global and made in narrative form and which is then translated into a global score and scores for the separate items as required on the evaluation form. Each of these areas can be subject to influences on the rater judgement and may contribute to variability between supervisors in the score they award.

This section will begin by reviewing the research that elucidates the expectations supervisors have of student performance. I will then move on to the process of observation and forming a picture of the performance, and then the process of comparing these to form a global

narrative evaluation. Lastly, I will consider the translation of that evaluation to scores on the form.

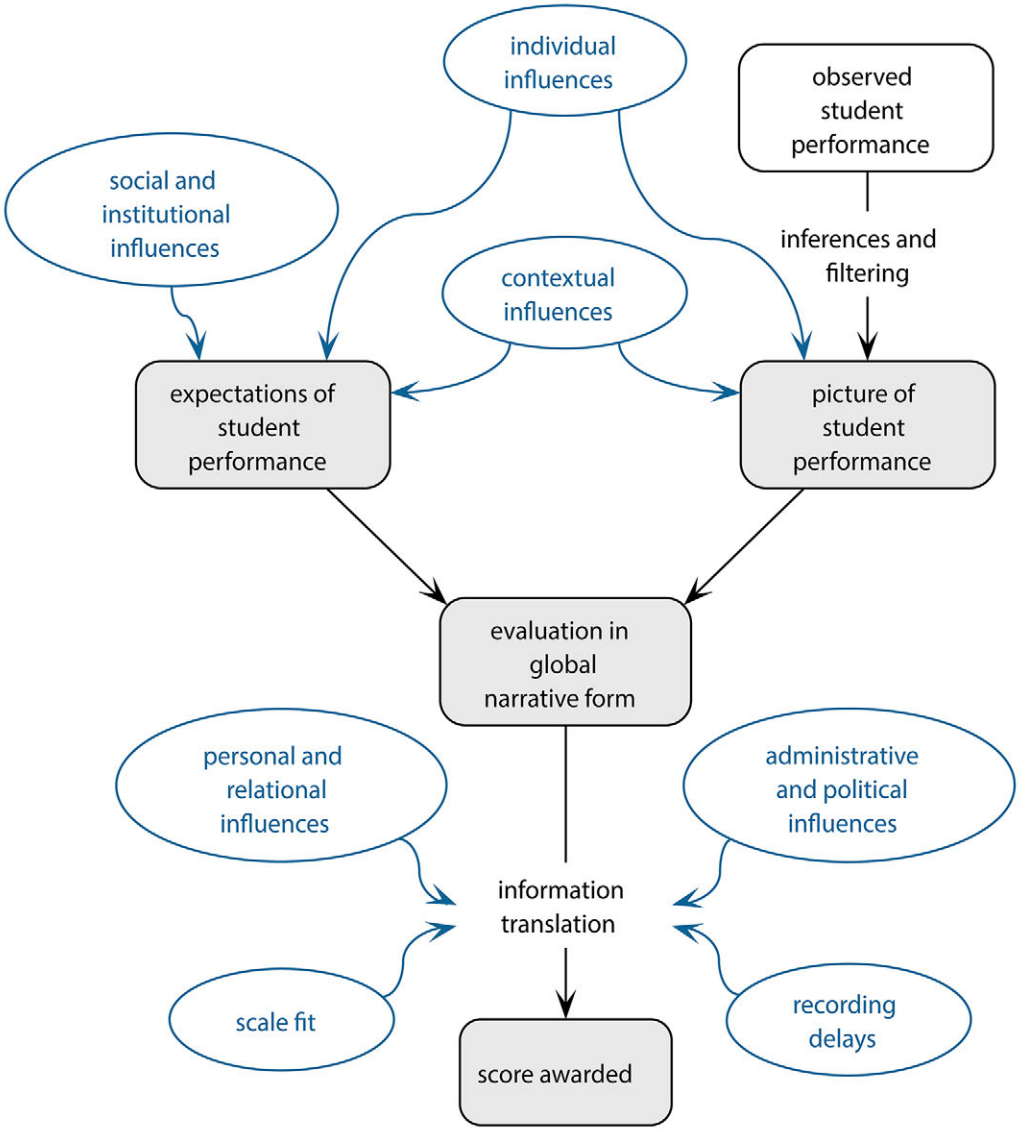


Figure 2.1: Concept map for the process of judgement in in-training evaluation. The main processes are shown in black and the main influences are shown in blue.

Expectations of student performance

Although the expectations for student performance are stipulated in learning outcomes, mark schemes, and competency frameworks, supervisors must construct a personal understanding of criteria and standards. Studies in which supervisors explain their thinking while judging student performances have shown that supervisors did not use competency assessment criteria as their main frame of reference. They found them ambiguous (Yeates et al., 2013). Instead they relied mainly on their own experiences as a student and as a practitioner, and the performance of other students they had supervised (Kogan et al., 2011; St-Onge et al., 2016; Yeates et al., 2013). Supervisors interpret and internalise these various sources of information in an ongoing process of refinement through experience. Socialisation with other supervisors is also important in developing a common understanding of expectations (Shay, 2004; St-Onge et al., 2016).

In simulated assessments, where raters observed recorded performances, supervisors considered a variety of aspects of performance including the student's factual knowledge, communication skills, diagnostic accuracy, the process of investigation that the student followed, their recognition of key aspects of the case, and patient safety (Kogan et al., 2011; Yeates et al., 2013). However, the features considered, how they were related to each other, and how they were applied, varied widely between supervisors and between evaluations (Govaerts, van de Wiel, Schuwirth, van der Vleuten, & Muijtjens, 2013; Kogan et al., 2011; Yeates et al., 2013). Contextual factors, such as the difficulty of the case, and the experience of the student with the procedure or situation, were also found to influence their expectations (Kogan et al., 2011).

A small number of studies in medicine and social work have stepped outside the context of assessment to investigate what is important to supervisors, which may be different to what supervisors are asked to assess (Bogo et al., 2006; Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010; Lavine, Regehr, Garwood, & Ginsburg, 2004; Rosenbluth, O'Brien, Asher, & Cho, 2014). Supervisor descriptions of excellent and weak students they had encountered on placements frequently concerned aspects of the student's disposition, personality, demeanour, or character. These included aspects such as intelligence, sense of humour, humility, buoyancy, shyness, tolerance, "teachability", cynicism, insecurity, commitment, arrogance, maturity, and adaptability that did not marry well with competency frameworks or were only partly

represented (Ginsburg et al., 2010; Rosenbluth et al., 2014). Other prominent themes that may not appear in competency frameworks were the student's impact on the supervisor in making their work easier or harder, their trustworthiness, and their enthusiasm for work or learning, work ethic, or energy (Bogo et al., 2006; Ginsburg et al., 2010; Lavine et al., 2004; Rosenbluth et al., 2014).

In two studies, the interviewers asked supervisors to describe particular students they recalled rather than describing all the things they thought were excellent or not excellent that different students might do. This enabled the way supervisors weighted different aspects of performance to be considered. Supervisors were found to vary in the way they weighted aspects of performance, and different aspects were considered for excellent and weak students (Ginsburg et al., 2010). Furthermore, the degree to which each theme was considered by supervisors was different for different students (Ginsburg et al., 2010). Work ethic, communication, and leadership were discussed much more frequently in relation to excellent students than weak students, and trustworthiness and response to feedback were discussed much more frequently in relation to weak students than excellent students. Poor performance in one area could affect the overall evaluation, but the degree to which poor performance could be balanced by good performance in other areas was different for excellent and weak students. In excellent students, relative deficiencies were most frequently seen in knowledge base or knowledge translation, but were also seen in organisational and time management skills, disposition, or discerning their own limits. In weak students, strengths were seen in knowledge base or intelligence, having pleasant personalities, or good organisational and time management skills.

Bogo et al. (2006) found that personal qualities of the student were more important to supervisors than their skills. Excellent and weak students were differentiated by "a constellation of personal qualities possessed by students that were perceived as affecting their approaches to learning, their interactions with others in the organization, their relationship with the field instructor, and their ability to develop relationships with clients" (Bogo et al., 2006, p. 589). Deficiencies in weak students were often attributed to lack of interest (Bogo et al., 2006).

From the research reviewed here, it is apparent that the expectations of supervisors are personally constructed understandings that are developed and refined over time. They are influenced by social, institutional, and contextual factors as well as the supervisor's own experiences as a student and supervisor. As a result, expectations are likely to differ between supervisors and from student to student. Studies of what is important to supervisors confirm these differences. They also show that supervisors' expectations extend beyond competency frameworks and assessment instruments to involve personal aspects of students, their trustworthiness, enthusiasm, and impact on the supervisor. As a result, supervisor expectations are highly personal and somewhat idiosyncratic, which is likely to contribute to interrater variability (R. G. Williams et al., 2003; Yeates et al., 2013). This conflicts then with the expectation that supervisors should use fixed and common criteria and standards for comparison of student performance. It also may be one reason why rater frame-of-reference training has not consistently been found to increase rating accuracy (Cook, Dupras, Beckman, Thomas, & Pankratz, 2009; Weitz et al., 2014) despite its benefits in helping supervisors identify more specific expectations (Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2015).

Forming a picture of student performance

The student's actions are the primary source of information informing the picture supervisors develop (St-Onge et al., 2016). A variety of situational factors will influence the student's actions, not all of which are under a student's own control or related to their own competency, but which have the potential to influence the picture the supervisor develops of the student's performance. These include the cases they are assigned and the actions of other team members. Other factors influence how many and which of a student's actions are observed by the supervisor.

The frequency of direct observation of the student will depend on a supervisor's other commitments, and how tired or distracted they are (R. G. Williams et al., 2003). The number of direct observations may actually be quite small. Reports by 10-20% of medical students that they are never observed performing activities such as collecting a history or completing a physical examination have been persistent (Association of American Medical Colleges, 2015; Burdick & Schoffstall, 1995; Daelmans et al., 2004; Scott, Irby, Gilliland, & Hunt, 1993; Stillman, 1986, 1991; Stillman et al., 1990). An observational study of senior emergency medicine

students indicated that the majority of time that supervisors spent interacting with them was while they presented and discussed the cases they were dealing with, rather than while they actually cared for patients (Chisholm et al., 2004). R. G. Williams et al. (2003) conclude that time spent with the student, as well as the frequency of observation, influences the quality of ratings, based on research in non-medical disciplines. Issues with the frequency and duration of direct observation of students on placement have been well recognised in medical and veterinary education and have led to programmes of assessment involving explicit and planned direct observation to supplement or replace in-training evaluation (Hecker et al., 2012; Kogan et al., 2009; Magnier et al., 2012).

The picture that a supervisor forms will also be influenced by what they notice when they observe the student. Studies aimed at gaining insight into what supervisors notice when they observe students in the workplace suggest that there is a marked difference between individuals. Yeates et al. (2013) found that raters paid attention to different aspects of student performance when viewing the same video, and called this differential salience. In essence, each rater judged a subset of aspects of the performance that was unique, therefore making it quite likely that they may reach different conclusions about the quality of performance. Several other studies have also shown that there is a great deal of variation in what supervisors notice, and that this can vary over the course of an observed performance, and between contexts (Govaerts et al., 2013; Mazor, Zanetti, et al., 2007; St-Onge et al., 2016). Case and context-specific comments showed that raters actively considered aspects of the case and context while thinking about the student's performance (Govaerts et al., 2013). Importantly, Govaerts et al. (2013) found that supervisors considered a number of dimensions and sub-dimensions for each student. As they conclude, this contradicts the conclusions from factor analysis studies that raters only use one or two dimensions. Dimensions were interlinked and considered together during evaluation, rather than one at a time, which may account for the high intercorrelation found in other studies, and suggests that rather than being halo error, the intercorrelation is construct-relevant.

One might expect that the differences between the aspects raters attend to is magnified when different disciplines or subdisciplines are considered, as the nature of the work may call for different key competencies. Conflicting findings leave open the question of the importance of subdiscipline as a cause of rater variability in medical education (McGill et al., 2013; Silber et al., 2004) and the question has not been examined in veterinary medicine. However, the

findings of a recent survey of Canadian veterinarians about the desirable personal characteristics of veterinary students give some insight (Conlon, Hecker, & Sabatini, 2012). Significant differences were found between small animal, equine, and mixed animal practitioners in the relative importance placed on communication skills, critical and creative thinking, empathy, and sound judgement. Differences in what is noticed by supervisors are therefore likely to occur both within and between subdisciplines in veterinary medicine.

Some evidence suggests that supervisors infer what they don't observe (Mazor, Canavan, Farrell, Margolis, & Clauser, 2008; Pulito, Donnelly, Plymale, & Mentzer, 2006; Stroud et al., 2015). Inferences may be based on case discussions with the student, reports of others, or a lack of information to the contrary. These inferences are based on assumptions that may not be warranted about relationships between the skill being assessed and the skill being observed, and may increase correlations in the scores of aspects of the performance intended to be assessed independently, as has been pointed out by Levine and McGuire (1971).

As well as inferences about a student's actions, supervisors make inferences about the reasons for them. Student's thought processes, intentions, feelings, personality, disposition, skills, motivation to improve, prior experiences, and preparation were frequently inferred (Govaerts et al., 2013; Kogan et al., 2011; St-Onge et al., 2016), and raters were not always aware that they were doing so (Kogan et al., 2011). Thus, the interpretations of the appropriateness of the student's actions could depend on the inferences supervisors made about the reasons for them. For example, Lavine et al. (2004) found that the same behaviours were viewed positively or negatively according to the supervisor's interpretation of the reasons for the behaviour.

One concern is that if inferences are used as a basis for judgement they may be wrong, as the supervisor may not have sufficient information to make correct inferences. Kogan et al. (2011) noted that raters rarely questioned their own assumptions or confirmed their inferences with the student (actor) they were evaluating in live performances, despite the opportunity to do so. Another concern is that the number of potential inferences is large and may be a significant contributor to variability in ratings. However, in a recent study, Gingerich, van der Vleuten, Eva, and Regehr (2014) found that, between raters, the range of inferences about a student's personality was limited. There was not a different viewpoint for each of the 34 raters in the study, but only 2-5 for each performance rated. Their findings indicate that there is a pattern

of consistency to the inconsistency between raters and Gingerich, van der Vleuten, et al. (2014) propose that these may be evidence of a small set of different but valid perspectives of the performance, that provide meaningful contributions to a full picture of performance, rather than reflecting rater subjectivity.

Thus, the research reviewed in this section suggests that the picture that the supervisor "sees" of the student's performance depends not only on what the student does, but what the supervisor is present to observe and notices about the performance. This may be a very limited sample of the student's performance and supplemented by inferences about the student's actions based on other observations and information from other sources. Supervisors pay attention to a number of different dimensions of performance, contradicting concerns of low dimensionality of the instrument. However, each supervisor has particular aspects of performance they consider salient, and what dimensions they pay attention to differs between supervisors, between occasions, and may also differ depending on the subdiscipline. Inferences that help a supervisor explain why the student behaved as they did are an important part of forming a picture. Supervisors also actively consider context in forming their picture. Together this means that the picture a supervisor develops of a student's performance is somewhat unique, and that different pictures can be developed by different supervisors even when viewing the same performance in the same context. Although this may be seen as a problem with rater judgement, it could also be seen as representing the range of perspectives that, together, would form a meaningful and comprehensive evaluation if they could be captured.

Comparing expectations and pictures to make an evaluation

The cognitive processes that I have conceptualised as involving comparison of expectations with the picture formed have been the subject of study by a number of research groups in medical education. Although portrayed in Figure 2.1 as if a separate process from forming expectations and a picture of performance, it is not. In reality, all three aspects are intertwined, inform each other, and are revisited as new information is added or conclusions are drawn. Thus, the discussion of the research in this section overlaps to a degree with the aspects already discussed.

St-Onge et al. (2016) studied the cognitive processes of supervisors as they rated recorded performances. They found evidence that raters searched for specific information as a basis for their decision, compared information to their expectations, made inferences to make sense of their observations, and weighed the relative significance of the observed elements in coming to an evaluation. Raters integrated a variety of information, including the observed performance elements and contextual information about the clinical situation and the student's experience. Govaerts et al. (2013) suggested that raters may use pre-existing mental models or schemas to help them categorise the information to make sense of it. The schemas are thought to guide the information that raters pay attention to and what is remembered about a performance. They found that schemas about the student as a person (person schemas), such as their personality or disposition, were developed very early on during observation of the performance, were informed by inferences, and differed between raters even when viewing the same performance. Govaerts et al. (2013) concluded that differences in the use of person schemas during rating could underlie the interrater differences in evaluations.

Schemas about the specific context were also used (Govaerts et al., 2013). Context-specific schemas may contribute to expertise in rating by facilitating the integration of contextual information into evaluation (Govaerts, Schuwirth, van der Vleuten, & Muijtjens, 2011). Other research has also shown that raters take account of context in making evaluations (St-Onge et al., 2016; Stroud et al., 2015). When assessing clinical communication, taking account of context has been found to result in higher ratings (Essers et al., 2013), but, because raters were also found to differ in how they incorporated contextual factors into their judgement, even when viewing the same recorded performance (Essers et al., 2015), taking account of context could also contribute to variability between raters.

Context is not only part of what a supervisor considers but it is also an influence on the supervisor and may therefore influence their evaluation in other ways. As noted earlier, context influences what is seen and noticed. It also contributes to mental workload and so influences the processes of evaluation. This will be discussed in a later section. Contextual factors may also influence the translation of judgement to a score, by affecting a supervisor's willingness to award a score that reflects what they really think, as discussed in the section on leniency-stringency. There are a host of other contextual influences that may affect a rater's judgement. For example, the rater's emotions can change implicit goals and the content and

depth of thought, and interact to affect judgement (Lerner, Li, Valdesolo, & Kassam, 2015). Contextual factors thus play an important role in all aspects of the rating process, influencing rater, as well as student, performance. Kogan et al. (2011) related this to situated cognition theory, which considers the holistic interplay and inseparableness of the rater's knowledge and thoughts from the context in which they are occurring. Both must be considered to fully understand how raters form an evaluation. Likewise, it is entirely appropriate that raters actively integrate considerations of the effects of context on the student into their evaluation. Not to do so would compromise validity because the context and motivation of the student gives us insight into the appropriateness of their behaviour (Ginsburg et al., 2000). Many commentators have therefore concluded that we must find ways to evaluate students that value context specificity rather than consigning it to error and attempting to eliminate it through standardisation (Ginsburg et al., 2000; Ginsburg, Regehr, & Lingard, 2004; Ginsburg, Regehr, & Mylopoulos, 2009; Govaerts et al., 2007; Hodges, 2013; Rees & Knight, 2007; van der Vleuten, 2014).

There is some research to suggest that in balancing and weighing all the observations, contextual factors, inferences, and interpretations against their expectations, raters first form a holistic, qualitative evaluation of student competency and then translate this to the overall score and item scores. Yeates et al. (2013) distinguished two cognitive processes: that of converting a narrative judgement into scores that reflected the scale descriptors; and that of deriving separate scores of each item based on their holistic judgement. In support of this, researchers have found that raters speak in holistic terms when explaining their judgement (Berendonk et al., 2013; Mazor et al., 2008; Yeates et al., 2013).

The formation of a holistic global evaluation is often thought of as based on impression formation and automatic, Type 1 cognitive processing; the connotation being that it is error prone as previously discussed. However, it is important to point out that holistic thinking is not necessarily a form of Type 1 processing. Holistic thinking attends to context, relationships, and associations (Buchtel & Norenzayan, 2009). It is contrasted with analytic thinking, which decontextualizes objects. Evans (2011) describes these as two different modes of thinking, as opposed to representing Type 1 and Type 2 processing. Both could be styles of Type 2 thinking that are deliberative and involve working memory. Both could also be automatic and require no controlled attention, especially with the development of expertise, and thus be examples of Type 1 thinking. Sadler's (2009b) description of holistic evaluation as a progressive and

complex process, involving attention to both particular aspects and the whole performance to build a picture of its quality, is consistent with a Type 2 process. It is also consistent with the observations of Govaerts et al. (2011) that both experienced and less experienced raters continuously reconsidered and revised their evaluation as new information came to hand and that this could result in a change in their evaluation through the course of observing the performance. Thus, it appears that although initial impressions are made, evaluation does not stop there, and there is an ongoing process of accumulating more information, balancing and weighing it, to arrive at a holistic judgement.

In this section, I have discussed some of the cognitive processes that combine the expectations a supervisor has with their developing picture of the performance of the student to formulate their evaluation. The use of pre-existing mental models may provide expertise that facilitates judgement but may also underlie differences between raters, as the development of mental models is informed by inferences about the student as a person and likely to be very idiosyncratic. The judgement process also seems to involve weighting and balancing of expectations, observations, inferences, and contextual considerations, and because the supervisor themselves is part of the environment, contextual factors will also influence the judgement process. The holistic judgement develops over the course of observation, and is subject to continual re-evaluation and revision. While rater expertise is likely to facilitate it being fast and automatic, holistic evaluation it is not necessarily an indicator of Type 1 processing. It is a complex integrative process and may involve a great deal of mental work.

Effect of cognitive load on making an evaluation

The research reviewed to this point makes it clear that raters are balancing a number of things when evaluating students. Tavares and Eva (2013) proposed that the rating process may cause cognitive overload because of the number of things that must be attended to and processed in order to make an evaluation and that this may contribute to variation between the judgements of different raters. They reviewed the literature on cognitive load theory, which proposes that although long term memory is unlimited, short term memory, through which new information interfaces with the long term memory, has a limited capacity. That capacity can be exceeded when there is more information that can be processed at one time. Expertise increases the capacity of processing so that more information can be managed. Overload

results in diminished performance and the use of strategies such as avoidance or simplification of the task. Simplification strategies include serial processing, where fewer aspects are considered at a time, and the use of heuristics. These strategies are associated with differences between raters and with rater error because not all the available information is considered. Tavares and Eva (2013) cited literature which demonstrates the decline in cognitive performance associated with increased mental workload in a number of disciplines, but pointed out the scarcity of research on mental workload relating to rating clinical performances.

Since then, research reported by Byrne, Tweed, and Halligan (2014) suggests that examining students in clinical medicine was associated with an excessive workload for raters. In further work, Tavares and Eva (2014) found that reducing the number of aspects to be assessed improved interrater reliability from 0.45 to 0.70, as well as the number of relevant behaviours raters identified, which they proposed was a result of reduced cognitive load. However, they were unable to demonstrate a change in measures of cognitive load, making the reasons for the differences uncertain.

Although not specifically looking at assessment, the work of another research group on cognitive load in relation to clinical problem solving is of relevance. Durning et al. (2012) found that manipulation of contextual factors in clinical scenarios significantly influenced the cognitive load and reduced diagnostic accuracy with a small to moderate effect. In another study the researchers found that when multiple contextual factors complicated the clinical scenario it could lead to key information being missed as the doctors evaluated the case, and postulated that this was a result of high cognitive load (Durning, Artino, Pangaro, van der Vleuten, & Schuwirth, 2011). These studies lend further support to the concept that increased mental workload, as a result of the influence of multiple contextual factors that must be considered, can affect expert judgement and therefore may affect evaluation of student performance on placements.

Further research is necessary to establish how cognitive demands change in different rating circumstances, and to establish whether mental overload negatively affects rating processes. As Wood (2013) commented, such studies are also likely to be of interest for what they reveal about cognitive processes involved in rating, and to what degree raters use Type 1 processing

or Type 2 processing. Dual process theorists consider that Type 1 (automatic) processing is the default thinking strategy but that Type 2 processing intervenes if initial intuitive decisions seem wrong or when the task is difficult, new, or there is motivation to utilise working memory (Evans & Stanovich, 2013; Kahneman & Frederick, 2005). Govaerts et al. (2007) also noted that Type 1 processing was likely to dominate in performance assessment, but that the demands of the task, the rating format, and whether the purpose of ratings is to evaluate overall performance or specific aspects would influence whether raters use Type 1 or 2 processes. St-Onge et al. (2016) noted that raters slowed down and re-examined their judgement when what they observed differed from what they anticipated. This behaviour is consistent with a switch from Type 1 to Type 2 processing, as needed, to make sense of the observations. Thus, Type 1 processes may be commonly used in rating irrespective of mental workload and it is therefore not clear if increased mental workload causes a change in the type of processing or just degraded performance. We should not assume that degraded performance indicates that Type 1 processing occurring because, as previously discussed, there is much evidence to show that both Type 1 and Type 2 processing are sometimes accurate and sometimes inaccurate (Evans, 2008). Indeed Gigerenzer and Brighton (2009) strongly argue, albeit somewhat controversially, that heuristic use enables intelligent decisions from limited information. Therefore in the normal conditions of rating complex performances of competency, in which aspects to be considered include intangible ones where information is necessarily limited, Type 1 processing may be not only the most commonly used strategy, but also the most appropriate. Govaerts et al. (2007) highlighted, in their review, research indicating that Type 1 processing can be highly accurate for overall evaluation, albeit less accurate than Type 2 processing for evaluation of specific aspects of performance.

In summary, a small amount of research suggests that the mental workload of making performance evaluations may be very high, especially in examination situations where performance is being observed and judged in a short concentrated period. It is likely that excessive mental workload results in less accurate cognitive processing and may contribute to interrater variability in evaluating performances. The development of expertise and use of Type 1 processing may be mechanisms through which mental workload of rating is managed. There are interesting avenues of further research to investigate these aspects. In addition, specific research is needed in relation to the mental workload of in-training evaluations because factors such as the complexity and stressfulness of the evaluation environment and

the duration of the evaluation period are very different to the standardised examination situations in which research has been conducted to date.

Translating a holistic narrative evaluation to a numerical score

Because their evaluation first seems to be developed in holistic and narrative terms, G. Regehr et al. (2011) suggested there may be a discrepancy between the representation supervisors construct and the way in which they are expected to document it. Evaluation forms present a set number of items and dimensions, categorised and organised in a specific way that may not match the cognitive framework of the rater. In addition, the aspects of competency they describe are overlapping and intertwined and not the mutually exclusive, separate items they appear to be on forms. It can then be a point of discussion whether, for example, an issue with poor ability to relate to a patient, falls under the domain of communication or professionalism, as Crossley and Jolly (2012) related. The level descriptors may also provide inexplicit benchmarks such as excellent, good, and satisfactory, or encourage comparison with other students, for example, by saying “above average”.

Crossley, Johnson, Booth, and Wade (2011) showed the improvement that could be gained by better aligning scale level descriptors with the concept of developing independence and expertise. They incorporated descriptors that related to how much supervision the student needed and found substantial improvements in reliability of scores on a variety of workplace-based assessment instruments. G. Regehr, Bogo, Regehr, and Power (2007) suggested that part of the problem of rater variability may relate to having numerical scales on evaluation forms. Their initial work in redesigning scales to be better aligned with the depictions and language used by supervisors when describing students was not as successful as that of Crossley et al. (2011). The scales did not differentiate students any more effectively, nor identify students with serious performance deficits. However, eliminating numerical scales and moving to system where supervisors matched performance to the closest vignette did improve the ability to identify poorly performing students. Numerical scales were also highlighted as a problem in the study by Kogan et al. (2011). Raters struggled to translate their judgement into numbers, particularly for the overall rating. Raters also reported finding it difficult to know what the numbers really meant in terms of the standard of performance and found scale descriptors inexplicit. There was a great deal of variability in the strategies used to derive an

overall numerical score which included averaging the aspects of performance, weighting aspects according to the purpose of the assessment, and using a non-compensatory system whereby a significant deficit in one area affected the evaluation of the whole performance.

The items and dimensions on the form may be only a subset of what supervisors consider. As I referred to when discussing supervisor expectations, studies have suggested that supervisors place importance on aspects that are not part of competency frameworks and unlikely to be reflected in items. Research that has examined the written feedback provided on evaluations similarly found references to aspects such as initiative, maturity, composure, and self-improvement where were not reflected in items (Frohna & Stern, 2005). Ginsburg, Gold, Cavalcanti, Kurabi, and McDonald-Blumer (2011) found references to the disposition, personality, or attitude of the student, the student's impact on staff, their level of improvement or progress over the placement, their level of performance in relation to peers, and predictions about their future. These aspects were very similar to those found in the interview studies as already discussed.

Ginsburg et al. (2011) also found that comments often spanned more than one dimension or item, which would make it hard for a supervisor to decide where to make their evaluation of that aspect of performance, and may result in them incorporating it into more than one item. Also, by presenting a list of items in a specific order, which are either explicitly or implicitly weighted, forms may not mirror the relative importance of aspects to a supervisor, as Govaerts et al. (2013) have observed. Indeed this would not be possible since the relative importance appears to differ depending on a variety of contextual factors such as the level of performance of the student. Hence it may be natural for the supervisor to work by mapping their holistic overall evaluation to the items on the form, as Yeates et al. (2013) found, rather than separately considering each aspect. Item intercorrelation may be the result of supervisors distributing their judgement amongst ill-fitting items, rather than indicating that supervisors consider few dimensions of performance.

These types of issues have prompted suggestions that it would be better to focus on capturing high quality narrative comments to supplement or replace numerical ratings (Govaerts & van der Vleuten, 2013; Hanson, Rosenberg, & Lane, 2013; Hodges, 2013). Hanson et al. (2013)

expressed the view that judgements are lost when translated into numbers, however there are conflicting research findings in regards to this.

Some research suggests that supervisors' narrative evaluations are well captured in scores, based on comparison of scores with rankings indicated by comments (Ginsburg et al., 2013) or how positive or negative the comments were (Frohna & Stern, 2005). Other research suggests that the narrative evaluation is not well captured in scores when students who are failing are considered. In qualitative studies of the feedback provided to medical students, both Cohen, Blumberg, Ryan, and Sullivan (1993) and Durning et al. (2010) found that comments more frequently indicated deficiencies than did the score awarded. Schwind et al. (2004) also found this to be the case in their study of the evaluations of medical students with serious performance deficiencies. Comments identified deficiencies far more often than scores, however the proportion identified was still small and contradicted by other evaluations indicating no deficiencies, or even excellent performance. This suggests that comments too may not well capture the supervisor's judgement and may be overly lenient. In another retrospective case control study[1] Guerrasio et al. (2012) found that students with serious performance deficiencies had a much greater proportion of evaluations with negative or ambiguous comments than students in good standing. All had received at least one evaluation with negative comments, however so had 71% of students with good standing.

These findings suggest that, while over the whole populations of students, scores capture much of the information present in comments, this is not the case for the minority of students with unsatisfactory performance. In those students, comments more often reveal performance issues than do the numerical scores, although not always. It may be that the issue with translation of narrative to numbers is greater for failing students than others. As discussed previously in the section on leniency-stringency, personal and relational issues as well as administrative and political issues may make supervisors reluctant to record what they really think in the numerical score. They may, however, provide a better indication of their evaluation in the comments, as indicated by a participant in the study by Kogan et al. (2011): "I

[1] In case control studies, cases with the outcome of interest are compared to a group of cases without the outcome of interest to look for differences that might indicate associations or causes (Elwood, 2007).

feel like I can express my dissatisfaction well in my comments without having to negotiate whether this was a [score of] 4 or 5 for the resident…" (p. 1055).

As previously mentioned, delays in rating can increase halo effects because they increase reliance on an overall impression rather than remembered details of aspects of performance (Murphy et al., 1993). Delays in rating are the norm for in-training evaluation. Paget et al. (2013) found that 50% of evaluations were completed more than 6 days after observation of the student. Furthermore, in-training evaluations usually cover performance over a long period of weeks to months, so the evaluation may include performances from a long time before rating. In an experimental setting Murphy and Balzer (1986) showed that delays in job performance evaluation of only one day increased item intercorrelation, but did not reduce the accuracy of ratings. Sanchez and De La Torre (1996) also found no decrease in the overall accuracy of job performance evaluations with delays, but showed that delays decreased the detail of strengths and weaknesses recalled and therefore would reduce the amount of specific feedback that could be given. Consistent with this, in a study of the operative performance evaluations of senior surgical students, R. G. Williams et al. (2014) found that 14 day delays in rating significantly reduced the amount of variation in item scores but that there was no substantial difference in mean scores. Delays in rating also substantially increased the proportion of evaluations with no comments or only non-specific comments. These effects of delayed rating are also consistent with supervisor comments that specific feedback was hard to give unless done immediately (Weijs et al., 2016) and concerns of supervisors about the accuracy of evaluations where recording was delayed (Dawson, Miller, Goddard, & Miller, 2013).

Together, these studies all support the idea that when rating is delayed, supervisors accurately recall their overall evaluation but not the detail of it. However Paget et al. (2013) did find an effect of delay on decreasing scores on in-training evaluations in a medical school. They proposed that this was an effect of leniency bias dissipating with time. Keeping notes has been recommended to help improve recall (R. G. Williams et al., 2003), as there is some evidence that diary keeping was associated with less lenient ratings and better discrimination of dimensions (DeNisi & Peters, 1996).

The research reviewed here suggests that the translation of a rater's evaluation from a holistic narrative to an overall numerical score on a rating scale and a series of item scores is problematic and may result in some variability between raters that is not part of the variability in their evaluations. As well as personal, relational, administrative, and political influences that may affect willingness to accurately represent their evaluation, poor fit of evaluation forms to cognitive frameworks and poor recall because of delays in recording, may contribute to issues with translation. These other factors may not affect accuracy of the overall grade, but may increase interitem correlation and decrease the amount of specific feedback given to students. Therefore, the value of the evaluation as a summative assessment may be unaffected, but its value as a formative assessment decreased. Researchers have shown that adjusting scale level descriptors to be better aligned with rater conceptions may improve the reliability of in-training evaluation scores and further research on the cognitive frameworks and processes of raters may improve our ability to align scales. Our conceptions of competency may also need revisiting and adjusting to improve alignment. Encouraging prompt recording and diary keeping may help supervisors provide more specific feedback, but may have limited effect if workload precludes sufficient time being spent on the evaluation process.

Summary of the process of evaluation

The literature reviewed to this point suggests that having formed a picture of student performance based on their observations and inferences, supervisors evaluate this in the light of their expectations. This process involves making further inferences about such things as the student's intentions, motivation, preparation, attitude, personality traits, and so on, that help the supervisor understand their observations. They consider the context of the student's performance, including the difficulty of the cases seen. They consider multiple interlinked dimensions of the student's performance at once, and weight these in ways that are different for every student and circumstance. With experience, they are able to integrate more contextual aspects, and make more inferences and interpretations. They develop a holistic and qualitative evaluation that balances these factors, and may include aspects that are not normally considered in competency frameworks or part of the evaluation criteria. In translating their judgement to numerical overall and item scores, they may be influenced by their own context and the context of the evaluation, including the consequences of the score they award for the student, the public, and themselves. It may be difficult for them to give very specific feedback if there is a long delay in recording their evaluation, but their overall

judgement is usually recalled well, if somewhat less leniently than it otherwise might have been.

Several aspects of the rating process appear to contribute to the variability of evaluations between supervisors (rater subjectivity), differences in leniency, and item intercorrelation (halo effect) that are seen psychometrically. The influences on rater judgement are highly personal and context specific. Cognitive processes, such as the mental schema used, also contribute variability, and are further influenced by context, mental workload, emotions, and rater expertise. Because each supervisor's cognitive representation may be different, it may not fit well with the items and descriptors on the form and lead to differences in the way evaluations are translated to scores. The personal response to the process of evaluation and to the administrative and political pressures and other contextual factors will be highly variable, even for one supervisor at different times, and may contribute to variability in leniency. How different supervisors make allowances for the busy stressful environment of evaluation, or the fact that they have not observed the student, may also produce differences in leniency.

Raters appear to form a holistic judgement that is likely, at least initially, impression-based and involving Type 1 processing, especially under high mental workload when multiple contextual elements need to be considered. While many different aspects are evaluated, the aspects are highly interwoven and dependent on each other, therefore it seems likely that intercorrelation between items is at least partly due to construct-relevant relationships between the aspects considered, rather than halo error. Intercorrelation may also be increased by delays in rating, which reduces the detail recalled while preserving the overall judgement.

The in-training evaluation in veterinary medicine

The research on assessment in veterinary education is very limited. A systematic review of the pre-2007 literature found only 51 papers that concerned veterinary assessment and of these, only five evaluated the assessment method (Rhind et al., 2008). None of the five papers studied in-training evaluations. Since then a further 47 papers have been published concerning veterinary assessment and of these, 14 evaluated the assessment method in some way. Six of the 14 papers studied the in-training evaluation. Although the large body of research in

medical education helps support practice in veterinary education, there is clearly a need for more research that specifically addresses assessment in this context.

This section will briefly review the six papers that concern the in-training evaluation in veterinary education in relation to what they reveal about rater judgement processes.

Evidence of leniency

The study by Walsh et al. (2012) gave some descriptive data that documented significant leniency in scores, with 81.63% of overall grades being *high quality* or *outstanding*, both of which are above the scale midpoint of *good*. Very few evaluations gave an overall grade of *unsatisfactory* (0.01% of evaluations). One item on the evaluation required supervisors to indicate the student's degree of progress towards graduating competency. More supervisors indicated concern about the student's progress (0.25% of evaluations) than awarded an unsatisfactory grade. The researchers proposed that because the overall grade contributed summatively to the students' final mark, supervisors were more lenient in scoring this but that they flagged potential issues using the item relating to progress towards new graduate competency. Walsh et al. (2012) also reported that return of evaluations to students was usually completed within 4 weeks. This suggests that there was often a significant delay of weeks. This study therefore provides some descriptive evidence that both leniency and delay in ratings are also seen with veterinary in-training evaluation.

Internal structure

One study has investigated the internal structure of an in-training evaluation to determine the dimensions assessed (Fuentealba & Hecker, 2008). The instrument comprised 21 items spanning four domains of knowledge, clinical skills, interpersonal skills, and professionalism. Using exploratory principal components analysis with orthogonal (varimax) rotation, the researchers concluded three factors were present, based on eigenvalues of greater than one and the amount of variance accounted for. The items comprising each factor suggested the dimensions represented concepts of professionalism, knowledge, and clinical skills. Factor scores were calculated by simple summation of equally weighted item scores for each factor. There were significant and substantial correlations (0.62-0.76) between the scores for each

factor. This study therefore presents evidence that supervisors can discriminate three dimensions of competency of veterinary students but that the dimensions are highly intercorrelated, suggesting there may be an underlying higher-order structure.

Relationship with other assessments

Two studies investigated the relationship of in-training evaluation scores to other assessment scores. Bateman et al. (2008) compared in-training evaluation scores with scores from an objective structured clinical examination (OSCE). They examined the agreement between scores on the two instruments for the very lowest and highest performing students using the kappa statistic and found little if any agreement (all values less than 0.2, whereas substantial agreement is indicated by values of at least 0.6). Possible reasons for the poor agreement identified by the researchers were real changes in student performance between the time of the placement and the end of year summative OSCE assessment, or rater errors due to consideration of construct-irrelevant factors. There were also some differences in what the two assessments were designed to assess, with professional conduct assessed in the in-training evaluation and not in the OSCE.

Roush et al. (2014) compared in-training evaluation scores with grade point average (GPA), class rank, and scores on the national veterinary licensing examination (NAVLE) using Pearson's correlations. In-training evaluation item scores were weakly to moderately correlated with GPA from the final pre-clinical year (0.25-0.44), but were not significantly correlated with GPA from the end of the clinical year. They were moderately negatively correlated with class rank (-0.43 to -0.55) and weakly to moderately correlated with scores on the licencing examination (0.26 to 0.42). These assessments are likely to have some difference in what they are designed to assess. For example, the veterinary licencing examination contains only a small amount (4-7%) of content related to the in-training evaluation items of zoonosis and biosecurity, and of client communication and ethical conduct (National Board of Veterinary Medical Examiners, 2015), so the poor correlation found between its result and the scores for these items is not unexpected. Correlation was much higher for items with content that was extensively covered in the licencing examination. The researchers also found strong intercorrelation between items on the in-training evaluation (0.52-0.94).

These studies indicate low to moderate relationship with other assessments commonly used in veterinary medicine, but there were differences in the concepts the instruments were designed to assess and therefore this does not necessarily indicate poor performance of the in-training evaluation. It suggests the in-training evaluation may be capturing aspects of performance the other assessments do not. The study by Roush et al. (2014) provides additional evidence of high item intercorrelation in the veterinary in-training evaluation.

Relationship of in-training evaluation scores with conceptions of and approaches to practice

Another series of studies investigated the relationship between in-training evaluation scores and veterinary student conceptions of and approaches to learning while on clinical placements and, later, their conceptions of and approaches to practice as new graduates (Matthew, Ellis, & Taylor, 2011; Matthew et al., 2010). The studies used a phenomenographic approach and mixed methodology. Although the purpose of these studies was not to evaluate the in-training evaluation, the researchers found that scores on in-training evaluation were predictive of higher quality conceptions and approaches, likely to result in better success in learning and in practice. Higher scores were significantly related to a cohesive conception of clinic-based learning including awareness of its complexity and contextual variation, with a large effect size of 0.70, and to a highly engaged and deep approach to learning that focussed on understanding and developing a holistic view of practice, with an effect size of 0.46 (Matthew et al., 2010). Higher scores were also significantly related to relational conceptions of practice, with a commitment to excellence and awareness of the interlinked aspects that contribute to professional decision making, with a large effect size of 1.06 (Matthew et al., 2011). There was also a nonsignificant tendency for association of scores with a reflective approach to practice, involving a commitment to reducing errors and application of holistic personal values to guide all aspects of performance. Taking part, accepting responsibility, and wanting to understand were key aspects to higher quality approaches. Higher scores on in-training evaluation were awarded to those with relational conceptions than those with multistructural conceptions.

These findings give significant insights into the constructs captured in in-training evaluation. They also provide evidence that supports the validity of in-training evaluation scores as indicators of potential success in clinical practice. If further research demonstrates that these

conceptions and approaches are of high value in ensuring success in clinical practice, then this would further strengthen the validity argument. In addition, the findings suggest that supervisors take account of how students integrate different aspects of competency, such as their medical knowledge, contextual elements, and the perspectives of others. Focusing on how aspects of competency are holistically combined may result in correlation between the aspects of performance captured in individual items of the in-training evaluation. This suggests that item intercorrelation is both expected and construct-relevant.

Student and supervisor perspectives on the in-training evaluation

Two studies have examined supervisor and student perspectives of veterinary in-training evaluations. Weijs et al. (2016) studied perspectives about the feasibility and utility of the in-training evaluation using focus group interviews. They compared the in-training evaluation to two other workplace-based instruments: the miniCEX and DOPS. Both supervisors and students found the in-training evaluation less valuable than the other two assessment methods; however, students thought it had the potential to be more valuable when personalised and specific feedback was incorporated, and that it had the advantage of giving a good overview of performance. Supervisors found it time consuming and logistically difficult to complete because of the need to collaboratively involve various instructors in finalising the evaluation. They found it difficult to provide personalised and specific feedback and felt that often this was just repeating comments they had already made verbally to the student. However, they noted the in-training evaluation was potentially valuable for clarifying areas that needed development. The researchers cautioned that the role of the in-training evaluation within a programme of assessment needs careful consideration of its formative or summative role, benefits and resources required.

Dawson et al. (2013) studied the impact of in-training evaluation on student learning and supervisor practices using focus group interviews with students and a survey of supervisors. The findings indicated that students perceived the in-training evaluation to be unreliable as it seemed unable to differentiate different levels of performance and they did not believe that supervisors took the process seriously. Consequently, they generally did not value the in-training evaluation, and tended to ignore the results, which usually came too late to be of benefit. Specific feedback about even one aspect of their performance improved the value of

the in-training evaluation to them. It seemed this reassured them that the supervisor was specifically remembering their own performance, and provided a justification for the scores awarded that made them more acceptable.

In contrast, most supervisors felt the in-training evaluation was reliable and that another faculty member would give the same or close to the same score, however concerns were raised about the accuracy of scores when recording of the evaluation was delayed. Supervisors identified strengths of the in-training evaluation in informing instruction about what students should be learning, improving objectivity of assessment, and holding colleagues accountable for assessment. The most frequently cited weakness was that it was too time consuming. There were also concerns that it inhibited utilisation of spontaneous teaching moments. The scale used in the in-training evaluation was based on levels of independence, and some supervisors commented on the difficulty in accounting for the stage of training when grading students. If the student was awarded a low grade simply because they were inexperienced and not yet independent, they felt this would affect the student's self-esteem and their own teaching evaluations.

These two studies indicate that there are significant issues with the perception and operation of the in-training evaluation, especially in regards to its formative capacity. Both studies noted that supervisors found it time consuming.

Comparison with self-assessment

Root Kustritz et al. (2011) compared students' self-assessed in-training evaluations with that of supervisors in a small animal placement. Supervisor scores were significantly lower on average than student scores and had a greater variance. Students with higher performance levels tended to under-rate their performance and students with lower levels of performance tended to over-rate their performance.

Summary of research on the veterinary in-training evaluation

Although there is very little research on the in-training evaluation, the findings of these studies indicate that issues are similar to those seen with in-training evaluation in medical education.

Leniency and delays in evaluation are issues in common. There is not always a strong relationship between scores on in-training evaluation and scores on other assessments, but the findings tend to suggest that the assessments are evaluating different aspects of performance by design and therefore this may not indicate construct-irrelevance. A lack of high quality formative feedback reduces the formative value of the in-training evaluation but also reduces the confidence students have in the scores awarded. However supervisors find it time consuming to give such feedback.

Although, as I reviewed, many studies in medical education have found that in-training evaluations assess only one or two dimensions, the single study done in veterinary medicine suggested three dimensions could be distinguished by supervisors. The finding of significant intercorrelation between factors also suggests a higher-order factor structure may be present which has not been investigated. Other research has given us important insights into the constructs being assessed in the in-training evaluation that suggest that it taps into conceptions and approaches to learning and practice that enable students to make deep connections and accommodate contextual and social complexity in clinical work.

The scarcity of studies on veterinary assessment indicates that there are many open questions and areas of research need.

Conclusion

The importance of assessing veterinary competency derives from a global focus on outcomes and accountability. However, it is an ill-defined and complex construct that involves abstract and unobservable aspects. Assessing competency is therefore a challenge and requires judgement. The in-training evaluation is a widely used but highly criticised instrument for assessment of competency. Variability that arises from rater judgement processes and contextual differences is a major theme, but there are also questions about what is being assessed that compromise the validity of scores. Research on rater judgement processes and the influence of context give us some insight into the mechanisms of rater and contextual variability and the interrelationships between dimensions assessed. However, very little of this research has been done in a veterinary context.

Chapter 3:
Research design

This research investigated how well the scores on in-training evaluation captured the aspects of veterinary student performance it was intended to assess, because this has been an area of criticism of the instrument. The research reviewed in Chapter 2 revealed competency to be a multidimensional but holistic concept. A conceptual framework, developed from the research literature, suggested that judgement of competency involves the supervisor comparing their view (or "picture") of the student's performance with their own expectations, in a holistic narrative judgement before translation to scores. A mixed method research design provided a way to gain insights into the supervisor's view and the constructs underlying scores so that aspects of this complex process could be illuminated. A critical realist perspective informed the research design, with an underlying constructivist epistemology and realist ontology.

In this chapter, I first present my theoretical perspective and discuss its inherent assumptions. The overall research design, or methodology, is then discussed, as well as how threats to validity and ethical concerns were managed; however, details of the methods used are not presented in this chapter. Instead, details of the methods for each phase of the research are presented separately within each of the next four chapters along with the findings of the research. This facilitates discussion of methods that depend on the findings of previous phases, as occurs in Phase 3 and Phase 4. It also facilitates interpretation of the findings of each phase in the light of limitations in the methods. The final part of this chapter introduces the research setting in which the phases of research took place.

Research perspective

My philosophical stance involves a combination of realist ontology and constructivist epistemology. It is thus consistent with the theoretical perspective of critical realism (Maxwell & Mittapalli, 2010). Critical realism takes the view that things exist independently of us experiencing them or knowing of them, but accepts that our knowledge of reality is uncertain (Outhwaite, 1987). Different people can hold different perspectives and concepts of reality.

Furthermore, since our concepts and perspectives are part of the structure of the world we live in, each of us experiences reality differently (Maxwell, 2012). More than one account of any phenomenon can be valid, however not just any account is valid. Reality constrains the conceptions that can be held true. Theories must be related to reality such that they are consistent with or can accommodate observations, but that does not mean that theories correspond exactly with reality.

Critical realists hold that mental phenomena, such as meanings, intentions, beliefs, motives, reasons, and so on, are real (Maxwell, 2012). This means that mental states and attributes can explain other phenomena but also be influenced by other phenomena, including the physical world. Because they are real, we are able to study people's concepts and perspectives of them, and the effects of their situation and actions on them. Critical realism is compatible with both the study of the quantitative aspects of causation (whether one variable causes differences in another variable) and the qualitative processes of causation (how one variable causes differences in another) (Maxwell, 2004; Maxwell & Mittapalli, 2010). For critical realists, knowledge building involves developing an understanding of how and why things happen, including what enables them to happen and what prevents them from happening (Outhwaite, 1987).

This focus on explanation and on understanding mechanisms makes critical realism a useful stance from which to pursue the goals of my research. A realist ontology allows for supervisors' observations, perspectives, cognitive processes, and feelings to be considered independent from me and able to be studied. It allows for supervisors' observations, perspectives, and feelings to influence and explain their cognitive processes and the evaluation they make. It supports statistical procedures such as factor analysis and regression which assume that latent variables are real and can influence other variables (Borsboom, Mellenbergh, & van Heerden, 2003).

A constructivist epistemology allows for the meaning that I create with this research to be a subjective interpretation, although based on careful and sufficient observation and constrained by the need to accommodate those observations in a credible manner. It also allows for the observations supervisors make to be perceived and conceived subjectively by them and given meanings that may differ from those of another observer. Varying views of

student behaviour and competency, as well as differences in context and circumstances, will naturally lead to variation in the value of that performance to the supervisor and the scores they award. Scores are not "true" properties of the real performance, but reflections that bring together supervisors' values with the performance they observe. This is the uniting of subjective and objective as described by Crotty (1998) as a feature of constructionism. Critical realism thus provides a consistent position from which to interpret scores as representations of a real performance and subject to social influences that affect how well the scores represent that performance (Wong, Greenhalgh, Westhorp, & Pawson, 2012).

In articulating a particular ontological and epistemological position, this stance contrasts with a pragmatic stance from which mixed method research is commonly undertaken. A pragmatic stance is one in which researchers do not adopt any particular theoretical perspective to inform the choice of methods, but use the combination of theories, perspectives, and methods best suited to answer the research questions (Johnson & Onwuegbuzie, 2004). Less importance is placed on the underlying philosophy. More importance is placed on the consequences and practical importance of the research (Cherryholmes, 1992). It is the practical use of theories rather than their "truth" that is an indicator of their value (Johnson & Onwuegbuzie, 2004). A pragmatic stance is similar in many ways to a critical realist stance including being consistent with there being different perspectives of reality and more than one account of a phenomenon or theory that is valid (Johnson & Onwuegbuzie, 2004). A stated advantage of a pragmatic approach is that it does not prioritise knowledge gained by any particular method (Biesta, 2010). However, Maxwell and Mittapalli (2010) argue that the underlying assumptions associated with particular types of analysis inevitably influence the researcher and may make it hard for researchers not to prioritise findings produced in certain ways.

Because a critical realist stance more closely mirrors my own perspective than any other, it was more appropriate to operate from that perspective and acknowledge the underlying assumptions. These assumptions, regarding the nature of reality and knowing, were compatible with combining quantitative and qualitative methods in a mixed method study (Maxwell & Mittapalli, 2010).

The mixed method research design

This research used a combination of qualitative and quantitative methods because this suited the research aim of investigating the constructs underlying scores and their relationship with what supervisors value and what the assessment is intended to assess. A mixed method approach enabled the combining of qualitative and quantitative methods synergistically to elaborate, enhance, clarify, and enrich the understanding gained from each. This purpose was described by Greene, Caracelli, and Graham (1989) as complementarity. It involves simultaneously examining different facets of the same phenomenon using different methods. Greene et al. (1989) likens this to "peeling the layers of an onion" to gain a deeper understanding of the phenomenon, and differentiates it from triangulation in which the different methods are chosen to complement the strengths and weaknesses of each and are applied independently to examine the same phenomenon. Complementarity is in keeping with the aims of critical realist research to develop explanation and understanding of the mechanisms behind occurrences (Outhwaite, 1987). Thus, a strength of the mixed method approach was the capacity to give a more complete picture and to highlight areas of either of convergence or contradiction which may indicate areas for further study of assessment of veterinary competency.

This mixed method design for this research had four phases (Figure 3.1). The first and fourth phases were qualitative descriptive studies and the second and third phases were quantitative. Each phase of the research addressed one of the four research questions described in Chapter 1. The use of qualitative and quantitative methods was governed by the nature of each question as I now explain. A detailed discussion of the individual data-gathering methods used, and their justification is provided later, when the research methods are discussed, in each of the next four chapters.

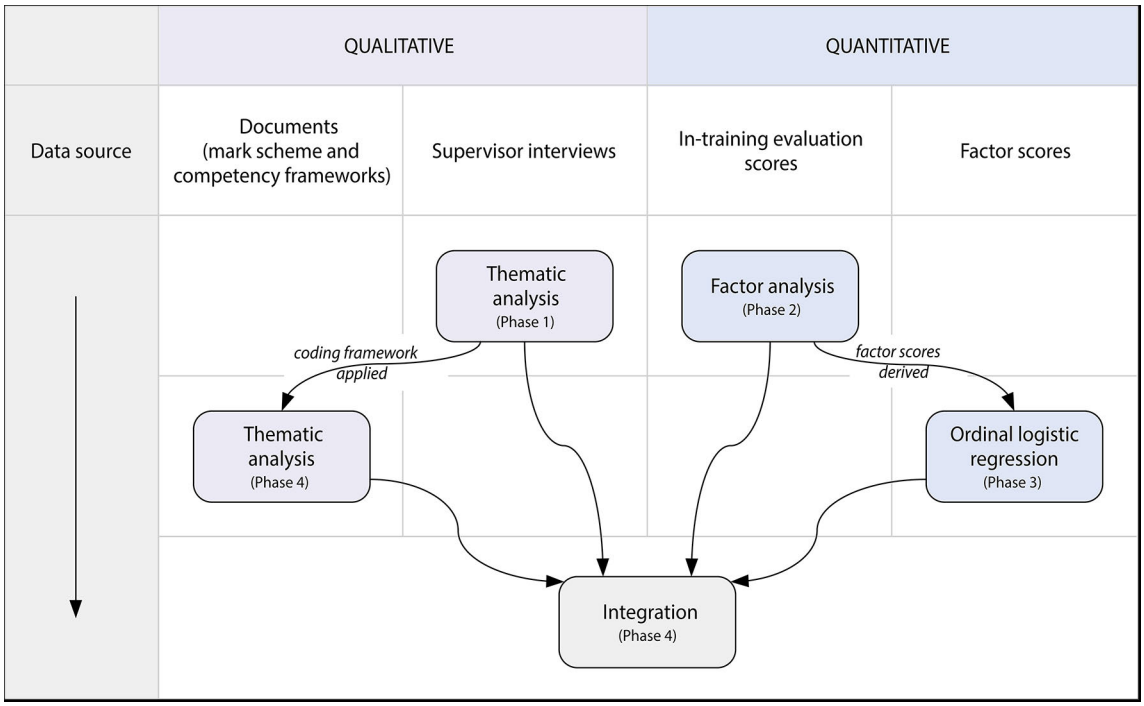


Figure 3.1: Sequence and relationship of qualitative and quantitative research phases designed to investigate how well the scores on in-training evaluation capture the aspects of student performance the instrument is intended to assess. Separate research questions are addressed in each of the four phases.

Phase 1 addressed the following research question: (1) What qualities of performance do supervisors value when making judgements about the competency of veterinary students on placements? Semi-structured interviews with supervisors were used to generate a rich descriptive picture, and then thematic analysis performed to investigate the range of aspects described and their relative importance. Qualitative research is especially useful for investigation of a participant's perspective (Maxwell, 2013) and thus was appropriate to investigate the views and values of supervisors.

Phase 2 addressed the following research question: (2) What is the nature of the dimensions captured by the items on in-training evaluation? This quantitative phase involved exploratory common factor analysis to identify and characterise the correlations between items on the in-training evaluation, as these suggest underlying constructs. Quantitative research is especially useful for identifying relationships between variables and their strength through statistical analysis (Creswell, 2012) and thus was appropriate for this phase of the research.

Phase 3 addressed the following research question: (3) How do the dimensions captured relate to the overall grade? This was another quantitative phase that was sequential to Phase 2 because it utilised data from it. Scores were derived for each dimension found in Phase 2 and then ordinal logistic regression was used to investigate how the dimensions related to the overall grade awarded. A quantitative approach enabled statistical techniques to be used to determine the strength of relationship between variables and was therefore appropriate for this phase of the research.

Phase 4 addressed the following research question: (4) What is the relationship between what supervisors value, the dimensions captured by the evaluation, and the competency frameworks? This was a final qualitative phase that integrated all four phases of the research. It involved thematic analysis of the mark scheme for the in-training evaluation and two relevant competency frameworks using the coding framework developed in Phase 1. The themes arising were then used to interpret and clarify the quantitative findings from Phases 2 and 3 and compare them, qualitatively, to the findings of Phase 1. This was an appropriate method of comparing the constructs informing scores with the constructs important to supervisors and the constructs of competency.

This was thus a complex mixed method design in which qualitative and quantitative phases were of equal importance, and converged in Phase 4, where the data from all four phases was integrated. Greene et al. (1989) recommended that methods should be implemented simultaneously for the purposes of complementarity. In my research, the timing of Phases 1 and 2 was theoretically concurrent, but determined pragmatically by practical factors such as availability of interviewees and datasets, and by the fact that I could only do one at a time. Phase 3 was dependent on data from Phase 2 and so was necessarily conducted after it. Phase 4 was performed last.

A mixed method approach was therefore well suited to the aims of this research, which involved both the consideration of supervisors' perspectives and their numerical representation in scores.

Validity of the research

Validity relates to the quality of research (O'Cathain, 2010). Both the use of the term validity and its meaning are debated, but I adopt the definition of Messick (1989) whose definition, given in the context of test score validity, was extended to other research by Dellinger and Leech (2007) thus:

> an overall evaluative judgment of the extent to which empirical evidence and/or theoretical rationale support the adequacy and appropriateness of interpretations and actions on the basis of data generated through any means. (Dellinger & Leech, 2007, p. 316)

Validity is an interpretation and a value judgement about how well the claim(s) being made are supported. A useful framework for thinking about the quality of evidence is outlined by O'Cathain (2010), based on the work of a number of researchers, which involves assessing the quality of research planning, design, data, analysis, interpretation, transferability, reporting, synthesisability, and utility. In evaluating the strength of claims, the areas that deserve most attention are those that are most likely to threaten the plausibility of the specific research project (Maxwell, 2013). For this research, there were several important threats including researcher bias, reactivity, sampling adequacy, missing data, and issues with data distribution. I have detailed how each of these was managed in the methods sections for each phase of research and discussed their implications when presenting the findings in the next four chapters. Here I provide some detail of my own background and influences, which is relevant to the researcher bias and reactivity threats.

I began this project because of my fascination with assessment. As a teacher I have found judging student work difficult and troubling. I have worried about being subjective and biased. I have worried about giving students the right level of challenge and determining a benchmark for comparison. I have worried about assessing fairly and giving good formative feedback. Through my reading to learn more about these issues, I discovered that the answers to these problems are not “out there” and that our assessment system in higher education is far from perfect. Statements like this from Edwards and Knight struck a chord.

> That schemes for the assessment of competence in higher education students are riddled with contradictions, problems and flaws there can be no doubt. (Edwards & Knight, 1995, p. 11)

I knew that I wanted to research this flawed system, to try to find solutions. Clearly then, I am invested in this research and brought to it preconceived ideas about the judgement process, informed by my own experiences and my reading. There was great potential for me to find what I expected in all phases of the research, but especially during the thematic analyses.

I took a careful and systematic approach to limit the effect of this as much as possible. I used verbatim transcripts to analyse interviews which I had checked for accuracy against the recordings and which participants had checked for accuracy and corrected if necessary. The audio recordings were accessed during coding to clarify participant meaning. I coded systematically and openly, and included all key ideas in all transcripts before determining patterns and themes. I utilised matrix summaries to cross check the coding and ensure it represented the data well. In summarising the findings, I provided sufficient detail as well as direct quotes, so that the conclusions are, I hope, transparent to readers. I discussed the findings and interpretations with my supervisors.

A strength I brought to this phase of the research was my experience as a clinical veterinarian and teacher, and my familiarity with the setting and supervisors. This gave me insider knowledge into the meaning of terminology and slang in common use; the way veterinary practices and practice-based teaching work; the type of work the students would be involved with; the conflicts between animal, client, and student that the supervisor must balance; and the administrative systems in place. This knowledge helped me derive meaning from the data.

As an insider, I also knew most of the interview participants as many worked in the same university department as I did. None were people I worked closely with, but they were all professional colleagues, even those I had not met. This facilitated a relaxed and open interview atmosphere, as did interviewing them in their own offices and workplaces. I used pre-prepared open questions (Appendix A, page 251) and asked for explanations and examples to try to ensure that I understood the participant's meaning. The interviews were audio recorded and I also took notes while interviewing.

Thus, although researcher bias and reactivity can never be eliminated, I was aware of their possible influence and used careful and systematic methods to limit their effect where possible. I also documented findings in a way that enables the reader to draw their own

conclusions. In these ways, I sought to support the strength of evidence for the claims made in this research.

Ethical considerations

I considered several ethical implications in conducting this research and took steps to minimise harm. The implications were: the use of previously collected data; issues of informed consent; maintenance of confidentiality and anonymity; risks of harm; ownership of data; and reciprocity (Punch, 2006). Risks of conflict of interest were minimal because I worked in a different part of the department, I was not a line manager of any member of the source population for interviews, and I was not involved in teaching or assessment of any undergraduate veterinary students at the time this research was conducted. I sought advice from the Kaiarahi Maori, College of Sciences, who identified no special risks for Maori participants.

The study involved the use of data in the quantitative phases of the research that had been collected for another purpose: that of assessment of students. No explicit consent for the use of the data for this purpose was given by the students and supervisors represented by the quantitative data. In order to minimise the risks associated with this aspect, specific consent to access the data was sought, and obtained, in writing from the Director, Student Management, Massey University (Appendix H, page 300). The data was also de-identified before I accessed it.

Consent was obtained in advance to collect the qualitative data. Interview participants were informed verbally and in writing of the research purpose (Appendix K, page 306), and participation was voluntary. Written informed consent for participation was obtained (Appendix L, page 308), and participants had the right to decline to answer any question or withdraw from the study at any point until analysis of the data was undertaken. Pseudonyms were used on the transcripts to protect the identity of individual interview participants and the transcriber signed a confidentiality agreement (Appendix M, page 309). After transcription, the transcripts were returned to participants who had the right to amend or edit any part, and they signed an authority to release the transcript once they had done so (Appendix N, page 310). All interview recordings, transcripts, student evaluation results and other data was securely stored.

Risks of harm arising from this research involved the reputation of the University, teaching hospital, its academic staff, and students. There was the potential for the research to suggest that previous and current assessment practices were not valid measures of student performance. The risk of harm to individual staff members and students was minimised by the maintenance of anonymity and confidentiality. The risks of harm to the reputation of the institution was minimised by the fact that any problems found were likely to be similar to those reported for many other institutions across a range of disciplines including medicine, and also likely to be typical of those found in other veterinary schools. In addition, the research findings had potential benefits to the institution, staff, and students in suggesting ways to improve practice. These risks were discussed with the Programme Director and the Head of Institute who were both supportive. Written permission to conduct the research, and name the institution, was sought and obtained from the Pro-Vice Chancellor of the College of Sciences and the Head of Institute of the Institute of Veterinary, Animal, and Biomedical Sciences at Massey University (Appendices I-J, pages 302-305). Approval for the research was obtained from the Massey University Human Ethics Committee before proceeding (Massey University Human Ethics Committee: Southern B, Application 13/94, Appendix G, page 299).

Setting for the research

The research was conducted in the veterinary school at Massey University in New Zealand where I am an academic. New Zealand has only one veterinary school. The veterinary degree at Massey University is a five-year Bachelor of Veterinary Science (BVSc). Upon graduation, Massey University BVSc graduates are entitled to be registered to practice as a veterinarian in New Zealand with no further assessment. Because the Massey University BVSc is internationally accredited, graduates are also entitled to practice, with no further assessment, in other jurisdictions such as Australia and the United Kingdom. Thus, Massey University's final assessment of veterinary students serves both to award a degree and to certify professional competency.

The final year of the programme is a clinical year in which students spend the entire year in workplace-based placements in different types of veterinary practices. The veterinary degree at Massey University allows partial streaming in the final year. Students must complete an 18-week core of placements that span all veterinary subdisciplines, but may then focus their study

on their area(s) of interest for the remaining placements. Most placements are 1-2 weeks in duration. Some placements are held within clinical practices and discipline units within the University and some are held in external veterinary practices. A core group of external practices is contracted by the University to supervise students, and their supervisors have close links with the University and receive training on teaching and assessment. About half of a student's placements are done at the University or in these core external practices. The remainder are done at practices of the student's choosing which may include core external practices, other practices that infrequently host students, and also may include practices overseas. There are thus distinct differences in background of the various supervisors, and a broad range of experience in clinical teaching and assessment of students.

The assessment for the final year comprises a mixture of examinations (including written, practical, oral, and computer simulation examinations), assigned work such as case reports and presentations, and workplace-based assessment. Two types of workplace-based assessment are used. One is a checklist of skills that must be completed. The other is an in-training evaluation, which is the subject of this research.

The in-training evaluation is conducted by supervisors at the end of every placement. Students are therefore assessed approximately 24 times over the course of the year. Practice varies between placements as to whether the evaluation represents the opinion of one supervisor, is collaboratively produced by several supervisory staff, or is an aggregated summary of independent evaluations performed by several supervisory staff. Which of these systems is used on a placement is not recorded. Evaluations for placements held at the University, or in core external practices, are captured electronically using a web-based portal. Evaluations made in other locations are captured on paper and entered into the electronic database by administrative staff. Supervisors are encouraged to complete evaluations promptly at the end of a student's placement. Many complete evaluations on the last afternoon or the following working day, however there is a wide variation in practice and some evaluations are very delayed.

At Massey University, most placements utilise the same in-training evaluation instrument, which is detailed in Chapter 5. It was locally developed and has 12 Likert-type items arranged in subscales spanning four domains of veterinary competency. All supervising veterinarians are

provided with information about the purposes of clinical placements, the expected performance of students, and the domains to be assessed, but specific rater training and calibration is not undertaken. Five levels of performance (excellent, good, satisfactory, marginal, and fail) are specified in items, with descriptors provided for each level in each item. An additional overall grade item, also ranging from excellent to fail, is completed by supervisors and has no level descriptors. No guidance is provided about how individual items should be weighted in contributing to the final overall score. This is left to the holistic judgement of the assessor.

The overall grade is converted to a pre-specified numerical value that contributes a small percentage to a student's final mark for the year. Marginal and failing overall grades require remediation by repeating the placement, and can prevent students from graduating. Therefore, the in-training evaluation at Massey University is a high stakes assessment for students. The in-training evaluation is also used formatively and supervisors are encouraged to provide formative feedback in relation to each subscale and overall. However, the focus of this research is on the summative aspects of the assessment and the meaning of scores.

Summary

In this chapter, I have presented the critical realist theoretical perspective that underpins this research, and justified the use of mixed methodology to address the research questions. I have given an overview of the research methods used, but the detailed discussion of these and their justification is left for the subsequent chapters where it is presented along with the research findings. I have discussed two of the threats to the validity of the research, being researcher bias and reactivity, but other threats such as sampling adequacy, missing data, and issues with data distribution are discussed in detail during discussion of the methods in each of the next four chapters. I have discussed how the chief ethical concerns of the use of previously collected data, issues of informed consent, maintenance of confidentiality and anonymity, and reputational risks, were managed to minimise their impact. Finally, the setting for the research has been detailed as a prelude to the four chapters which now follow, in which the details of the methods used and the findings for each phase of the research are provided.

Chapter 4:
What supervisors value

Findings of interviews with supervisors—Phase 1

The aim of this phase of the research was to gain insights into what supervisors value in veterinary student performance. As conceptualised in the framework presented in Chapter 2, the picture supervisors develop and their expectations of student performance form early steps in the process of judging performance. How these are translated into scores on in-training evaluation and how they align with the intentions of the assessment can then be investigated in later phases of this research, to inform our understanding of the process of judging the competency of veterinary students, and the validity of the scores. Individual interviews were therefore used to explore supervisors' depictions of students performing at different levels of competency. The design mirrored research that has been conducted in medical education and social work education (Bogo et al., 2006; Ginsburg et al., 2010). These previous studies gave insight into the range of aspects considered by supervisors, and how these were weighted and balanced for students of differing performance levels in these disciplines. They revealed that not only did what was considered differ for different students, but that it was not always well aligned with the intentions of the assessment. Therefore investigating what veterinary supervisors value is an important first step in investigating how well the scores on this in-training evaluation instrument capture the aspects of student performance we intend to assess.

Research method

Interviews

Semi-structured interviews were used to collect data for this phase of the research. Interviews were an appropriate method for collecting information about people's views and perspectives (Creswell, 2012; Merriam & Tisdell, 2016). Open questions allowed supervisors to respond in the manner they wished, which therefore helped give a sense of the relative importance of the

aspects they discussed. The semi-structured format enabled me to phrase and rephrase as necessary to help understanding of the questions. It enabled me to redirect, clarify, or seek more detail, if needed, to help my understanding and so that I gained as full a picture of their views as possible. Individual interviews were most suitable because the supervisors were likely to feel comfortable expressing their views one-on-one and because the purpose was to understand individual views in-depth.

The interview protocol (Appendix A, page 251) was adapted with permission from that used by Ginsburg et al. (2010) with supervisors of medical students. It approached the research question indirectly by asking supervisors to describe students they had previously encountered that they considered to have been excellent, weak, and marginal on placements. This was a way of getting a sense of what supervisors valued, without asking them directly about it. It also separated the context of the questions from the context of the in-training evaluation instrument, and from any competency frameworks. The aim was to tap into what supervisors considered and valued, rather than what they thought they were supposed to consider and value. This meant the themes arising were not externally derived, but those in use.

The wording of questions was changed slightly from the protocol used by Ginsburg et al. (2010) so that it applied to veterinary students. Another change was to ask for descriptions of marginal students, rather than average students. Ginsburg et al. (2010) has asked for descriptions of average students, but found that these did not elicit additional themes. The term “marginal” was chosen instead because the correct categorisation of marginal students is an area of difficulty and it was thought this may elicit valuable descriptions which would help clarify the difference between good students and weak ones. During the interview, participants were asked to describe a particular student they remembered, not an amalgamation of various students. This was to ensure the description was not an idealised—and potentially unrealistic—one, and to capture examples of how supervisors weighted or disregarded aspects of the student performance in forming an opinion of them. After supervisors had given their initial description of what made the student excellent/weak/marginal, I prompted them to clarify aspects of what they had said and to elaborate with examples of what they had observed. I then prompted them to cover any aspects of performance they had not already talked about such as the student’s knowledge, application of knowledge, technical and animal handling skills, communication, and work with others, wherever these were relevant to the placement. They were also prompted to describe

strengths or weaknesses if they had not already described any. A pilot of the adapted interview protocol with two staff members demonstrated that the questions functioned well to elicit descriptions of students in all three categories. The pilot interviews were not included in the analysis.

I conducted the interviews with each participant in person in their own office or workplace and audio-recorded them, with their permission. The interviews were then transcribed verbatim by a research assistant who had signed a confidentiality agreement. Pseudonyms were used to protect the identity of participant supervisors. I verified and corrected the transcripts by listening to the recordings. Each transcript was also verified by the participant who had the opportunity to clarify or correct any part of the transcript. Only a few participants had changes and these were minor, such as typographical errors. Both the transcripts and the recordings were uploaded into NVivo (version 10, QSR International) for analysis. This allowed concurrent analysis of audio and written versions to aid in the capturing of the meaning intended by participants as conveyed by tone of voice, pauses, and non-verbal sounds.

Participants

Interview participants were purposively selected from those supervisors who had been performing in-training evaluations of final year students from Massey University at the time of the study. Only those supervisors from the core (compulsory) placements were included for two reasons. Firstly, supervisors on core placements had supervised large numbers of students over the year, and so had a great deal of experience from which to form their views and draw examples. Secondly, since all students had to participate in the core placements, each supervisor was likely to see a spectrum of students with a range of interests, behaviours, and personalities.

Of the 54 supervisors who supervised core placements, I planned to interview 15 participants because this was a manageable number that had been sufficient to reach saturation of themes in a previous study of a similar design (Ginsburg et al., 2010). In order to capture a range of views, supervisors from across the range of subdisciplines and sites (on-campus or off-campus) were selected. The number selected from each subdiscipline and site was proportional to the time spent in each subdiscipline and site by students on the core placements. Thus, a greater

number of supervisors were selected from subdisciplines and sites in which students spent more time during their core placements. Only supervisors who were veterinarians were included because others would not have the same educational and professional experience, and thus would have had different influences on their views. Only supervisors with at least two years of experience supervising students on placement were included, to ensure they had adequate experience on which to draw for the interviews. Lastly, participants were also selected to provide representation from both genders and a range of years of experience.

Of the 15 supervisors invited to participate based on these criteria, all agreed to participate. There were four female and 11 male participants. Three were non-academic staff in private practices and 12 were academic staff working on campus. There were four supervisors from the production (farm) animal subdiscipline, two from equine, five from small animal (dog and cat), two from pathology, one from diagnostic imaging and one from anaesthesia.

Thematic analysis procedures

Constant comparison analysis (Leech & Onwuegbuzie, 2007) was used to investigate the interview data. This is the method of choice when research questions are general and overarching (Leech & Onwuegbuzie, 2007). In constant comparison analysis aspects of responses are categorised (coded) and codes are then grouped into themes (Leech & Onwuegbuzie, 2007). The codes used were generated from investigation of the transcripts and recordings (i.e. emergent) in an iterative process. Dual coding was permitted.

Within NVivo, I read and initially coded the transcripts to identify key ideas. These were segments of descriptions that seemed significant either because they were recurrent, were emphasised by a participant, linked with learning outcomes and competency frameworks, or had been mentioned by other authors previously in veterinary science or other disciplines. The linked recordings were used to clarify aspects of the transcripts as necessary to understand the supervisor's meaning. This initial coding generated 131 key ideas, some of which were overlapping or related. I then sorted the key ideas into groups according to the similarities in the ideas. This resulted in 12 themes which summarised the key ideas. Three framework matrices were then created according to the 12 themes, one for each type of student (excellent, weak, marginal). The framework matrix was a table with the 12 themes as rows and

the participants as columns. In each cell of the matrix, I summarised the theme from the original transcripts. This process allowed me to check that the themes were an appropriate grouping for the key ideas. The matrix framework summaries for each participant were then themselves descriptively summarised to form the descriptions and definitions of the themes. Descriptive information about the way each theme was used, illustrated with excerpts, was also assembled.

Descriptive information about how often a theme was used by supervisors for each type of student they described was also assembled. Together with the emphasis given by supervisors, this was used to give an indication of the relative importance of themes to supervisors. Because some themes were prompted for if not discussed spontaneously by supervisors, the frequency of theme use in the initial spontaneous descriptions was assembled separately from the overall frequency. The use of each theme by supervisors was also categorised as being positive (indicating a strength) or negative (indicating a weakness) or mixed (indicating the theme had both positive and negative aspects) and these were quantified for descriptive analysis. Themes that did not concern the student's current performance, for example themes that concerned the supervisor, were not included in the analysis of positive and negative comments.

Findings

Themes described by supervisors

Many comments by supervisors were descriptions of the student's abilities, behaviours, attitudes, and personal characteristics. Nine distinct but interrelated themes could be identified in these comments. Supervisors also went beyond descriptions of the students themselves to comment on the student's impact, prospects, and the difficulty in judging their competency, resulting in an additional three themes. These are defined in Table 4.1.

Each theme was present in the initial spontaneous descriptions offered by at least three supervisors (Figure 4.1). Most supervisors used each theme at some point in their descriptions of at least one of the three students they described (Table 4.2). All themes were mentioned

both by academic and non-academic supervisors. This demonstrates that the thematic framework developed was useful for capturing the thoughts of supervisors.

Table 4.1: Themes arising from supervisors' descriptions of excellent, weak, and marginal veterinary students, and their definitions.

Thematic area	Theme	Definition
Engagement		Enthusiasm for and participation in the activities of the placement.
Trustworthiness		Honesty, reliability, taking responsibility, and ability to discern own limits and act within them, including perceptions about own abilities
Discipline-specific knowledge and skills	Knowledge	Discipline-specific knowledge
	Applying knowledge	Problem solving, planning, decision making, report writing, giving presentations
	Technical skills and animal handling	Discipline-specific technical and animal handling skills including safe handling of animals
Relating to others	Communication	Character and effectiveness of communication and willingness to communicate, listening, body language
	Social interactions	Social awareness, relationships with others, ability to work with others
Personal functioning		Management of emotions and personal issues, managing fears, shyness, confidence
Caring for animals		Showing concern for animals under their care, paying attention to the animals, concern for animal welfare
Other aspects	Impact	Effects of the student actions and attitudes on the supervisor or other staff
	Prospects	Predictions about the future career or employment of the student
	Difficulty in judging	Aspects of student performance and the assessment process that made it difficult to assess students

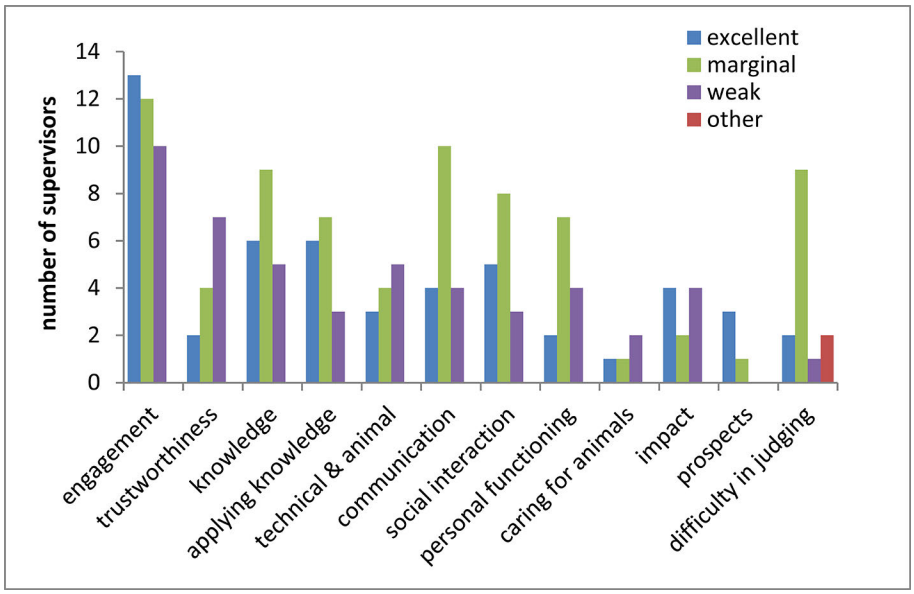


Figure 4.1: Number of supervisors using each theme in their initial spontaneous descriptions of excellent, marginal, and weak veterinary students.

For example, 13 of 15 supervisors described aspects of engagement in their spontaneous descriptions of excellent students; 12 of 15 described it in marginal students; and 10 of 15 described it in weak students. The category of "other" refers to comments made after description of excellent, weak, and marginal students was complete, when supervisors were asked if there was anything else they would like to add.

Table 4.2: Number of supervisors using each theme in their initial spontaneous descriptions or within all (both spontaneous and elaborated) descriptions of any student.

Thematic area	**Theme**	Spontaneous	All (spontaneous and elaborated)
Engagement		15	15
Trustworthiness		8	13
Discipline-specific knowledge and skills	Knowledge	11	15*
	Applying knowledge	10	15*
	Technical & animal	7	14*
Relating to others	Communication	12	15*
	Social interaction	9	15*
Personal functioning		10	13
Caring for animals		4	9
Other aspects	Impact	7	10
	Prospects	3	8
	Difficulty in judging	10	14

Note. Asterisks indicate themes that were prompted for if not spontaneously discussed by supervisors and which therefore could be expected to be present in the elaborated descriptions of all supervisors if the theme was relevant to the placement.

For example, 10 of 15 supervisors spontaneously used the theme of personal functioning in their descriptions of at least one of the three students they described, and a further 3 used it during clarification or elaboration of their descriptions, but this theme was not specifically prompted for.

Frequency of student strengths and weaknesses

The initial spontaneous descriptions of excellent students were almost all positive in nature, and of weak students were almost all negative (Table 4.3). Descriptions of marginal students tended also to be negative, but there were a greater proportion of positive and mixed aspects than in weak students. After further elaboration, the balance of positive and negative aspects altered little for excellent and weak students, but the proportion of positive comments about marginal students increased (Table 4.4).

If negative comments were made about an excellent student, supervisors usually balanced them with positive aspects of the same theme. Also, generally, the weaknesses of excellent students tended not to be very negative. For example, Adam mentioned a “slight lack of self-confidence”, a theme that was also picked up on by Gary, and Jean spoke of an excellent student as being “quite a serious, studious looking person”. Ed described an excellent student who lacked “the practical experience of actually doing”. Knowledge that “wasn’t quite up to scratch” was mentioned by Walter and echoed by Ben, Harry, and Todd. There was often a sense that these were only relative weaknesses, for example, Ben explained that “no one knows everything” and didn’t see that “necessarily as a weakness”. Martin and Simon could not identify weaknesses in the excellent students they described.

In weak students, the strengths mentioned were sometimes significant strengths, for example “base knowledge was great”, “nice and friendly”, and “diligent”, but more often were not very strongly positive, for example “affable”, “fine with clients”, and “she was okay”. The strength of two weak students related to improvements made after they failed the placement. Julia and Oscar could not identify strengths in the weak students they described.

Table 4.3: Presence of positive, negative, and mixed themes in descriptions of excellent, weak, and marginal students: spontaneous descriptions.

	excellent	weak	marginal	total
positive	44 (85%)	5 (9%)	12 (18%)	61
negative	2 (4%)	48 (84%)	46 (68%)	96
mixed positive and negative	6 (12%)	4 (7%)	10 (15%)	20
total	**52 (100%)**	**57 (100%)**	**68 (100%)**	**177**

Note. The themes of impact on the supervisor, prospects of the student and difficulty in judging competency are not included in these totals. Percentages are rounded to the nearest integer.

For example, 85% of themes used in descriptions of excellent students were positive, 4% were negative and 12% were a mixture of positive and negative.

Table 4.4: Presence of positive, negative, and mixed themes in descriptions of excellent, weak, and marginal students: spontaneous and elaborated descriptions.

	excellent	weak	marginal	total
positive	105 (85%)	15 (13%)	29 (28%)	149
negative	4 (3%)	95 (80%)	60 (59%)	159
mixed positive and negative	15 (12%)	9 (8%)	13 (13%)	37
total	**124 (100%)**	**119 (100%)**	**102 (100%)**	**345**

Note. The themes of impact on the supervisor, prospects of the student and difficulty in judging competency are not included in these totals. Percentages are rounded to the nearest integer.

Use of themes by supervisors

Engagement

The theme of engagement was used to refer to a student's enthusiasm for and participation in the activities of the placement which could include specific learning activities such as discussions and presentations, a range of professional activities, or generic work such as fetching equipment or cleaning up. Supervisors sometimes referred to specific tasks but they often spoke more generally. Frequently they described attitudes or dispositions (such as motivation, commitment, laziness) or unobserved behaviours (such as thinking and reflection), complemented by descriptions of observed behaviours (such as initiative, thoroughness, punctuality, doing extra reading, asking relevant questions), or standards of performance (such as work quality).

Engagement was the theme mentioned by more supervisors than any other in their spontaneous descriptions of all types of students (Figure 4.1), and all supervisors used this theme in describing at least one student in their spontaneous descriptions (Table 4.2). This suggested it was very important to all supervisors and a memorable aspect of student performance. For Walter it was the "first thing that made them really excellent". Adam, Ben, Julia, Harry and Gary all directly stated its importance to them over all other factors. Gary explained that "when it comes down to it, that is all I really care about. I want them to be interested, I want them to be engaged, I want them to be motivated to learn and I want them to enjoy what they're learning in my rotation."

Engaged students were ready and eager; they volunteered to do things, asked questions, and seemed to be enjoying themselves. As Adam described, "they're obviously happy to be there". Walter, Julia, and Simon also emphasised the eagerness of the engaged students they described, which included being eager to do things that were boring or that other students did not want to do, even at night. For example, Walter described "how unbelievably enthusiastic they were, to do something, it didn't matter what time, what time of day or place even, they'd be keen to, to go out and do things". Todd and Harry also described engaged students who looked for things that needed to be done and tried to be helpful.

Even when not actively doing things, students demonstrated their engagement through their body language, making eye contact, and paying attention. They also showed they were actively thinking and reflecting on what was going on by "asking questions that let you know that they were synthesising the information and taking it a step further" (Ben), and as Martin and Walter also described. Engaged students also sought further information from reading, as Ruth and Jean mentioned.

Some supervisors also mentioned evidence of feedback-seeking, self-assessment, and wanting to improve by very engaged students. Nancy described a student "actively seeking feedback on how he was doing and what he could improve". Oscar described a student who asked for "tips" about the areas she felt she was having trouble with, and then deliberately practiced and improved on those areas after feedback. Walter described a student who self-identified areas for improvement and deliberately challenged herself in those areas by her choice of topic for a presentation.

Amongst students who were not engaged, there were different ways of not engaging. Some did not participate in the placement activities and others participated but nevertheless did not seem to be learning. Non-participation ranged over the activities of the placement as well as learning activities, and often involved both. The unengaged students described by Harry, Gary, Julia, Jean, and Oscar did not volunteer to take on work. As Julia explained, she "had to assign her cases most of the time". Similarly, Harry said that his student "wouldn't actually do a lot of those things unless she was asked to" and Gary said that "they just stand back and let other people do it". Jean described a student who would "do a minimum amount I suppose, leaving the majority of the work with, or the dirty work if you like, to the other students". When directed to participate some students "had to be prompted several times and then it would still not happen" as Ruth explained, and Martin and Julia described similar scenarios. Gary found the requirement to repeatedly prompt a student to do things very frustrating, explaining "It's very frustrating as a, as a teacher to have to feel like the only way they're going to do something is if you're standing there."

Students would also sometimes arrive late or not at all as Jean, Wayne, and Ruth referred to. Unengaged students also contrasted with engaged students in not asking questions and not voluntarily participating in discussions, as Jean, Ruth, Ed and Walter described. Wayne described having to "drag out" a discussion with such students.

Three different types of attitude and behaviour were described by supervisors in relation to students who did not participate. For some, supervisors inferred an attitude of not caring and not wanting to learn. Such students appeared not to be making an effort or trying hard and some seemed to act as though they thought the experience was a waste of their time. They would do the bare minimum required to pass the placement. To Ruth it seemed they were "lazy". Julia, Gary, Jean, and Ruth also spoke along the lines summed up by Walter in these comments:

> Straight up uninterested and, and everything, like, the whole, yeah, their body language, you know, even to the level of the way they dressed, you know showed like, I just don't care kind of attitude. Finding any way to be a little bit, to do the least possible but get a pass mark. Finding, you know, like an easy way around doing something that's been asked of them to do. Just to get the box ticked essentially. (Walter)

An attitude of not caring was also implied by students appearing not to do any further reading, or giving any deeper thought to the meaning of information being discussed. This could be in relation to presentations or reports, as Julia and Gary described, or to the significance of clinical findings as Jean described. Some supervisors, such as Gary, commented on a lack of effort shown, others that work was not done well (Wayne) or that it was superficial, not thorough, and things would be missed (Jean). Gary did not think his student even cared that his performance was poor, based on their reaction to being corrected about a lack of effort, and Ruth described a student not improving his work when given the opportunity.

Others who were not participating appeared to supervisors to think they knew it all already and did not need to practice. Martin and Wayne both mentioned this attitude in the students they described, and Martin's student even voiced this explicitly when confronted about their lack of participation, saying "yes I know it all, I just didn't have to be there saying it" which Martin thought very arrogant. Simon and Wayne described students who were dismissive of feedback and would make "light of [their] failings" as if anything they did not already know or could not do was unimportant. Ed described a student who would argue assertively and disrespectfully with the opinion of the supervisor about the best way to do things and was reluctant to follow suggestions.

Still others who were not participating were extremely quiet and reserved, or would just stand and not do anything. They would hang back or hide behind others and would look down or avoid eye contact to avoid being involved. Supervisors inferred various reasons for these behaviours. Some students seemed very shy. Gary described such a student at length, who was "incredibly shy …and you have to force them to talk" but similar students were also described by Harry, Oscar, Walter, and Martin. Some students seemed to lack the confidence to participate or to keep trying when something was difficult, as Harry and Oscar found. Julia and Gary both talked of students who needed explicit and detailed instructions in order to perform tasks. In his student's case, Gary thought this reflected a lack of motivation and initiative. Julia thought it reflected a lack of confidence. With other students, supervisors wondered if they lacked fundamental skills that would enable them to understand what to do or to think and make decisions. This included lacking sufficient English comprehension. Nancy described a student who seemed to be unable to act independently to make decisions and had a tendency to "dither". Ed assumed at the time that a student "wasn't doing anything"

because she did not know what to do, he also wondered if it could have been that she "was so lost in the whole procedure and stressed by it, she wasn't thinking any more".

Another group of students did participate in activities but nevertheless did not seem to be learning. Todd and Simon both described students with misdirected interests who appeared to be highly engaged but in aspects that were not directly relevant to the applied situations they were working in. They would "go off completely on a tangent" (Simon) and be overly focussed on some aspects to the exclusion of what was important. Other students were participating but did not appear to be taking in feedback. Nancy described a student who was constantly seeking (and receiving) feedback but whose performance remained poor. Julia described a student who "would go away and look things up" but Julia was "never convinced that she was able to work out the significance of the things that she was learning". Harry described a student who "did seem to be trying but it just didn't seem to, didn't seem to sink in with him". Harry wondered if the student "had this sort of underlying belief in himself that he actually knew what he was doing when he didn't".

Engagement, and the various ways of not being engaged, seemed to be very important determiners of excellence and weakness, but the descriptions of three particular students illustrated how students did not have to like a subject in order to be highly engaged (Box 1). Although every student described by supervisors was different, these three students had something in common, as each of them wanted to be a large animal (farm) veterinarian and was focussed on that pathway, but was working on a small animal (dog and cat) placement. Each behaved very differently on the placement. Harry's student was excellent. He approached the small animal placement with enthusiasm. There is a striking difference between his attitude, and his engagement with learning and the work of the placement, and those of the weak students described by Wayne and Jean. It is not really surprising that poor engagement should have such a negative effect. By acting as though they don't care about something their supervisor thinks is important, students risk offending their supervisor, and negatively affecting the supervisor's view of them. Gary articulated this idea explicitly and suggested that students would be best advised to act interested even if they were not.

Box 1: Three vignettes of students who were primarily interested in large (farm) animal work, during their time on a small animal (dog and cat) placement. These vignettes are included to illustrate that engagement did not require liking for a subject area.

Harry, describing an excellent student.

The first thing Harry said about this student was that "he was very hard working, got stuck in and was very helpful and offered to assist with everything". This included routine tasks such as cage cleaning. He spoke of the student's capability, initiative, and conscientiousness. The student was fairly quiet, but still interactive as he "asked the right questions without asking too many questions". The student was polite with clients and shook hands when it was appropriate to do so. He listened and observed quietly and would "stand in the right places". Although there were gaps in his knowledge this student was able to think through problems and come up with 2-3 "good valid options" that were realistic in the context of the situation and sensitive to client needs. The student admitted when he didn't know much about an area, but would also try to think through the answers to things he did not know. The student had good technical and animal handling skills and was calm and confident with the animals. He would ask for help if he needed it, and, combined with his capability and conscientiousness, this meant he could be trusted with a fair degree of responsibility, which was of great help to the supervisor at a time when the clinic was short-staffed. "He could just sort of handle things."

Wayne, describing a weak student

The first thing Wayne said about this student was that he "basically made it seem like he knew everything… that he thought he already knew it". Despite this, the student was often wrong and had poor ability to apply knowledge. He was hard to teach because, when corrected, he would dismiss the feedback and make light of his failings, brushing them aside as if they were unimportant because he intended to work in a different discipline when he graduated. He would turn up late some days and other days would not come at all. He had to be asked more than once to do things and then did poor quality work. He was not liked by the other students in his group. Despite describing the student's technical skills as "fine" and that his communication with clients "wasn't too bad", when asked about whether this student had a strength Wayne replied "No. We didn't think he had a strength." In fact, Wayne thought the student was potentially dangerous, saying of him "he's not the sort of person that I would want to employ because he doesn't know his limitations and he's wrong in a lot of things. He, you know? He thought he knew, [but] he was actually wrong."

Jean, describing a weak student

The first thing Jean said about this student was that "they essentially had a bad attitude and a lack of enthusiasm and we see this sometimes in students who have got all the brains but

just choose not to use them". The student acted as though "they just wanted to get through the programme, so they would do the bare minimum". He was not thorough and would miss things, would show up late and leave work to others. "If an extra case came in they'd be the last person to volunteer for it." Although not neglectful or dangerous, the student did not seem to care about the animals, interacting with them very functionally. His knowledge levels seemed good, although he would not answer questions unless pushed. However he was not well able to apply his knowledge and Jean thought he was not "perceptive", and could not see the "bigger picture".

Trustworthiness

The theme of trustworthiness was used to refer to aspects of the student's performance that engendered trust or distrust in supervisors. It included honesty, reliability, taking responsibility, and the ability to discern their own limits and act within them. Lack of trustworthiness seemed to be an important theme in weak students. It was the theme referred to by more supervisors than any other except engagement in their initial spontaneous comments about weak students (Figure 4.1), and was discussed quite often in elaborated comments. It often appeared to cause real concern, with supervisors using the term dangerous in descriptions of four students. In contrast, trustworthiness was not mentioned spontaneously by many supervisors in regards to excellent students, but, although not specifically prompted for, was a feature of descriptions elicited on further elaboration of several excellent students. Importantly, where it was mentioned by supervisors, excellent students were always trustworthy, and weak and marginal students always were not, which may indicate that it is a strong indicator of the overall evaluation of a student in the minds of supervisors. Almost all supervisors referred to aspects of trustworthiness in one or other of the students they described, and just over half of them did so in their initial spontaneous comments (Table 4.2).

Students who were trustworthy seemed to know their own limits. They were able to act independently but did not overstep a boundary that the supervisor thought was appropriate, as Todd explained when describing an excellent student: "She was a confident but not overly confident student. She also had a sort of sense of not, when something would be overstepping like I shouldn't go there." Ed, Harry, Jean, and Oscar all described students who would readily admit when they did not know and would ask for help or check with supervisors when it was

needed. There was a sense of the amount of checking being well judged – neither too little, nor too much: “They bothered you when it was appropriate to. When I say ‘bother you’ I don’t mean it was a nuisance, but they came to you when it was appropriate” (Jean).

In addition, descriptions of students who were trustworthy often indicated they were very capable (Harry, Ruth, Simon, Todd) and tended to make the correct decisions, even for things they were checking with supervisors (Ed, Todd). Nancy and Jean described students who could be relied upon to check on patients without being asked. Jean, Nancy, and Todd spoke of students who could explain and convey information to clients without going too far or adding to client concerns.

Students who were not trustworthy displayed a spectrum of behaviours. One group described by Ed, Harry, Martin, Simon, and Wayne tended not to know their limitations, be overly sure of themselves, and even “cocky” (Wayne) or “arrogant” (Simon). Harry described a student who “tried to do everything by himself to start with” and was “probably a little on the reluctant side to ask for help”. Ed, Simon (for two students), and Wayne went as far as describing such students as dangerous because they would endanger the lives of their patients, themselves, or others: “They had a lot of experience but their experience led them to be over-confident in their ability to do the job and to the extent that they were dangerously so. They were too sure that their decisions were the right decisions.”

Other students were not trustworthy because supervisors could not depend on them to notice or seek help if something were to go wrong or if they did not understand what to do. The students described by Ben, Ed, and Jean were inattentive and not thinking about what was going on, or lacked sufficient knowledge and capability to identify when something might need to done or to know how to start. Ed described being “very uncomfortable about leaving them alone with a patient”.

Another group of students were not trustworthy because they could not be relied on to be thorough or perform tasks when asked. Martin, Julia (two students), and Ruth all described students who did not do things they were asked to do. One of the students Julia described was also not thorough and would miss things, something that Jean described in a student as well. Oscar described a student who he could not “depend on” because she seemed not to listen

and would get the wrong idea about what he wanted. He "didn't get a feeling that she would ask straight away if she was, if there was something unclear". This contrasted with the student Nancy described, who was constantly checking with her and not able to act independently, indicating that there could be either too much or too little checking.

Finally, some students were not trustworthy because they appeared to supervisors to be saying what the supervisor wanted to hear, as Julia and Oscar described, or they appeared to be misrepresenting situations as described by Ruth and Walter.

Discipline-specific knowledge and skills

The knowledge and skills that were particular to veterinary science made up another group of themes and are considered here together.

Knowledge

The theme of knowledge was used to refer to the breadth and depth of discipline-specific knowledge of veterinary science. Supervisors spoke of knowledge and/or understanding and frequently qualified these to distinguish them from more applied and practical knowledge, using terms such as academic, core, basic, background, factual, theoretical, or knowledge of principles. This type of knowledge was sometimes also distinguished by reference to the didactic component or preclinical coursework or to specific preclinical subjects such as pathophysiology. Sometimes supervisors referred to student's knowledge in terms of intelligence, smartness, cleverness, giftedness, or "brains". Discipline-specific knowledge was mentioned by many supervisors spontaneously when describing students (Table 4.2). However it was not the most important contributor to excellence, as suggested by statements made by Nancy and Harry and this from Gary, which indicated that engagement was far more important than being correct.

> They asked good questions, they answer my questions, even if they don't have the right answer, they still had a good crack, and even if they got the wrong answer they were able to demonstrate that they had made an effort to think about it. (Gary)

The finding that knowledge was the most frequent weakness of excellent students and the most frequent strength of weak students (Figure 4.2) also suggested it was not the most important contributor to perceptions of overall excellence and weakness. Although some

excellent students demonstrated a very high level of knowledge and understanding, students could still be excellent with gaps in their knowledge and some excellent students were described with only average to good knowledge, and not necessarily the best grades. In a similar way, weak students did not always have poor discipline knowledge. Walter described one weak student's "phenomenal" knowledge base as "his massive saving grace". Marginal students frequently had average or good levels of knowledge but some were described with poor knowledge.

Usually students demonstrated their knowledge or lack of it by the correctness of their answers to the supervisor's questions. For example, Oscar related an example of a student with gaps in her knowledge, who "didn't know what it [a particular disease] was or the cause, what you could do about it, what does it mean for the animal". Wayne described a student who other students saw as an asset and would "turn to him if they got stuck on something" because he always knew the answer. Adam explained that determining a student's level of knowledge was also about the type of questions they would ask, that "show good understanding for the level they're at". This was echoed by Gary when he said that with "a question you can demonstrate a lot more intelligence sometimes than with an answer you know", and Harry and Ed also spoke of the appropriateness of questions students would ask. Nancy spoke of frequent silences when describing how she could see that knowledge was deficient in one student. Sometimes supervisors suggested it was difficult for them to determine the knowledge level of students. In some cases, such as a student Gary was describing, this was because the student was not participating. As Nancy put it, "it's hard to assess knowledge if they're not confident enough to express an opinion". Oscar also mentioned that "it's very hard for us to assess the level of knowledge" which seemed to be because the clinical setting was more about applied knowledge. Frequently supervisors referred to their awareness of a student's performance in other assessments or rank in the class in determining their level of knowledge.

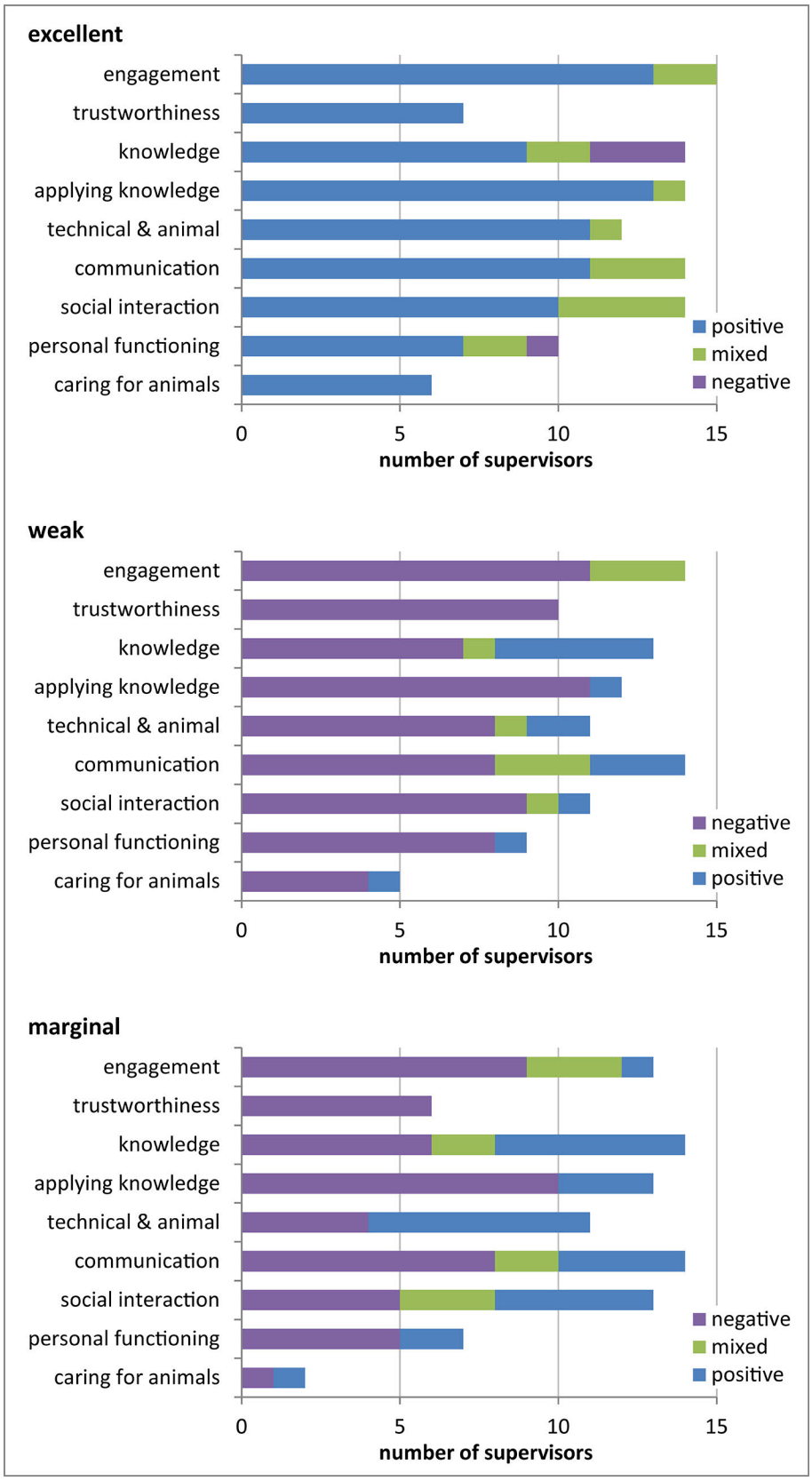


Figure 4.2: Frequency of positive, negative, and mixed depictions of themes occurring in all descriptions (both spontaneous and elaborated) of excellent, marginal, and weak students. Mixed depictions contained both positive and negative aspects.

Applying knowledge

The theme of applying knowledge was used to refer to cognitive skills such as problem-solving, planning, decision-making, and determining the significance of findings or options. It also included skills such as report writing and giving presentations. As a theme, the ability to apply knowledge was discussed spontaneously by many supervisors, especially in relation to excellent and marginal students (Figure 4.1). When mentioned, most excellent students described had reasonable or good abilities in applying knowledge, and most weak and marginal students had poor ability to apply knowledge or gaps in their ability to apply knowledge (Figure 4.2).

Supervisors spoke sometimes in terms of the qualities of thinking. For example, Ben described how a student would organise concepts "into a more complex thought and question and maybe something that I had or hadn't even thought of before". Todd spoke of a student saying the wrong thing in answer to a question "because his mind was very jumbled about". This theme of orderly or disordered thinking was also spoken of by Gary, Jean, and Nancy when describing student's abilities to give focused, logical, and organised explanations in presentations, clinical notes, and explanations to clients.

Another aspect of applying knowledge related to analysis and synthesis or integration of knowledge, putting things together, making connections, relating information, suggesting options, and solving problems. Ben spoke of the ability to "attack a problem". Some supervisors spoke of higher-order thinking (Julia), co-ordination of knowledge (Adam), and deductive reasoning (Martin).

Supervisors also described the ability to understand the meaning of findings, consider the whole picture, and distinguish the relative importance of information so that they could make a plan that was "realistic" (Harry) or "feasible" (Walter). Harry referred to the importance of "picking the better options" and this was also discussed by Jean in terms of students being able to focus on more likely than less likely possibilities. She contrasted a student who was good at this with one that had caused clients a lot of anxiety by over-emphasising remote, but more serious possibilities. Like Jean and Walter, Ed also emphasised the importance of students being able to "integrate their physiology and work it out to make a decision based on the whole picture of the patient". The sense of "putting an amalgam of ideas together and coming

up with a plan or recommendations" (Martin) was conveyed by many supervisors. The step of being able to apply information to a clinical situation was emphasised by Todd: "He was a gatherer, but not, really not able to then translate that in any kind of useful action for the patient or the client."

An additional thread concerned being able to predict what was going to happen and re-plan as things went along. Jean spoke of the ability to "deviate from the plan", based on integrating and interpreting the significance of "what they're seeing". Todd also spoke of the "ability to anticipate what comes next" and Ed referred to "thinking about what's going on in this particular patient" and "what does this individual patient need?"

Technical skills and animal handling

These two themes were used to refer to the discipline-specific technical and animal-handling skills, including the safe handling of animals. Supervisors commonly spoke generally of technical skills, manual, or practical skills, but sometimes referred to specifics such as physical examination, venepuncture, catheter placement, surgical skills, necropsy skills, or pregnancy testing (by manual palpation). Technical or animal handling skills were mentioned by about half of the supervisors in their spontaneous descriptions (Table 4.2).

Technical skills were not relevant to every placement as some did not involve technical work, and even for some of those that did, there was a sense that a high level of technical competency was not expected or important in that discipline or at the student's stage of training. Gary thought that "the technical aspects of what I do are not important… I can imagine the technical skills would be a lot more important for, you know, other clinical rosters." and Wayne said that "we have low standards…. we're not expecting them to be great surgeons". Nancy explained that a student's technical skills were "about where we would expect the students, you know, still learning". A few other supervisors made it clear that rather than the absolute standard, it was a trajectory of improvement that was most important to them. Oscar was particularly impressed with a student who worked on her technical skills and improved, saying "you can actually see improvement which is brilliant I think". Nancy and Simon both described the opposite situation in two weak students who were not making any progress.

Animal handling skills were also not relevant to every placement and were infrequently mentioned in spontaneous descriptions of students. However, they seemed important to those who did discuss them and especially so when they impacted the supervisor, affected the trustworthiness of the student, or when caring for animals was involved. Animal-handling skills spoken of included understanding and interpreting an animal's behaviour; being quietly confident and able to move the right way and make the right noises; carrying, handling or restraining in ways that reduces stress on the animal; applying common sense around animals; and paying attention to the animal. For example, Oscar described a student who "read the animal's behaviour well enough to be able to handle it competently as the least stressful manner". Adam described a student as "quietly confident and [had] the ability to understand the animal's behaviour so they stand in the right place, [and] make the right noises at the right time". Having a calm and quiet approach was also commented on by Harry.

Poor animal handling skills were sometimes associated with apparent fear or lack of confidence. As Adam describes, students sometimes seemed to be "fearful of large animals". As a result the students may reluctant to get involved or become "quite panicky so they tend to make sudden movements, to be loud, to make noises at the wrong time". Sometimes, however, poor animal handling skills were related to overconfidence and being too abrupt and forceful with animals, such as with a student Harry spoke of. Simon also spoke of overconfidence and lack of common sense leading to students, "not taking the necessary precautions", "just being a little bit too casual" around the animal and putting themselves "in dangerous positions".

Most excellent students for whom technical skills were mentioned had good technical skills, and most weak students for whom they were mentioned had poor technical skills. Technical skills tended to be good in marginal students when they were mentioned, but were sometimes poor. In a similar pattern to the technical skills, animal handling skills, when mentioned, were always good in excellent students, tended to be poor in weak students and tended to be good in marginal students.

Relating to others

Two interrelated themes of communication and social interaction concerned the student's interactions with and ability to relate to other people, and will be discussed here together.

Communication

This theme was used to refer to the character and effectiveness of communication by the student, their willingness to communicate, listening behaviour, and body language. Supervisors spoke of students communicating and asking questions, having discussions, and conversations. They discussed communication with staff, clients, and other students. In marginal students, communication was the theme referred to by more supervisors than any other except engagement in their initial spontaneous comments, but few supervisors mentioned it in regards to excellent or weak students (Figure 4.1). Most excellent students communicated well and any negative aspects were balanced by positive comments (Figure 4.2). Most weak and marginal students had elements of poor communication.

Students communicating well stepped forward and introduced themselves, as Oscar and Nancy explained. They shook hands "when the people wanted to shake hands" (Harry). A number of supervisors described them as friendly. They were talkative and interactive but not too much so, as Walter described, and could communicate well at different levels with all sorts of people. Nancy spoke of a student communicating clearly, "in a manner that the client will believe what they're saying" and of giving clients "directions and structure for the consultation". Jean described a student who was "good at explaining things in a logical way, using everyday language, not dumbing it down, but in a way that the client could understand, and he was good at building agreement as to a plan and a way forward".

As well as contributing to discussions, good communicators were quiet when it was appropriate, "not sort of shy quiet, just sort of respectful quiet if that makes sense" (Harry) and "patient enough to listen to what was said by someone else" (Oscar). Walter and Oscar both described situations where students had been very tactful and diplomatic in choosing the appropriate moment to discuss difficult issues and had been sensitive in their discussion of them.

Some students not communicating well were inappropriately "quiet and hesitant to contribute" (Martin) and "wouldn't really join in conversations or discussions, and would seem to be quite withdrawn" (Adam). Conversations or discussion would have to be "dragged out" (Adam, Wayne) or students would only speak if asked a question (Walter). Sometimes supervisors surmised that students who weren't responding to what was being said were not listening (Oscar, Simon), but another reason inferred by Jean was poor English comprehension because English was a second language. However, sometimes a student's body language implied they were not interested in communicating, as described by Walter and Gary.

Supervisors sometimes described students with poor communication skills. Wayne referred to students who "struggle to just start a conversation" because "they just haven't learned conversational skills". Todd and Nancy both described students who could hold a conversation well enough but were not convincing or authoritative when speaking in a professional context. Poor skills sometimes involved reading people's body language or between the lines of what they were saying. Julia described a student who "couldn't read people very well and she seemed to just annoy them quite easily".

Sometimes students were quite talkative. Harry described one whose chatty conversation and discussion on a wide range of subjects was generally stimulating but "occasionally it was over the top". Julia described a student who would frequently "interrupt [discussions] to tell me, everybody, things that she had found out". Two students argued with supervisors. One, described by Ruth did so reasonably and professionally but another, described by Ed, was very assertive and bordering on being impolite with staff, "questioning their judgements, second-guessing, inappropriate replies to questions". One student described by Oscar seemed quite judgemental although polite. Oscar found him difficult to communicate with because he seemed to constantly be saying what he thought Oscar wanted to hear rather than what he really thought.

Social interactions

This theme was used to refer to the student's social awareness, relationships with others, and ability to work with others. Social interactions were mentioned by about half the supervisors in their spontaneous descriptions of marginal students, but by fewer supervisors in relation to excellent or weak students (Figure 4.1). Most excellent students had good social interaction

skills and any negative aspects were balanced by positive comments (Figure 4.2). Most weak students had elements of poor social interaction and marginal students were very mixed.

Supervisors described students who excelled at social interaction as working well with others, including their peers, veterinarians, and technical and lay staff. Oscar described them as interested and open-minded towards clients. They understood the social and business situation of clients and farmers, and the role of the animal in different contexts, and were able to use that to behave sensitively and make realistic suggestions and plans. For example, Adam described how students "understand why you're there, what role you play in the system I suppose rather than just, you're there to fix a sick cow, or to get an answer. They understand how the procedure fits in with the farming enterprise and what the farmer's trying to achieve." Harry similarly spoke of a student with excellent awareness of the social context of clients.

Another aspect of excellent social interactions described by Oscar was that of being respectful of others and not judgemental, in relation to a situation where the client's actions had not been the most appropriate. Several supervisors described situations where students had shown a good sense of the social situation and when it was appropriate to talk and when to hold back. This included knowing when it might not be the best time to ask questions (Ben and Walter), when an issue was sensitive and they should check with the supervisor before saying something to a client (Jean and Todd).

Students with excellent social interactions treated others well, without seeming superior. As Simon described, they "didn't think that they were better" than other students or technical staff, which Adam also alluded to. They also did not try to dominate discussions, or take all the best work as described by Jean, Ben, Gary, and Wayne. Here Jean explains: "But the thing I liked about this student is they knew a lot but they would just offer up an appropriate amount of stuff. They didn't dominate things, they didn't try and poach the better surgeries."

Ben, Harry, Wayne, Gary, and Ed all talked of students who were friendly and likeable and as Ed said "I got the feeling she was well-regarded by her classmates." In addition, Jean, Ruth, Simon, and Wayne described students who worked well in a group and were valued team members. Todd described a student who would organise work that involved several other clinical staff members "without ever offending anyone". Simon described a student who was a

supportive friend to another student in the group who was having personal difficulties, uncomplainingly doing an after-hours shift for them.

In contrast, descriptions of poor social interactions included students who did not work well with others. They did not pull their weight and left work to others, as Jean, Wayne, and Nancy described. In her case, Nancy attributed this to the student being hesitant and lacking self-confidence rather than unwillingness. Julia, Nancy, Simon, Todd, Wayne, and Ed described students who tended not to be well-liked and did not fit in, and whom others found annoying, frustrating, or unpleasant to be with. Such students were often isolated as Julia described, and "quite quickly they left her out". Two students also irritated their supervisors, Oscar and Walter, who felt they were arrogant.

Some students did not respect the opinions of others. They would belittle or challenge other opinions and be dismissive or judgemental. Julia's student did not know "when to rein back her strong opinions". She would challenge others and was a "bit scathing of their opinion". Simon described a student who "dominated in many sort of group scenarios with often wrong information [and] didn't allow other people to express their opinions".

Some were very self-absorbed and not thinking about how they could help others. Todd described a student as not understanding the role of a clinician to help people and their animals and having "no recognition that there was a responsibility beyond to himself". One of the students Walter described appeared to give no thought to how he could be helpful.

Personal functioning

This theme was used to refer to emotional and personal issues that might affect a student's performance such as fears, shyness, confidence, and other emotional issues, such as relationship issues, or illness. Personal functioning was mentioned by about half the supervisors in their spontaneous descriptions of marginal students, but by fewer supervisors in relation to excellent or weak students (Figure 4.1). For most of the excellent students about which it was discussed, personal functioning was good, and for most of the weak and marginal students about which it was discussed, personal functioning was poor (Figure 4.2).

Supervisors frequently spoke about the confidence of students they were describing. Sometimes supervisors spoke in general terms about confidence (Wayne, Nancy, Simon) or lack of confidence (Adam, Oscar), but more often they would speak about confidence in specific situations. For example, Adam, Simon, and Harry all referred to students being calm or confident around animals, which contrasted with students described by Adam, Todd, Nancy, and Harry who were fearful or lacked confidence around animals. Similar contrasting examples of student's confidence and willingness to do things and lack of confidence, hesitancy or tentativeness in doing procedures, talking to clients, or discussing things in groups, could be found in the descriptions made by Ed, Gary, Harry, Jean, Julia, Martin, Nancy, Oscar, Ruth, Todd, and Wayne. A characteristic of some excellent students was their confidence in identifying problems and making decisions as Jean, Todd, and Wayne described. Other excellent students were slightly less confident (Ed, Ruth) but Nancy described a weak student with a more significant lack of confidence to make decisions. This student would constantly seek feedback and reassurance about what they were doing. Some supervisors specifically mentioned not being overconfident as a strength of some students (Ed, Todd, Walter) and students who were overconfident were criticised by Ed, Gary, Harry, and Oscar.

Personal functioning also related to the way students managed personal issues so as to maintain their performance. Harry and Ruth each described a student who was ill and Simon described two students with personal difficulties. Supervisors found it hard to know whether to make allowances for performance in these cases and this is discussed further later.

Caring for animals

This theme was used to refer to concern for the animals under one's care and included paying attention to the animal and being empathetic towards it. It encompasses a sense of caring for the animal, wanting to help it and actually liking animals. The specific mention of aspects of caring or consideration differentiated it from descriptions of good animal handling skills. As a theme, it was only mentioned by a minority of supervisors (Figure 4.1), however, when it was discussed, especially when it was lacking, it was with a sense of importance and was amongst the main reasons for being weak in three weak students and one marginal student. When mentioned, caring for animals was never weak in excellent students, however strength in caring for animals was not able to compensate for other weaknesses in one weak student.

Harry, Nancy, and Wayne each used the term empathy in relation to the way students cared for, or did not care for animals. Supervisors described students reassuring, patting, or cuddling animals. Jean seemed to expect this as normal behaviour for veterinary students, as she described here when speaking of a student who didn't do this: "They treated the animal as a physical being with a bit of possible pathology.… Whereas most people would take the cat out and they would give the cat a cuddle." Wayne graphically contrasted students who carried animals with distaste "like a 3-week-old dead sardine" while explaining how an excellent student cared about animals.

Paying attention to how an animal was doing was another aspect referred to by some supervisors in different ways. Nancy described a student who kept checking on their patients in the hospital. Ben described how a student who cared for animals would look for ways to make them comfortable: "I would say conscientious towards the patients. ….. I guess thinking about how we can comfortably move the patient onto the table without them getting stressed and struggling." Harry spoke of a lack of attention to the animal and adjustment of behaviour to suit its needs, for example by backing off or slowing down: "He'd keep trying to do things with them and they were obviously not coping with it."

As Jean mentioned above, seeing the animal as more than just a disease or an interesting problem to solve was also important. Todd also spoke of a student who treated patients as if their only purpose in being there was to help him learn: "It seemed like he was never able to understand that 'I'm, my job here is not just for me to learn, but for me, while learning, to pass on information of value to clients, and taking care of my patient'."

Other aspects

Three additional themes were present in descriptions that were not observations of the students themselves and are discussed here together.

Impact on the supervisor

Impact was used to refer to the effects of the student's actions and attitudes on the supervisor or other staff. Several supervisors referred to a student's impact in their spontaneous comments (Figure 4.1) and a few more during elaboration of other aspects of performance (Table 4.2). Some students had a positive impact. They were enjoyable to have around on

placements because they were “easy to teach” (Todd), were interesting to talk to (Simon), or were helpful or useful (Harry, Todd, Walter), or decreased the supervisor’s workload (Simon). Students’ attitudes, social interactions, and personal functioning also seemed to play a role in a student’s impact, as Simon spoke of when he explained that “having to deal with kind of personality conflicts and attitude problems—that just makes it kind of a little bit arduous”.

Other students had a negative impact because they were more work to teach. Gary spoke of a student needing close supervision and direction in order for them to complete their work. Simon spoke of a student needing close supervision because they were being dangerous. Several supervisors mentioned the extra time needed to engage and enthuse a student (Ben), to push them into doing things (Harry), to answer their extra questions (Julia), or to work through a problem with a huge amount of step-by-step support (Jean). Todd described a student who was frustrating because “fairly routine elementary task[s]… wouldn’t get done in a reasonable period of time” and Oscar described a student who was “very hard to work with” because he couldn’t depend on her asking if something was unclear and she did not seem to be listening.

Sometimes supervisors spoke in terms of their own emotional response to the student’s behaviour. Students were frustrating when they didn’t seem to understand (Todd), or when they would not do things unless closely supervised (Gary), and irritating or annoying when they seemed to be lazy (Gary, Julia, Walter), or when they made assumptions about what the supervisor or the client wanted to do (Oscar). It was sometimes “painful” (Jean) or “overwhelming” (Nancy) to deal with the students when they required a lot of effort.

Some students caused dilemmas for supervisors. Todd worried that he might have intimidated a student who was not able to apply his knowledge to solve problems, although he recognised that other teachers also found the same issues. Gary wondered whether he should speak to students about needing to concentrate and pay attention or just ignore their poor behaviour. Gary also spoke of his struggles to teach very shy and quiet students, and how it was especially hard when there was more than one as they would band together.

Prospects of the student

Prospects referred to comments made by supervisors about the future potential and likely employment or further educational pathway of the student. Several supervisors mentioned the prospects of the student to enrich their descriptions (Table 4.2) but only a few did so in their spontaneous initial comments (Figure 4.1). Jean, Wayne, and Oscar each illustrated the excellence of a student's performance by the fact that they or their colleagues would have liked to employ the student after graduation. Similarly, Julia, Martin, and Wayne illustrated a student's weakness by the fact that they would not want to employ them after graduation or would have suggested others didn't if asked for a reference. Sometimes supervisors framed their comments about a student's prospects in terms of potential success in the workplace. Nancy spoke of her concern about a weak student, saying that "she'd need a very supportive environment and discussions around stress management and time management, and, yeah, personal skills". In contrast, Todd spoke of how his prediction that an excellent student would be highly successful had come true. Jean and Todd made comments relating to the branch of veterinary work they thought the student was suited to. In their view the balance of strengths and weaknesses of these students did not preclude them from veterinary work at all, but just from certain roles. Wayne similarly expressed the idea that students who perform poorly in one placement may do well in other veterinary roles.

Difficulty in judging competency

Although they were not directly asked about this in the interview, all except one supervisor made reference to the difficulty of assessing students while they were on placement. Some of the comments related to aspects of the student performance that made it hard to determine an overall grade and some comments related to the assessment process itself or to the consequences of the decision.

One aspect of the student performance that made assessment more difficult was general mediocrity. Students who were clearly excellent or clearly weak were easy to grade, however those in the middle were more difficult, as Wayne and Adam both explained. Part of the difficulty of assessing mediocre students was that their performance may not have stood out in the group of students that were on a placement at any one time. Supervisors may have had trouble remembering details of the performance when it was time for grading them and would "scratch their heads and go, mm, well they're probably alright and just give them a, a middle of the road grade, neither good nor bad" (Wayne).

Students who were very quiet and would only speak when spoken to were also very frustrating and difficult for supervisors to assess. Harry explained that “it was really hard to assess her, she was just very, very quiet, very reserved” and Gary described how he found it “impossible to know. Should I just fail them, because they haven’t said anything all week? Or do I, you know, am I optimistic about what they’re actually taking in and pass them?”

Another difficulty related to student performance was assessing students who were weak only in particular areas. Jean commented on the difficulty of knowing how to grade a student with excellent theoretical knowledge but poor ability to apply it. Ed and Ruth spoke of difficulties grading students when the area of weakness was a nontechnical skill such as a personal or social characteristic, or communication. Ruth felt it was “wrong” that communication and engagement should influence her judgement as much as she felt they did. Wayne felt that assigning a fail grade was not easy even when he felt strongly that a student was potentially dangerous because they did not know their limitations, describing it as a “struggle”.

Judgement of performance was also difficult when supervisors felt it was influenced by factors that should not influence their judgement, such as student illness or cultural differences. For example, Harry discussed being unsure whether a student “had some particular health issues that were actually influencing her overall performance” and Gary asked “how much leeway do we give them for their language?” Another difficulty supervisors faced was the potential that the student’s behaviour had been influenced by their own behaviour and whether this should be accounted for in the overall judgement. Martin described being “a little bit tolerant” when a student’s poor engagement “might have just been her interaction with [him]”.

The assessment process itself sometimes contributed to difficulties in assessing students. One aspect was the limited contact supervisors might have had with a particular student during the placement. Ben and Wayne both colourfully commented on the short time over which students are observed and the limited opportunity this gives for making a judgement: Ben described how it would take something extreme “like someone would have to punch me in the face I think to think they should fail in a week’s time”; Wayne compared the process to speed-dating. Nancy commented on the need for direct purposeful observation of students because otherwise the lack of “concrete evidence” made it “hard to make a decision”.

Another difficult aspect of the assessment process was differences in opinion of staff members on the merits of a student's performance. Oscar described how sometimes there would be "four good assessments for one student and one poor one" amongst him and his colleagues in the practice and how "the fact that I think that someone is excellent doesn't mean that they are excellent is someone else's eyes". Todd described a situation where he was "adamant" a student should fail but "others didn't agree with me". Wayne thought it no surprise that a student might fail one placement but not another as he felt it indicated "a bias or an individual character trait that pisses somebody off". However, that such differences in opinion could lead to some unease was hinted at by Simon when he expressed a sense that a decision to fail was better justified when it was in agreement with the opinions of others. In trying to account for the different viewpoints of a group of supervisors in making an overall decision, Oscar felt he often ended up awarding a grade that did not reflect his own opinion.

The sensitivity of the assessment instrument to capture the fine detail of a performance was also a source of difficulty mentioned by two supervisors. Oscar commented on the idea that capturing the grade in terms of five overall levels did not capture the richness of an individual performance and the differences between students. In contrast, Wayne thought that the instrument was only suitable for making broad judgements and not capturing particular detail. Wayne, in particular, seemed very uncomfortable with the validity of the judgements he was being required to make, stating that he thought "the whole thing is rubbish but we just fill it in because we have to".

Lastly, the implications and consequences of the decision to award a particular grade made it difficult for supervisors at times. Walter expressed difficulty arising because failure implied some sort of extreme weakness, which made it hard to fail students who just had an overall poor performance, saying that "a fail just seems like, she'd have to have done something so horribly detrimental, you know, or cause harm or just, you know, need something that so severe." Nancy spoke of concern about "the consequences of a fail or marginal grade for that student". In another comment, Nancy also spoke of the difficulty in giving honestly negative feedback to a student because "to only give them negative feedback would be devastating to them". Todd referred to the consequences for society of making the wrong decision because there was no other backstop (such as an independent licencing examination) to ensure that only competent students could become licenced to practice.

Interrelatedness of themes

There was a great deal of co-occurrence of themes, with supervisors discussing more than one at once. Dual coding was required to capture themes successfully. For example, communication and social interaction frequently co-occurred, as illustrated when Ed was describing the way a student spoke to staff when asked to do something:

> Actually borderline, they were not very polite. They certainly didn't see a teacher-student relationship. They didn't seem to grasp that as a concept. (Ed)

In this example, Oscar also made reference to social interactions when describing a student's communication and included elements of engagement as well:

> And you know straight away the people that are easy with their communication and the people aren't showing for whatever reason they don't interact, like maybe they're not interested. She was definitely not a person who was very interactive during those tutorials. (Oscar)

Here, Jean is discussing a student who could be trusted to communicate with clients in a way that did not increase their anxiety about their animal, and who showed an understanding of how to manage a social interaction with a client:

> Sometimes students pick up on client anxiety and almost buy into that anxiety to show them they're taking things seriously. And this particular student would be a little bit more relaxed and he also wouldn't come out and say something like that without discussing it. (Jean)

When explaining how a student communicated with staff, Nancy also touched on themes of application of knowledge (in planning) and engagement (in thinking things through):

> So part of [communication with staff is] about discussion of cases, that they can summarise it and communicate the important points and put together a plan for diagnostics and treatment. Not necessarily that it's right but that it's presented in a logical systematic fashion that they've obviously thought through it. (Nancy)

Further examples can be seen in quotes throughout this chapter.

In addition, certain theme and valency combinations tended to be present in the descriptions of the same student, suggesting an association, even though they were not necessarily spoken of together. Engagement and trustworthiness were two themes that seemed associated in this way. For most students in which both were discussed by supervisors, those with positive engagement were also considered trustworthy and those with negative engagement were not.

This relationship is not unexpected, as the behaviours of students who were not engaged, such as acting as if they did not care, thought they already knew more than they did, or were not confident enough to act, would naturally not engender trust. In addition, if students were engaged but the engagement was too late, ineffective, or not consistent, they were not considered trustworthy.

Similarly, there was an association of the valency of engagement with impact. For students where supervisors discussed both engagement and impact, those with positive engagement always had a positive impact and those with negative engagement always had a negative impact. Once again, this intuitively makes sense as students who are engaged are helpful and pleasant to teach and those who are not require extra work and attention from supervisors. Though the numbers were small, there often seemed to an association between the valency of engagement and communication, and between engagement and social interaction. Such relationships are not surprising because students need to communicate and interact socially in order to demonstrate engagement. A strong association between the valency of engagement and applying knowledge was also seen. Therefore, engagement was an underlying thread in several themes, which could contribute to a supervisor's opinion of students in these other areas.

There were also associations between the valency of trustworthiness and other themes. Positive impact or caring for animals was always associated with trustworthiness and negative impact or caring for animals was always associated with a lack of trustworthiness. Positive aspects of applying knowledge, communication, social interactions, and personal functioning also seemed to contribute to trustworthiness in many cases, and negative aspects of technical and animal skills with not being trustworthy. Again these associations are not surprising, but serve to indicate how different aspects of performance can be quite distinct but yet interrelated.

Differences between what was important in excellent and weak students

There were differences between what was important in excellent and weak students. The most common positive aspects of excellent students were their engagement and their

application of knowledge (Figure 4.2). Engagement and application of knowledge were also the most common weaknesses of weak students, closely followed by trustworthiness. However after engagement, more supervisors spoke of trustworthiness than other themes in their spontaneous descriptions of weak students, suggesting these were the things that first came to mind when thinking of weak students (Figure 4.1). It may be that while application of knowledge was weak in the majority of weak students, it is a common weakness and does not differentiate weak students from average ones. In support of this a similar number of marginal students had poor application of knowledge to the weak students, however whether this is also true of more average students would require further study. Engagement and applying knowledge were also the most frequent weaknesses of marginal students and technical and animal skills, knowledge, and social interaction were their most frequent strengths. However in their spontaneous descriptions, more supervisors spoke of engagement, communication, and then knowledge than other aspects of student performance, suggesting that these are what first came to mind the most often with marginal students.

Supervisor use of observation, explanation, and inferences

Supervisors' descriptions were a mixture of observation, explanation, and inferences. They often related what had happened or gave examples of what the student did or said. For example, Simon described a student "holding things upside down" and Jean described how a student would "tell you what he had found". Supervisors also described events and behaviours they had not directly observed, but inferred from subsequent behaviour. For example, supervisors often inferred that students had been reading and studying, by the knowledge they would display the next day. Ruth explained that she inferred a student could collect information from a client well, based on the information the student related to her later. Reports from other staff members were also important sources of information. Oscar described how a student had behaved in a tutorial as related to him by another staff member. The nursing or technical staff, and sometimes clients, also provided information in the form of observations and opinions. Supervisors made inferences about the thinking processes of students. For example Martin described seeing the "cogs turning in his mind" and Walter talked of "getting that information, processing it and then delivering it back, whether it's an answer or a question, but the thought process, you can tell is, has gone on". Supervisors' descriptions also contained attributions about students' attitudes and dispositions. For example, Walter inferred that a student "didn't really care about communicating with the farm

manager" and Harry inferred that a student "wanted to learn". Julia described a student that she thought had a "constant desire to keep me happy by saying the right thing". Inferences about the student's attitude and disposition were often mixed together with descriptions of observations and information from other sources as supervisors explained what happened and what they interpreted at the same time.

Strengths and limitations of Phase 1

This phase of research investigated the qualities of student performance that supervisors value. The research design had been used previously in medical and social work educational research (Bogo et al., 2006; Ginsburg et al., 2010). It involved semi-structured interviews with supervisors, which elicited descriptions of excellent, weak, and marginal students they had worked with. The aim was to address the research question by building up a picture of what supervisors considered, and what was important to them. By doing this outwith the context of the assessment instrument, it aimed to tap into values in use, rather than intended.

Semi-structured interviews and content analysis were appropriate methods to meet the aim of this phase. The open nature of the questions generated rich descriptions of the themes that were important to supervisors and gave insights into supervisors' thoughts and attitudes, their expectations of students, and what student competency meant to them. However interviews also have limitations as a data collection method, as Silverman (2014) discusses. As an account given by the interviewee, they reflect the understanding of participants about what the interviewer wants them to discuss, and what they choose to, or think important to, say. They are limited by the expression of experiences, emotions, and values as language, influenced by norms of expression, and interpretation of that expression by the interviewer. The resulting understanding is therefore a shared construction of the supervisors and me.

The sampling was from a range of subdisciplines, genders, and ages and was able to capture a range of views. Sampling appeared to be sufficient because all key ideas that seemed of importance to supervisors were able to be captured in the thematic framework, and each theme was used by at least three supervisors. There was a great deal of commonality in theme use by supervisors, even across subdisciplines. However, it must be acknowledged that the

sample size was small and there may be supervisors who hold other views that have not been captured.

Another limitation is that supervisors only described one student from each category of excellent, weak, and marginal. Therefore, the students described might not represent all the patterns of performance that students can display. In addition, the semi-structured interview method potentially limited the supervisor's ability to say all that they might have said, by constraining them to discuss particular students in a particular order. However, these limitations were also an important strength of the research design for this phase. By asking supervisors to recall and describe a particular student, as was also done by Ginsburg et al. (2010), it gave insight into the holistic picture that supervisors held of student performance and how supervisors trade-off aspects of performance.

Another aspect of the research design that proved useful was asking about marginal students. Because marginal students are on the borderline between failing and satisfactory performance, the themes described were often similar to those of weak students and in this respect did not add much to the analysis. However, supervisors' descriptions of marginal students were very helpful in illuminating the difficulties of judging competency.

The questions were framed so that they did not ask directly about competency or assessment criteria. Although an indirect approach to determining the meaning of student competency to supervisors, a strength of this design was that it removed the constraints of these frameworks. This meant it was more likely that supervisors spoke of their own views than what they felt they were supposed to think. However, it is clear that supervisors were thinking of the assessment framework when discussing students, particularly given that almost all of them raised issues about the difficulty in grading students on placements. Therefore, the responses were not isolated from the assessment criteria and competency frameworks and may have been influenced by them.

The analysis presents my interpretation of the data and other interpretations are possible. Aspects of my analysis which strengthen the credibility of the findings include: (1) the use of verbatim transcriptions and audio recordings to code from; (2) the verification of transcripts by participants; (3) the systematic application of methods; (4) the coding of a large number of

different potential themes in all interviews before looking for patterns and commonalities and consolidating themes based on similarity, frequency and emphasis given; and (5) the use of matrix summaries to verify coding. As a veterinarian researcher and academic with previous clinical experience, I brought to the analysis a discipline-based understanding of practice-based teaching, assessment, the veterinary competency frameworks, and the administrative environment that aided interpretation of the data. However, the framework of themes is the meaning I made of the data. Others may make a different meaning and indeed, in other similar studies, the clustering of key ideas within themes has been slightly different, resulting in a slightly different framework (Ginsburg et al., 2010; Rosenbluth et al., 2014). Having presented a rich description, I enable readers to make their own meaning, so as to increase the utility of the analysis.

Summary

Out of the interview data emerged a broad set of 12 themes that indicated areas of importance to supervisors when considering the performance of veterinary students on placements. Two important and overarching themes were engagement and trustworthiness. Engagement referred to the student's enthusiasm for and participation in the work and learning activities of the placement. Every supervisor discussed engagement spontaneously during interviews and it was often described first or stated to be important. While positive engagement was extremely valued, negative engagement contributed to a poor opinion of students. One group of unengaged students did not actively participate, and seemed to either not care, think they knew it all already, or were extremely quiet and reserved. Another group participated but nevertheless did not seem to be learning, either because they had misdirected interests or did not seem to be able to take in feedback and improve. Three contrasting examples of student engagement in similar situations, described by supervisors, illustrated that engagement was not the same as liking a subject area.

Trustworthiness was discussed less frequently by supervisors, but a sense of its importance was conveyed by the type of words used, and it was an important contributor to a student's performance being judged to be weak. It was the theme discussed spontaneously most frequently after engagement in weak students. Students who were untrustworthy were never considered excellent and students who were trustworthy were never considered weak or

marginal, so it was a clear differentiator of student performance. No other theme was consistently positive in excellent students and negative in weak and marginal students. Students who were trustworthy were honest, reliable, took responsibility, discerned their own limits, and acted within them. Students who were trustworthy were often also very capable. In contrast, students who were untrustworthy displayed a variety of behaviours including being overly sure of themselves and potentially dangerous to themselves, others, or animals. Others could not be relied upon to notice problems or seek help, or to perform their tasks or do them well, or to understand what was required. Still others seemed to be dishonest in saying what supervisors wanted to hear or misrepresenting situations.

Other themes discussed by supervisors included expected areas of competency such as the student's application of knowledge, knowledge, technical, and animal skills, communication, and social interactions. Of these, application of knowledge was a frequent strength of excellent students and weakness of weak and marginal students. Knowledge was frequently traded off, and was the most frequent weakness of excellent students and most frequent strength of weak students. A high level of technical skill was not expected. Animal handling skills were important to some supervisors, especially when they impacted the supervisor, affected a student's trustworthiness, or involved caring or not caring for animals. Communication, or lack of it, was referred to more than any other theme after engagement in supervisors' spontaneous comments about marginal students, but was less frequently spontaneously spoken of in excellent and weak students. Social interaction was also less frequently discussed spontaneously in relation to excellent and weak students, than in marginal students.

Two other aspects of student behaviour sometimes discussed by supervisors were a student's personal functioning and caring for animals. Personal functioning referred to a student's management of their fears, confidence, shyness, and other personal and emotional issues including illness. Where these affected student performance, supervisors had difficulty knowing whether to account for them in making their overall judgement. Caring for animals was an infrequently used theme, but it seemed to influence the supervisor's judgement positively or negatively, and was an important cause of weakness in some students. It was distinct from animal handling skills in that it encompassed an empathy and liking for animals.

Further themes that were not observations of the student themselves were impact on the supervisor, prospects of the student, and the difficulty in judging competency. Several supervisors referred to a student's positive impact in helping get work done or making teaching more enjoyable, or negative impact in increasing workload for the supervisor or causing irritation, frustration, or dilemmas. The prospects of students in terms of their likely employment or future educational pathway was also spoken of by supervisors, especially to illustrate aspects of their descriptions. A final theme that was important to supervisors was that of the difficulty in judging the competency of students. All but one supervisor discussed this even though they were not directly asked about it. Judging performance was more difficult when there was an absence of demonstrably excellent or weak behaviours and attitudes. Very quiet students were also difficult to assess, as were those who were weak only in particular areas, especially when these were nontechnical areas. Dilemmas such as whether to account for issues such as illness or language, and whether the supervisor's own behaviour may have influenced the student's performance also made judgement more difficult. Supervisors also commented on the limited observation from which to form an opinion, the differences in opinion from different staff members about a student's performance, and the consequences of the decision for the profession and the student.

Many of the themes were interrelated and overlapping. Supervisors often spoke of more than one theme at once, and the level of performance on one theme was often associated with a certain level of performance on another theme. In particular, both engagement and trustworthiness were linked with each other and with many other themes.

Supervisors considered a balance of positive and negative factors when making their judgements. Excellent students did not have to be perfect, however their weaknesses tended not to be very negative or only relative weaknesses. The strengths of weak students were sometimes significant strengths. While engagement was important in all types of students, there were differences in what was next most important in excellent students (application of knowledge), weak students (trustworthiness), and marginal students (communication). All aspects of student performance except trustworthiness could be compensated for by performance in other aspects.

Supervisors descriptions were mixtures of their observations, explanation, and inferences. Reports from other staff members were an important supplement to their own observations. Supervisors used many inferences to supplement their observations and often related both together as part of explaining their opinion. Inferences were made about behaviours they had not directly observed as well as about unobservable aspects such as the student's attitude or disposition.

These findings give some insight into the picture supervisors form of a student's performance, the implicit criteria they use to judge its quality, and the difficulty in judging student competency. The next phases of research aim to gain greater understanding of how this picture relates to the scores supervisors award and to the intentions of the evaluation. The next step is to examine the constructs underlying the scores which is the subject of the next chapter.

Chapter 5:
Dimensionality of in-training evaluation

Findings of the factor analysis—Phase 2

In this phase of the research, the aim was to determine the dimensions of performance that were captured by the in-training evaluation. Although the instrument had been designed to evaluate four domains of performance (professional attitude, clinical skills, knowledge, and communication), whether, in use, it reflected these four domains had not previously been determined. As discussed in the literature review, previous studies in medical education have shown that usually only one or two dimensions were being evaluated in in-training evaluations (Table 2.2, page 33). The one previous veterinary study found three dimensions were being evaluated, even though the instrument had been designed to assess four domains (Fuentealba & Hecker, 2008). These previous studies demonstrate that it cannot be assumed that an instrument assesses the domains it is intended to assess. If it does not, there is uncertainty about what is informing scores and conclusions about performance drawn from the scores are not credible. Therefore, establishing the dimensionality was an important step in assessing the validity of in-training evaluation scores at Massey University. This was performed using exploratory common factor analysis.

Research method

Nature of the in-training evaluation instrument

The in-training evaluation instrument comprised 12 Likert-type items intended to span four broad domains of competency: knowledge (3 items), clinical skills (4 items), communication (2 items) and professional attitude (3 items) (Appendix B, page 253). The items are detailed in Table 5.1. This set of 12 items was used to assess students on most subdiscipline placements. Some specialised subdiscipline areas (pathology, anaesthesia, small animal surgery, and radiology) used adapted sets of 8-12 items to cater for specific differences in the type of work

being performed, although the items spanned the same 4 domains of competency. These subdisciplines were not included in the factor analysis because they had differing items.

Table 5.1: Domains and items on the in-training evaluation.

Domain	Item	Description
Clinical skills	Animal	Animal handling and patient care, including empathy and manual skill.
	Technic	Technical skills, proficiency, including care of patient
	Exam	Physical examination skills including accuracy, thoroughness, organisation
	History	History taking from clients, including thoroughness, organisation and completeness
Knowledge	K_use	Use of knowledge including ability to integrate findings and theory, determine differential diagnoses and make plans
	Kknow	Knowledge of the discipline
	SDL	Self-directed learning including use of resources and effort in sourcing information.
Professional attitude	Assignd	Performance of assigned tasks including completeness, effort, and initiative in going beyond the bare minimum required but not beyond own limitations.
	Judgemt	Professional attitude and judgement, including maturity of demeanour, responsibility, reliability, dependability, interest, motivation, knowledge of own limitations, punctuality, personal presentation, care for animal welfare.
	Partic	Participation including team work, interest, dependability, availability, contribution to discussion.
Communication	Comclin	Communication and interactions with clinical staff, including respectfulness, reliability, responsibility, teamwork, sensitivity to other's concerns, maturity, tidiness and professionalism.
	Client	Communication and rapport with clients, including respect, empathy, integrity and ability to convey medical information and establish trust.

Each item on the in-training evaluation was evaluated on a scale comprising categories of excellent, good, satisfactory, marginal, and fail. An option for not applicable was also included. Descriptors characterising performance at each of these levels was provided for each item. In addition to the individual items, an overall grade was also awarded on a scale of excellent, good, satisfactory, marginal, and fail. Level descriptors were not provided for the overall grade

and there was no specification of weighting to be applied in formulating an overall grade. Hence, supervisors awarded the overall grade based on their own global evaluation of the student. The in-training evaluation also provided space for written feedback for each of the four domain areas and the overall evaluation.

Sampling and data preparation

A retrospective convenience sample of all in-training evaluations for every veterinary student from the 2012 and 2013 academic years was obtained. The data was obtained from electronic databases into which it had either been directly electronically entered by supervisors located on-campus, or transcribed from paper forms that had been submitted by supervisors located off-campus. The data was downloaded from the databases and transferred to a series of Excel (2010, Microsoft Corporation) spreadsheets. It was anonymised by a research assistant by assigning a unique identifier to each student. Thus, the anonymisation method protected the identity of individual students while enabling serial evaluations of the same student to be identified.

Each response for the Likert-type items on the in-training evaluation was converted to a numerical score as follows: excellent = 10, good = 8, satisfactory = 6, marginal = 4, fail = 0. These values were those in use at Massey University to convert the overall grade on the in-training evaluation to a mark for assessment of students. Items given an evaluation of not applicable were considered as missing data but coded to distinguish them from evaluations that were just missing.

All in-training evaluation data including overall and item scores, covariates, and variables used in the next phase of research were assembled into one Excel spreadsheet. The accuracy of the assembled spreadsheet was checked by manually comparing the data from each evaluation from a random sample of 5% of the students against each separate spreadsheet of original (deindentified) data. Summary data was tabulated and graphically displayed to detect outliers and unexpected values and these were examined and corrected if they were found to be transcription errors. Missing data was sought and obtained from other databases where possible. The date of each placement was used to derive a variable that denoted the week of each placement in relation to when each student first began placements.

Missingness analysis and management

A detailed analysis of missingness was undertaken to support both quantitative phases of the research and is presented in full in Appendix C (page 257). This involved determining the frequency and distribution of missing data, examining the relationships between missingness and other variables, and exploring differences between evaluations with missing data and those that were complete in order to determine the influence of missingness on the validity of the analysis. Missingness not only reduces the sample size, but it can introduce bias if the missingness is a result of an underlying effect that influences the effects of interest (Tabachnick & Fidell, 2013). This can invalidate inferences and limit the generalisations that are possible because the missingness may confine the study to a sample that is not representative. The degree of bias introduced by missingness is related to the proportion of missingness, the strength of association between missingness and data (White & Carlin, 2010), and how the missingness is managed during the analysis.

Missingness was managed using principled methods wherever possible and is summarised here in brief. Detailed results of the missingness analysis are provided in Appendix C (page 257). A low proportion of 7.2% of evaluations were missing the entire evaluation or all item scores. These missing evaluations could not be included in the factor analysis, however were largely confined to a small number of placements in which the in-training evaluation was normally not used. Therefore, their exclusion was not thought to bias the findings significantly.

Amongst evaluations that contained at least one item score, a low frequency (7.9%) of items were missing, mostly because they were scored not applicable. However, these were distributed amongst nearly half of all evaluations (46.5%). Missingness was concentrated in certain items: participation in placements (Partic), history taking (History), physical examination (Exam), self-directed learning (SDL), and client communication (Client). This missingness was managed by maximum likelihood estimation during construction of the correlation matrix for factor analysis using the SAS (version 9.4, SAS Institute) procedure MIXED. Maximum likelihood estimation is a recommended method for managing missingness in the dependent variable (Schafer & Graham, 2002). It enables unbiased estimates and standard errors to be calculated, taking account of information in covariates, as long as the missingness is unrelated to the value of the missing variable. In this case, however, the missingness analysis suggested that item missingness was potentially related to the value of

the item. In other words, the value of the item may have influenced whether the item was missing. This could therefore lead to biased estimates. In order to mitigate this, covariates that were found to be strongly related to the value of the item were also included in the model.

A third type of missingness was present due to differences in the number of evaluations for each student (unbalanced data). Significant imbalance arose because the number of times a student was evaluated depended on their choice of electives, and whether they chose to include extra placements over and above the requirement. This was effectively managed during construction of the correlation matrix for factor analysis by linear mixed modelling with maximum likelihood estimation using PROC MIXED. This type of modelling can accommodate differing numbers of evaluations and account for the uncertainty produced in the calculation of standard errors.

Justification for common factor analysis over other techniques

The number and nature of the dimensions Massey University's in-training evaluation actually assessed had not been previously determined. It was therefore appropriate to conduct an exploratory analysis to determine the number of factors and how each item on the in-training evaluation contributed to them, rather than testing a single hypothesised structure using confirmatory procedures (T. A. Brown, 2006). The common factor model was used rather than the principal components model, because this research assumes that there are underlying factors that relate items to each other. The common factor model divides the total variance of each item into that which is in common with other items and that which is unique to that particular item, including error variance (Gorsuch, 1983). In contrast, principal components analysis divides all the variance between the factors. This means that conceptually it does not model the presence of latent common factors. Principal components analysis is useful to determine which reduced combination of variables retains the most information and best reproduces the data but is less well suited when the aim is to uncover the structure of unobserved latent factors (Fabrigar et al., 1999). By modelling common factors separate from measurement error, exploratory common factor analysis provides a solution that is more realistic (Fabrigar et al., 1999) and potentially more replicable in other samples (Gorsuch, 1983) or in future confirmatory factor analyses (Floyd & Widaman, 1995). Also common factor analysis has been found to give less biased loadings than principal components analysis (Snook

& Gorsuch, 1989). The latter method tends to have systematically inflated loadings, especially when there are few variables and therefore may suggest items are related to factors when this is not the case. In practice, however, either method usually leads to the same conclusion when reliabilities are high and the number of variables is low (Velicer & Jackson, 1990).

Factor analysis method

Factor analysis involves a number of steps each with a number of options. The options are reviewed in Appendix D (page 275), and the methods used are also detailed. Here a brief outline is provided.

The use of ordinal data potentially violates distributional assumptions of factor analysis which assumes variables (in this case item scores) are normally distributed and residuals are also normally distributed and independent (Tabachnick & Fidell, 2013). The significance of this violation was evaluated by examining the degree of skewness and kurtosis of the data. Estimates of parameters such as eigenvalues and pattern and structure coefficients were considered accurate with maximum likelihood estimation if skewness and kurtosis were between -2.0 and 2.9 because levels less than this have been shown to cause little or no bias with maximum likelihood estimation (Muthén & Kaplan, 1985). Standard errors were considered accurate if skewness and kurtosis were between -1.0 and 1.0. Because values fell outside this range, the more conservative significance level of α=0.01 rather than the usual α=0.05 was used in order to limit type I errors, because standard errors tend to be overly small in that case (Muthén & Kaplan, 1985). The large sample size was considered suitable for factor analysis (Fabrigar & Wegener, 2012).

The dataset contained multiple observations on each student, sometimes by the same supervisor or supervisory group. The observations were therefore not independent. To account for the intercorrelations between these repeated measurements, the method of Cook et al. (2010) was used. This involved using a linear mixed model to estimate a correlation matrix that appropriately accounted for the repeated measures, then performing the common factor analysis on this adjusted correlation matrix. For the purposes of comparison, a Pearson correlation matrix of the raw data was also calculated, in which the dependency between repeated measures on the same student over time was ignored.

The suitability of the adjusted matrix for factoring was assessed by checking for sufficiently large correlations and sufficiently small partial correlations to suggest the presence of common factors. In addition Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy were assessed. Extraction of factors was undertaken using maximum likelihood estimation in SAS PROC FACTOR with squared multiple correlations as the initial communality estimates. Factors were rotated obliquely using direct oblimin rotation. The 99% confidence intervals around both the structure and pattern coefficients were used to determine whether loadings were significantly different to zero, significantly different to 0.4 (an arbitrary moderate loading) and significantly different to 0.6 (an arbitrary strong loading).

Determining the number of factors to retain is challenging and therefore a number of procedures were examined, as has been recommended (Bandalos & Finney, 2010; Beavers et al., 2013; MacCallum, 2009; Ruscio & Roche, 2012). The goal is to find the number of factors that best represent the underlying constructs, acknowledging that models are not true representations, but, rather, are useful simplifications (MacCallum, 2003). Fabrigar and Wegener (2012) suggest that the number of factors retained should be that which gives a substantially better fit than one fewer factor, but not appreciably worse than one more factor, and provides interpretable factors that relate to constructs of interest. Rotated factor solutions were therefore produced for different numbers of factors and examined to determine their interpretability and model fit.

Procedures used for determining the number of factors were those which were most likely to be accurate given the characteristics of this data determined by a preliminary analysis. These characteristics were a large sample size; items assessed using a 5-option Likert-type response format; mild-moderately skewed and kurtotic distribution; few factors; few variables per factor; strong factor loadings; and strong interfactor correlations. Where procedures indicated different numbers of factors to retain, the greatest consideration was given to procedures reported to be more accurate for this type of data, and to the usefulness and interpretability of the solution. Some procedures were also included for comparison purposes, even though they are known to be less accurate, because they are commonly reported in the literature on factor analysis of in-training evaluations. Thus the procedures undertaken were the eigenvalues greater than one rule, the scree test, parallel analysis, minimum average partial method, chi squared test of model fit and chi squared difference test, Tucker-Lewis index (TLI), root mean squared error of approximation (RMSEA), evaluation of residuals and partial

correlations and interpretation of the importance of factors. The use of the variance explained was considered but was not performed. A detailed discussion of factor analysis methods used is presented in Appendix D (page 275).

Higher-order factor analysis method

Items on the in-training evaluation were rotated with an oblique rotation during factor analysis because it was expected that the factors would be correlated. Interfactor correlations, which are part of the output when oblique rotations are performed, were then examined to determine if this was likely. Interfactor correlations greater than 0.5 were considered substantial as these are often reported to indicate strong interfactor correlation (Garrido, 2012; Warne & Larsen, 2014; Yang & Xia, 2015).

Gorsuch (1983) recommended that higher-order factor analysis be undertaken whenever an oblique rotation is used in order to fully characterise the factor structure. To do this the interfactor correlation matrix from factoring the items was used as the sample correlation matrix for higher-order analysis. A common factor analysis method was used to extract higher-order factors. This consisted of maximum likelihood estimation with squared multiple correlations used as the initial communality estimates, as was done in the factor analysis of the items. If more than one higher-order factor was possible, an oblique rotation using direct oblimin rotation was performed, allowing the higher-order factors to be either correlated or uncorrelated. In determining the number of higher-order factors to retain, solutions were evaluated based on the strength and pattern of factor loadings, the interpretability of results and examination of the scree plot. Statistical tests for the significance of factors were not used because the distribution of first order factors depends on the rotation used and standard errors cannot be calculated (Gorsuch, 1983). Because the number of higher-order factors, if present, was anticipated to be small, even higher-order factors with as few as two substantial loadings were considered potentially important. This approach has been suggested to be appropriate by Gorsuch (1983).

Method for construction of three-dimensional graphs

The adjusted correlation matrix was graphed in three dimensions to illustrate the patterns of correlations amongst the items. Factor analysis is a method of exploring the intercorrelations between variables (Furr & Bacharach, 2014). A group of variables that are highly correlated with each other have something in common which is the latent factor (McDonald, 1985). There may be multiple groups forming clusters of intercorrelated variables with less correlation between groups. Each cluster represents a different latent factor. Clusters of intercorrelations can often be seen when examining the numbers in correlation matrices, and this is a basic form of factor analysis but becomes difficult when there are more than a few variables (Furr & Bacharach, 2014). Graphing the correlation matrix in three dimensions, with the degree of correlation forming the *z* (height) axis, allowed visualisation of the correlations between variables more easily than examining the numbers.

The graph was constructed in Excel. A three-dimensional surface chart type was used with the two sides of the correlation matrix as the *x* and *y* axes. The *z* axis was formed by the correlation values. This gave a terrain-type view with the height of the contours given by the degree of correlation. Colours were used to indicate the height of contours, as is done in contour maps, by colour coding various ranges of correlations. The order of variables within the correlation matrix was arranged so that variables with similar correlations were grouped as close to each other as possible while maintaining relationships with other variables. For ease of visualisation the diagonals were excluded from the graph.

Findings

Sample

The data comprised 3466 evaluations performed on 197 students who were in their final year of the veterinary degree in 2012 and 2013. Each student was evaluated a mean of 17.6 times, with most students receiving between 12 and 24 evaluations and one student receiving 35 evaluations because they repeated the year. The distribution of scores for each item on the in-training evaluation is shown in Table 5.2.

Table 5.2: Distribution of scores awarded for each item in 3215 evaluations in which at least some items were scored.

Variable	N	N Miss	Mean	SD	Min	Median	Max	Skew	Kurtosis
Animal	3150	65	7.77	1.42	0	8	10	-0.43	0.88
Technic	3132	83	7.78	1.38	0	8	10	-0.67	1.64
Exam	2855	360	7.74	1.38	0	8	10	-0.63	1.74
History	2599	616	7.77	1.48	0	8	10	-0.51	1.08
K_use	3124	91	7.73	1.40	0	8	10	-0.88	2.84
Kknow	3167	48	7.69	1.35	0	8	10	-0.51	1.19
SDL	2288	927	8.11	1.44	0	8	10	-0.67	1.13
Assignd	3162	53	8.36	1.35	0	8	10	-0.58	1.33
Judgemt	3193	22	8.37	1.44	0	8	10	-0.73	1.31
Partic	2880	335	8.47	1.45	0	8	10	-0.75	0.75
Comclin	3184	31	8.51	1.49	0	8	10	-0.83	0.94
Client	2817	398	8.26	1.44	0	8	10	-0.70	1.02

Note. Skewness ranged from -0.43 to -0.88 and kurtosis ranged from 0.75 to 2.84. Abbreviations: N: number of items scored; N Miss: number of items with missing scores; SD: standard deviation; Min: minimum; Max: maximum; Skew: skewness.

Factor analysis

Suitability of the matrix for factoring

All items on the evaluation were highly intercorrelated with correlations ranging from 0.40-0.73 (mean 0.51) even after the intercorrelations between repeated measures on each student had been removed (Table E.1 in Appendix E, page 291). Before accounting for the repeated measures, intercorrelations were somewhat higher with ranging from 0.47-0.77 (mean 0.59). A moderate proportion of variance could be attributed to common factors, with initial estimates of communalities ranging between 0.43 and 0.62 (Table E.2 in Appendix E, page 291).

The fact that all correlations in the matrix for analysis were greater than 0.3 suggested that common factors may be present. Bartlett's test of sphericity was highly significant (p<0.001) and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.93, both of which confirmed the suitability of the matrix for factor analysis. Examination of the partial correlations controlling for other variables revealed generally small values (median 0.05) suggesting the presence of common factors, but with some higher partial correlations (maximum 0.48). Five partial correlations were over 0.3.

Extraction and rotation of factors

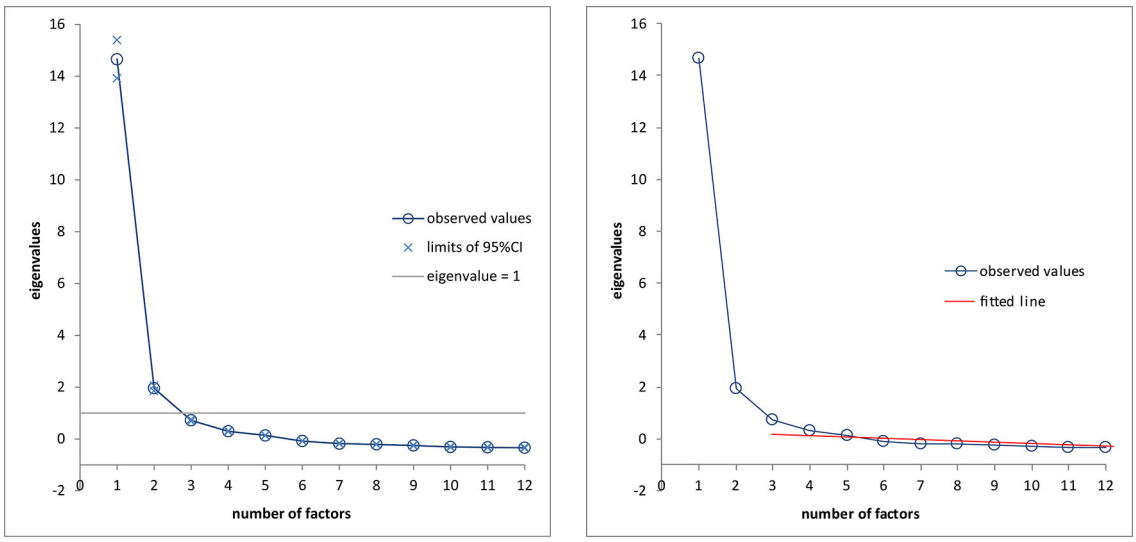
Extraction and oblimin rotation of 1, 2, 3, and 4 factors was performed. Extraction of 5 factors resulted in a Heywood case[2]. The results of analysis for the number of factors to retain are shown in Table 5.3 and Figure 5.1. Two factors were indicated by the eigenvalue greater than one rule, parallel analysis, the minimum average partial (MAP) test, and the size of residuals. More than four factors were suggested by the scree test, the chi squared test of model fit, the chi squared difference test, and RMSEA.

Table 5.3: Number of factors suggested by each method used to determine the number of factors to retain.

Method	Number of factors suggested
Eigenvalues > 1	2
Scree test	5
Parallel analysis	2
Minimum average partial (MAP)	2
Residual size	2
Chi squared test of model fit	more than 4
Chi squared difference test	more than 4
Tucker-Lewis index (TLI)	4
Root mean squared error of approximation (RMSEA)	more than 4

[2] Heywood cases occur when estimated communalities equal or exceed 1 and are improper solutions because communalities are squared correlations, and should always lie between 0 and 1 (SAS Institute Inc., 2014). Communalities represent the variance accounted for by the variable and therefore a communality of 1 or greater implies that the variable accounts for 100% or more of the variance, which is not possible (Fabrigar et al., 1999; McDonald, 1985). Heywood cases are therefore nonsensical and factor solutions containing them were therefore discarded from this analysis, following the advice of several authors (T. A. Brown, 2006; R. B. Kline, 2011; B. Thompson, 2004; Wothke, 1993). There are numerous causes of Heywood cases in exploratory factor analysis including a non-normally distributed data, specifying too few or too many factors, and small samples (Wothke, 1993). McDonald (1985) considers the most common reason for the occurrence of Heywood cases to be factors having only one or two strongly loading variables with the rest having very small loadings, close to zero. It is important therefore to have at least 3 or 4 strongly loading variables for each factor. This then provides a natural limit to the number of factors that it would be possible to define from a certain number of variables, even if there were more latent factors present. In this dataset of 12 variables this natural limit is 4 factors.

Both residual and partial correlations became progressively smaller as more factors were fitted (Figure 5.1). With two factors extracted 88% of residuals were less than 0.05 and 82% of partial correlations were less than 0.1. Improvement in residual size with extraction of more than two factors was somewhat attenuated compared with that seen between extraction of one and two factors.

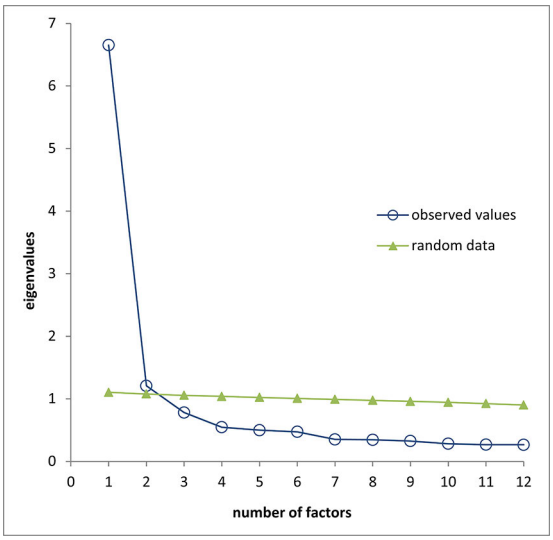


Eigenvalues greater than one rule

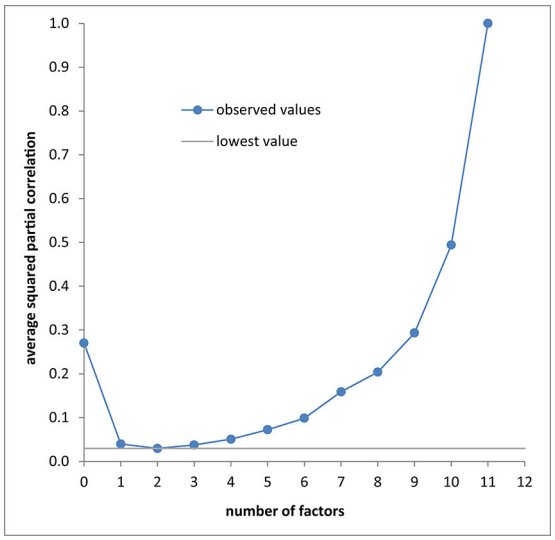
Scree test

Figure 5.1: Graphs depicting the results of procedures to determine the number of factors to retain (continues on next page). The eigenvalue greater than one rule retains all factors with eigenvalues greater than one. The scree test retains all those above a line fitted by hand. Parallel analysis retains all those greater than randomly generated eigenvalues. The point of inflection indicates the number of factors in the MAP test. Residual covariances ideally fall below 0.05 and the number of factors is a trade-off between gains in reducing residual values and the increased complexity of an additional factor. Partial correlations should ideally have no significant remaining correlations. For the TLI and RMSEA the number of factors is the fewest that produce a TLI of greater than 0.95 and an RMSEA of less than 0.02.

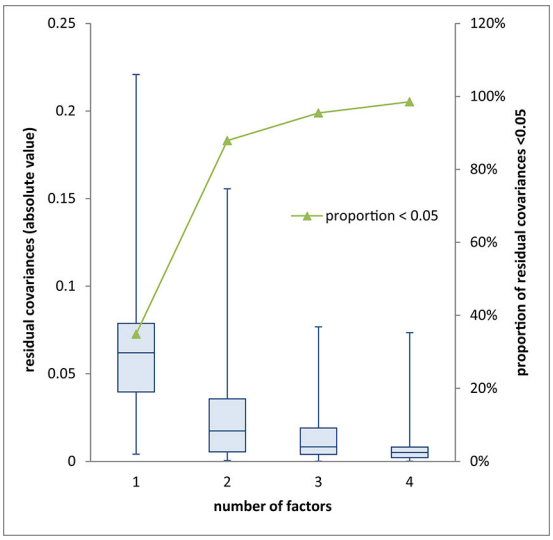
Abbreviations: MAP test: minimum average partial test; TLI: Tucker-Lewis index; RMSEA: root mean squared error of approximation.



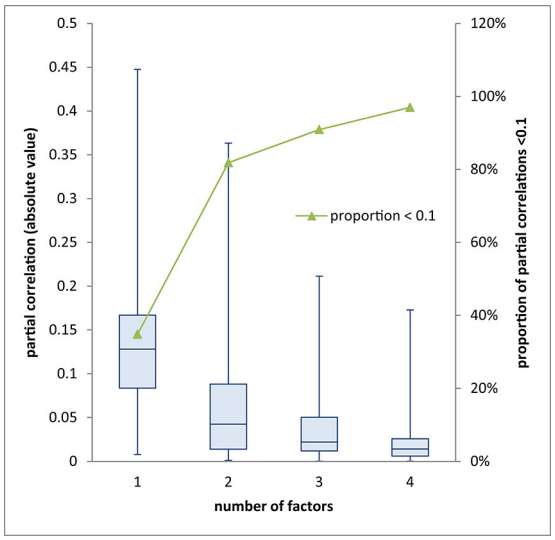
Parallel analysis



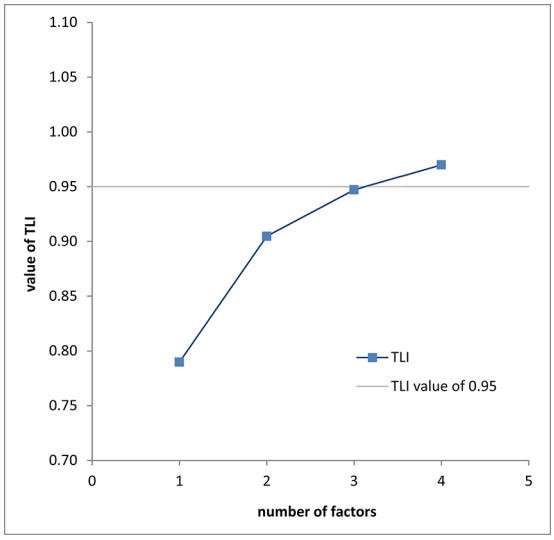
MAP test



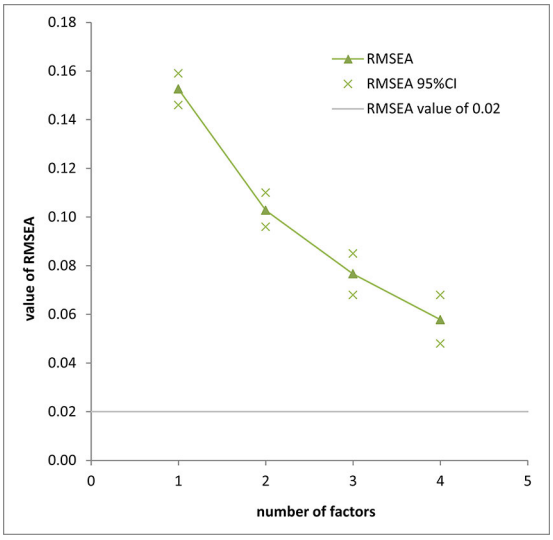
Residual covariances



Partial correlations



TLI



RMSEA

Figure 5.1: Continued.

Factor solution

Strong factor loadings produced interpretable solutions with relatively simple structure with extraction of 1, 2, 3, and 4 factors (Table 5.4, with additional tables Appendix E, page 292). As more factors were extracted the number of items loading on each naturally declined so that when four factors were extracted, one factor was only defined by two items and the other three factors by three and four items each. Only a 2-factor solution had all factors defined by at least 3 strongly loading items. The 2-factor solution grouped clinical skills and knowledge items together in one factor and professional attitude and communication items together in another factor (Table 5.4). The item relating to self-directed learning was ambiguously cross-loaded across both factors. A 3-factor solution produced separate factors for the clinical skills and knowledge items, retaining the combination of professional attitude and communication items in the third factor. The self-directed learning item was cross-loaded on the knowledge factor and the professional attitude-communication factor. A 4-factor solution separated the items according to their groupings of clinical skills, knowledge, professional attitude, and communication. Similarly to before, the self-directed learning item was cross-loaded on the knowledge factor and the professional attitude factor.

Table 5.4: Factor pattern coefficients produced by extraction of 1-, 2-, 3-, and 4-factor solutions on the adjusted Pearson correlation matrix, using maximum likelihood estimation and direct oblimin rotation.

Item domains	Items	1 factor	2-factor solution		3-factor solution			4-factor solution			
Clinical skills	Animal	0.69	0.57	0.16	0.70	-0.09	0.14	0.71	-0.08	(0.04)	0.11
	Technic	0.71	0.68	0.08	0.72	(0.00)	0.07	0.72	(0.01)	(-0.02)	0.09
	Exam	0.70	0.90	-0.12	0.82	0.11	-0.11	0.79	0.09	(0.03)	-0.11
	History	0.74	0.82	(-0.01)	0.72	0.12	(0.00)	0.70	0.12	(0.06)	(-0.04)
Knowledge	K_use	0.73	0.73	0.07	0.15	0.68	0.07	0.10	0.77	-0.07	0.11
	Kknow	0.69	0.73	(0.01)	(0.04)	0.83	(0.00)	(0.03)	0.81	(0.04)	(-0.04)
	SDL	0.66	0.35	0.36	(0.00)	0.39	0.38	(-0.01)	0.41	0.33	(0.05)
Professional attitude	Assignd	0.72	0.10	0.68	0.10	(0.02)	0.66	0.12	(-0.02)	0.77	-0.05
	Judgemt	0.76	(0.05)	0.78	(0.02)	0.05	0.76	(0.04)	(0.05)	0.68	0.12
	Partic	0.75	(-0.01)	0.82	(-0.03)	0.05	0.81	(-0.01)	0.07	0.62	0.20
Communica-tion	Comclin	0.73	-0.10	0.90	-0.04	(-0.03)	0.88	(-0.01)	0.06	0.22	0.67
	Client	0.72	0.07	0.71	0.12	(-0.02)	0.69	0.16	0.07	(0.02)	0.66

Note. Shaded cells indicate coefficients that are significantly greater or equal to 0.4 (α=0.01). Coefficients not significantly different to zero are presented in brackets.

Higher-order factor analysis

Interfactor correlations with oblimin rotation were all significantly greater than or equal to 0.5 (α=0.01) suggesting that common factors may be present (Table 5.5). Higher-order factor analysis with rotation of one higher-order factor produced a solution in which all four first-order factors loaded highly (0.74-0.85) onto the higher-order factor (Table 5.6). There were not enough first-order factors to examine a 2-factor higher-order solution, and the solution did not converge.

Table 5.5: Interfactor correlations produced by extraction of 2-, 3-, and 4-factor solutions on the adjusted Pearson correlation matrix, using maximum likelihood estimation and direct oblimin rotation.

	2-factor solution		3-factor solution			4-factor solution			
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	1	0.73	1	0.72	0.69	1	0.74	0.65	0.53
Factor 2	0.73	1	0.72	1	0.59	0.74	1	0.59	0.44
Factor 3			0.69	0.59	1	0.65	0.59	1	0.73
Factor 4						0.53	0.44	0.73	1

Note: the correlation of 0.44 in the four factor solution had 99%CI 0.33-0.53 and therefore was not significantly different from 0.5 (α=0.01).

Table 5.6: Factor pattern coefficients produced by extraction of one higher-order factor on the intercorrelation matrix from the 4-factor solution, using maximum likelihood estimation.

Factors from 4-factor solution	Domain of items in each factor	Items in each factor	Higher-order factor pattern coefficient
Factor 1	Clinical skills	Animal	0.81
		Technic	
		Exam	
		History	
Factor 2	Knowledge	K_use	0.75
		Kknow	
		SDL	
Factor 3	Professional attitude	Assignd	0.85
		Judgemt	
		Partic	
Factor 4	Communication	Comclin	0.74
		Client	

Three-dimensional correlation matrices

Examination of the three-dimensional plot of the adjusted correlation matrix showed that from a high plateau of intercorrelations of 0.40-0.50 rose two distinct “mountains” each with two and three separate peaks respectively (Figure 5.2). These peaks mapped well to the dimensions the in-training evaluation aimed to assess (Figure 5.3).

Evaluation of the three-dimensional graphs of the partial correlations remaining after extraction of each factor provided a visual demonstration of the extraction of factors (Figure 5.4). Extraction of one factor removed the base correlation from the matrix but preserved the pattern of intercorrelations between variables to produce a similar graph to the full matrix. After extraction of two factors the patterns were still present and there was substantial remaining correlation amongst the variables associated with knowledge. Extraction of three factors, however, removed almost all correlation associated with the knowledge variables. With extraction of four factors, remaining correlations reflected bivariate correlations between the clinical skills of history taking and physical examination which are often performed together, and between technical and animal-related clinical skills. Essentially all other correlation was captured by the factors when four factors were extracted and the 4-factor model appeared to have an excellent fit.

The factor structure mirrored the three-dimensional structure found when the adjusted correlation matrix was graphed with the strength of the correlations forming the height (Figure 5.5). The factors in the 2-, 3-, and 4-factor solutions were consistent with the peaks in the three-dimensional graph and the way these peaks were clustered. When it is considered that the base of the three-dimensional graph begins at a substantial correlation of 0.4, a single factor solution can also be conceptualised with all correlations being part of one large “mountain”.

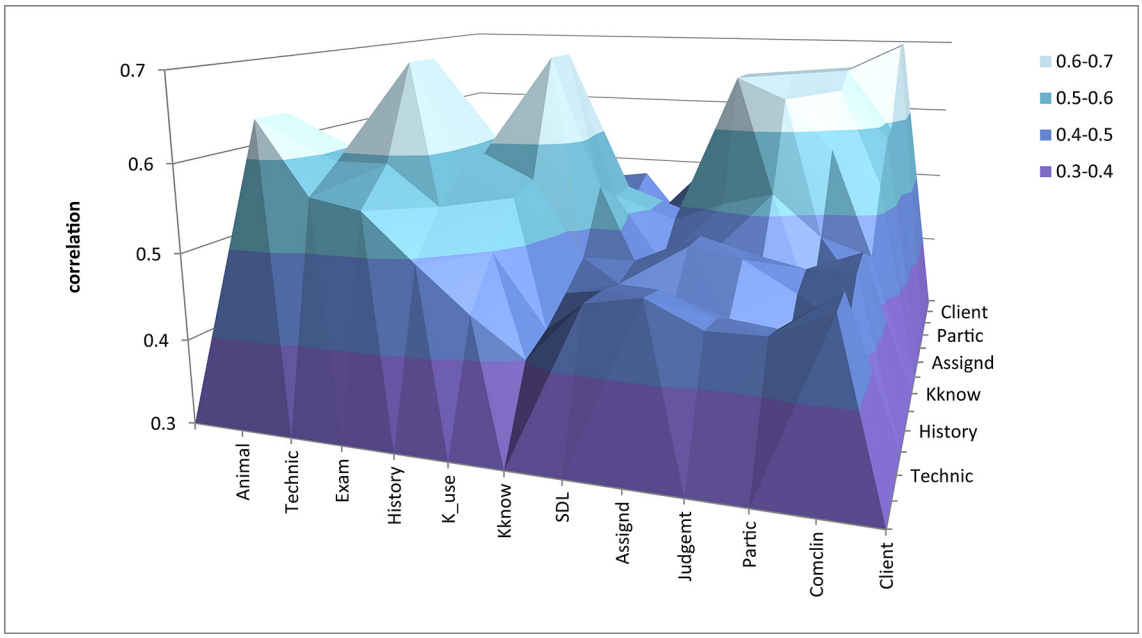
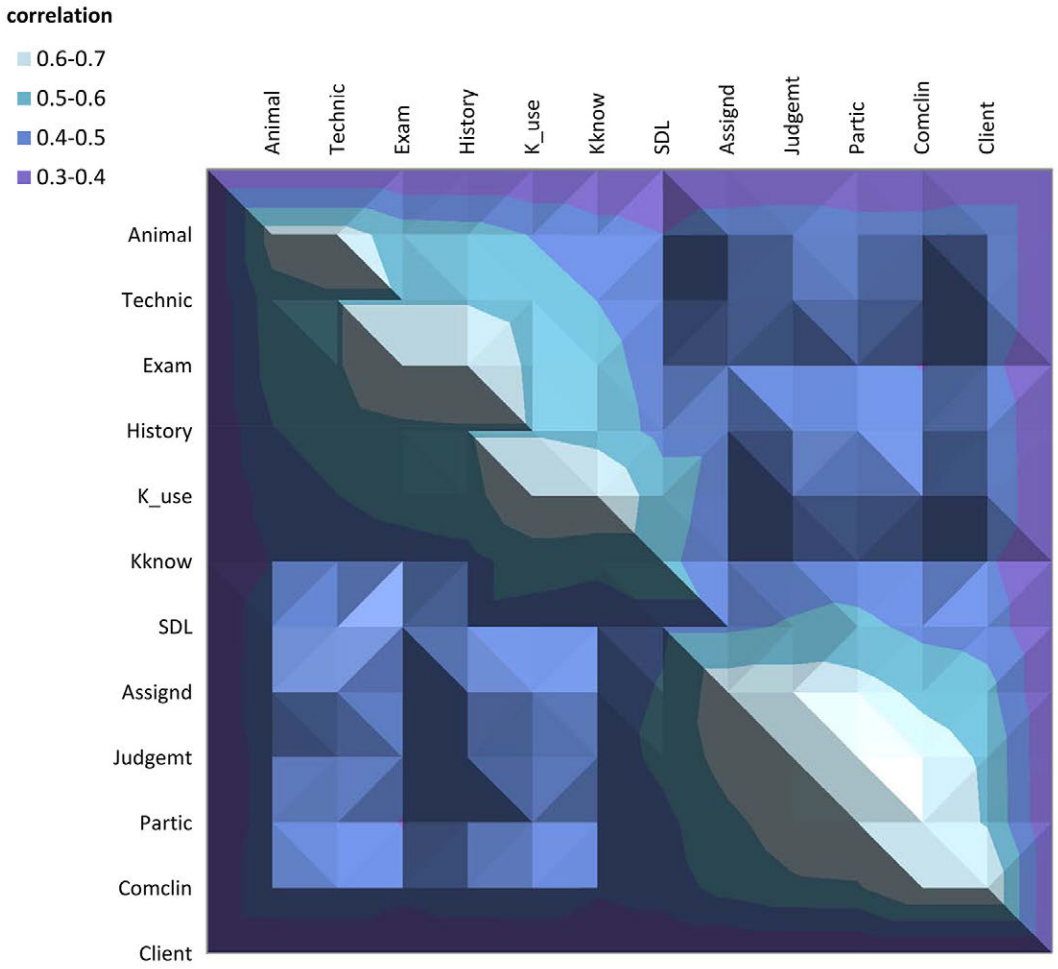
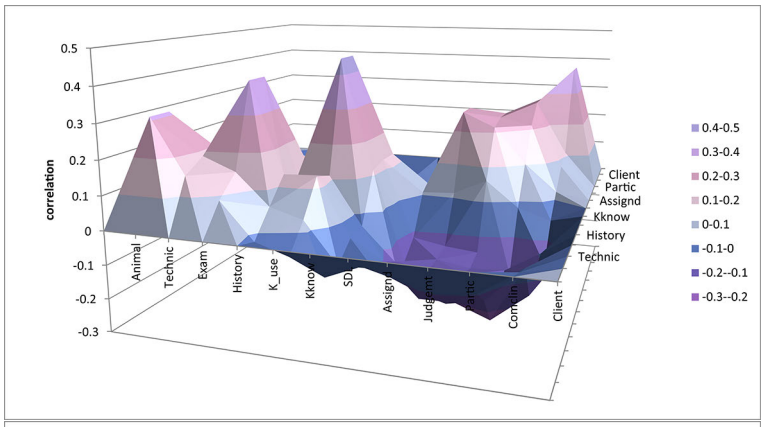


Figure 5.2: Three-dimensional graph of the adjusted correlation matrix. The correlation matrix forms the base of the figure, with the degree of correlation forming the height axis.

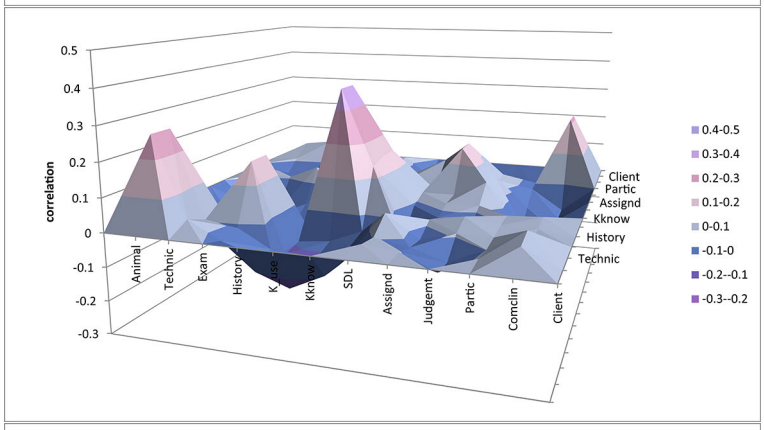


Item domain	Item	Animal	Technic	Exam	History	K_use	Kknow	SDL	Assignd	Judgemt	Partic	Comclin	Client
Clinical skills	Animal		0.64	0.56	0.55	0.50	0.44	0.40	0.47	0.48	0.45	0.45	0.49
	Technic	0.64		0.59	0.59	0.55	0.50	0.42	0.47	0.47	0.45	0.44	0.49
	Exam	0.56	0.59		0.71	0.58	0.55	0.44	0.46	0.44	0.43	0.40	0.43
	History	0.55	0.59	0.71		0.59	0.56	0.47	0.47	0.49	0.48	0.45	0.48
Knowledge	K_use	0.50	0.55	0.58	0.59		0.73	0.54	0.45	0.49	0.48	0.46	0.49
	Kknow	0.44	0.50	0.55	0.56	0.73		0.52	0.44	0.46	0.44	0.40	0.42
	SDL	0.40	0.42	0.44	0.47	0.54	0.52		0.48	0.51	0.53	0.47	0.45
Professional attitude	Assignd	0.47	0.47	0.46	0.47	0.45	0.44	0.48		0.67	0.64	0.58	0.53
	Judgemt	0.48	0.47	0.44	0.49	0.49	0.46	0.51	0.67		0.67	0.65	0.59
	Partic	0.45	0.45	0.43	0.48	0.48	0.44	0.53	0.64	0.67		0.67	0.59
Communica-tion	Comclin	0.45	0.44	0.40	0.45	0.46	0.40	0.47	0.58	0.65	0.67		0.71
	Client	0.49	0.49	0.43	0.48	0.49	0.42	0.45	0.53	0.59	0.59	0.71	

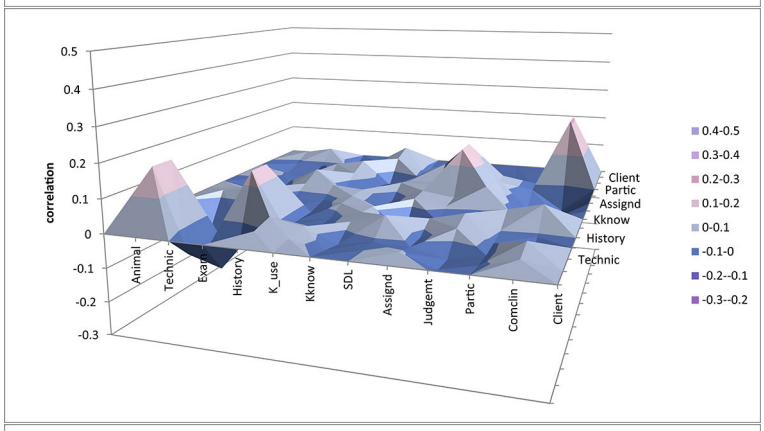
Figure 5.3: Comparison of the adjusted correlation matrix displayed in three dimensions from above and in a table form. Colours indicate the strength of correlation and the same colours are used in both parts of the figure. The three-dimensional graph is a rotated (aerial) view of the same graph presented in Figure 5.2.



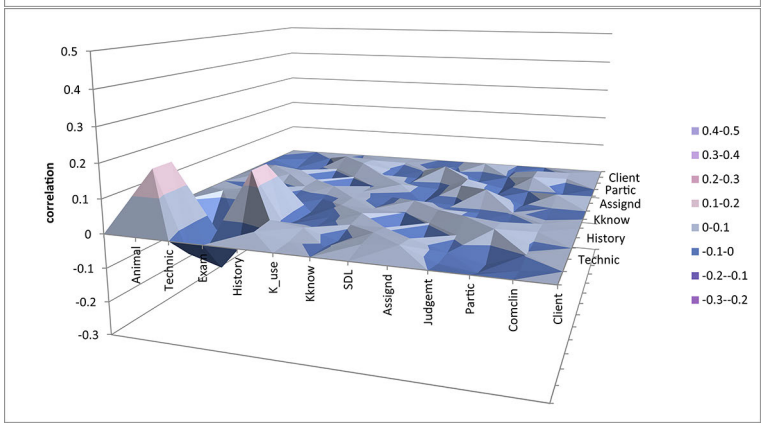
Partial correlations after extraction of one factor



Partial correlations after extracting two factors



Partial correlations after extracting three factors



Partial correlations after extracting four factors

Figure 5.4: Three-dimensional graphs of partial correlations remaining after extraction of factors. The correlation matrix forms the base of each figure, with the degree of correlation forming the height axis.

Item domains	Items	1 factor	2-factor solution		3-factor solution			4-factor solution			
Clinical skills	Animal	0.69	0.57	0.16	0.70	-0.09	0.14	0.71	-0.08	(0.04)	0.11
	Technic	0.71	0.68	0.08	0.72	(0.00)	0.07	0.72	(0.01)	(-0.02)	0.09
	Exam	0.70	0.90	-0.12	0.82	0.11	-0.11	0.79	0.09	(0.03)	-0.11
	History	0.74	0.82	(-0.01)	0.72	0.12	(0.00)	0.70	0.12	(0.06)	(-0.04)
Knowledge	K_use	0.73	0.73	0.07	0.15	0.68	0.07	0.10	0.77	-0.07	0.11
	Kknow	0.69	0.73	(0.01)	(0.04)	0.83	(0.00)	(0.03)	0.81	(0.04)	(-0.04)
	SDL	0.66	0.35	0.36	(0.00)	0.39	0.38	(-0.01)	0.41	0.33	(0.05)
Professional attitude	Assignd	0.72	0.10	0.68	0.10	(0.02)	0.66	0.12	(-0.02)	0.77	-0.05
	Judgemt	0.76	(0.05)	0.78	(0.02)	0.05	0.76	(0.04)	(0.05)	0.68	0.12
	Partic	0.75	(-0.01)	0.82	(-0.03)	0.05	0.81	(-0.01)	0.07	0.62	0.20
Communica-tion	Comclin	0.73	-0.10	0.90	-0.04	(-0.03)	0.88	(-0.01)	0.06	0.22	0.67
	Client	0.72	0.07	0.71	0.12	(-0.02)	0.69	0.16	0.07	(0.02)	0.66

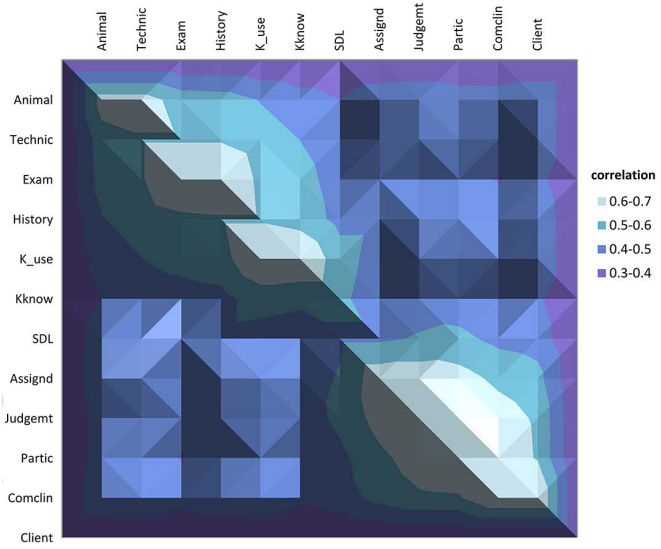


Figure 5.5: Comparison of the factor structure with the adjusted correlation matrix displayed in three dimensions from above. The factor structure is presented with variables in the same order as displayed in the three-dimensional graph. The three-dimensional graph is a rotated (aerial) view of the same graph presented in Figure 5.2.

Note. Shaded cells indicate coefficients that are significantly greater or equal to 0.4 (α=0.01). Coefficients not significantly different to zero are presented in brackets.

Interpretation of the number of factors

The aim of Phase 2 was to determine the dimensions of performance that were captured by the in-training evaluation and to describe the items that contributed to each dimension. The in-training evaluation had been designed to assess four constructs thought to be important dimensions of veterinary competency, but whether it did assess these constructs had not been determined. Previous research on other in-training evaluation instruments in medical and veterinary education had usually found fewer factors, casting doubt on the ability of supervisors to discriminate more than a few dimensions of performance. The findings of this phase suggest a complex factor structure that is interpretable in more than one way. It is appropriate to discuss these interpretations at this point because factor scores are used as data in the next phase of this research. Therefore, in this section I will begin by discussing the interpretation of the factor solution, before moving on to discuss the strengths and limitations of the analysis, and summarising the findings.

The number of factors present is a key aspect of the internal structure of a set of data, but it is also very challenging to determine (Fabrigar & Wegener, 2012), and methods of doing so are reviewed in detail in Appendix D (page 280). Determining the number of factors requires the consideration of the results of multiple procedures (Bandalos & Finney, 2010; Beavers et al., 2013; MacCallum, 2009; Ruscio & Roche, 2012). Agreement between procedures lends support to a solution (Ruscio & Roche, 2012), however, because characteristics of the data influence the accuracy of procedures (Garrido, 2012), not all procedures are of equal value and rules of thumb should be avoided. The solution should represent a conceptually and statistically useful simplification of the data (Fabrigar & Wegener, 2012). The usefulness and interpretability of the solution, as well as the accuracy of each procedure in this data, were therefore considered in determining the number of factors present.

The rotated solutions after extraction of different numbers of factors showed that each produced an interpretable solution that approximated a simple structure. Achieving a simple structure is a goal of factor rotation. Ideally, each variable loads strongly on one factor and insignificantly on other factors (Gorsuch, 1983). As seen in Table 5.4, in all solutions most items had strong loadings of greater than or equal to 0.6 on only one factor with all other loadings less than 0.4 and in many cases not significantly different from zero. Only one item was crossloaded and this occurred in each of the 2-, 3-, and 4-factor solutions. The item

represented self-directed learning, and, as well as loading much less strongly than other items, it consistently co-loaded with items encompassing knowledge and items encompassing professional attitude. This suggests that the item has something in common with both knowledge and professional attitude, either because the construct spans both or because supervisors variably link it with one or the other. The item may thus be ill-defined or difficult for supervisors to judge, and indeed this was the item that was most frequently not evaluated (Table 5.2).

In this data the factors were highly intercorrelated (Table 5.5) and hence each item influenced a number of factors as indicated by the strong crossloading seen in the factor structure coefficients (Table E.3 in Appendix E, page 292). Factor structure coefficients represent the total contribution of a variable (an item) to a factor, including the unique contribution of the item as well as indirect contributions through the correlations between factors (Gorsuch, 1983). In contrast, the factor pattern coefficients (Table 5.4) and reference structure matrix (Table E.4 in Appendix E, page 293) show the unique relationship between a factor and the items, and thus do not include effects mediated through correlations between factors. The clear simple structure evident in both of these demonstrated that a component of the variance of each item was independent of other factors. The findings are, therefore, best explained by a higher-order structure, with one overarching factor and between two and four subfactors. A 1-factor solution could also be considered because the high degree of interitem and interfactor correlation suggests a single general factor (Gorsuch, 1983).

As well as being statistically interpretable, all solutions were also conceptually interpretable. A 4-factor solution grouped items into the four domains of knowledge, clinical skills, communication, and professional attitude that the in-training evaluation was designed to assess. With the addition of the higher-order general factor, it suggests a general competency construct overarching the four dimensions of performance and thus is consistent with conceptualisations of competency as both unitary and multidimensional (Le Deist & Winterton, 2005). A 3-factor solution combined the items from the communication and professional attitude domains into one factor, with separate factors for the clinical skills items and for the knowledge items. A similar 3-factor solution, encompassing professionalism, clinical skills, and knowledge was found in another study of veterinary in-training evaluation (Fuentealba & Hecker, 2008) and 3-factor solutions have also been documented in medical training (Gough et al., 1964; Hojat et al., 1986; McGill et al., 2013; Skakun et al., 1975). A 2-factor solution

grouped communication and professional attitude items together and clinical skills and knowledge items together. Such as a split is not surprising and reflects a dichotomy between discipline-specific skills and nontechnical skills. It has also been found by other authors in medicine (Dielman et al., 1980; Forsythe et al., 1985; Ginsburg et al., 2013; Maxim & Dielman, 1987; McLaughlin et al., 2009; Nasca et al., 2002; Silber et al., 2004; Verhulst et al., 1986; Woloschuk et al., 2010; Woloschuk et al., 2014).

Thus each of these solutions is consistent with potential underlying constructs. Each solution is therefore theoretically appropriate. In addition, each solution is supported by the factor solutions from other research. Therefore various solutions of one to four factors and a higher-order structure are consistent and interpretable solutions. The results of statistical procedures to determine the number of factors gave conflicting results (Table 5.3). As reviewed in Appendix D (page 280), the two methods that were most likely to give accurate determinations of the number of factors to retain, based on previous research and given the characteristics of this data, did not agree. These were the parallel analysis, which determined two factors were present, and the Tucker-Lewis index (TLI), which determined four factors were present. Other indices also indicated either two factors or at least four factors being present. Therefore, these procedures provide good evidence for both a 2-factor and a 4-factor solution. Based on these findings, the data could be variously interpreted as revealing one dimension or a number of dimensions with an overarching higher-order dimension.

Displaying the data three-dimensionally provided a way to visualise all of these interpretations, using the imagery of a landscape. The overall shape is a single high mountain that from a distance would appear as one, but that on approach has two main mountain peaks arising from a high plateau. Each of these is formed by two or three smaller peaks respectively. Each of the peaks maps to the factor solutions for 2-, 3-, and 4-factor solutions and even suggests a 5-factor solution may be present, but undetected because there were only 12 items.

These findings suggest the factor structure is more complex than previously reported and that concluding that any one particular factor structure is present would be an over-simplification of what is actually a complex landscape of separate but related dimensions. A few researchers, for example Cook et al. (2010) and McGill et al. (2013), have also discussed the possibility of more than one interpretation of the number of factors present in studies of workplace-based

assessments, although they settled on a single solution. Most researchers, however, present the results as unproblematically indicating a single best solution. The idea that more than one solution may be equally "correct" may challenge usual practice but is easily appreciated through the visual imagery of the mountain in the three-dimensional graph. The "correct" number of factors may depend on the perspective adopted, as indicated by the purpose of the analysis. For a rich understanding of how supervisors use the instrument, the 4-factor higher-order structure is the most complete, and was therefore used in subsequent phases of this research. For other purposes, such as generalisation to other contexts or datasets, the 2- or 1-factor model might be more useful because they are less specific and a looser model fit, and therefore likely to be able to represent other data.

Strengths and limitations of Phase 2

Strengths and limitations of the factor analysis methods

In addition to the procedures used to determine the number of factors present, factor analysis involves a number of other methodological options, which are reviewed in detail in Appendix D (page 275). A strength of this research was its use of options that, as far as possible without specialised software, best addressed the purposes of analysis and aligned with the characteristics of the data. A common factor model was appropriate to explore data where underlying constructs governing performance were hypothesised and had not been previously determined. Maximum likelihood estimation enabled both the significance and the substantiveness of the relationship of items with factors to be considered, instead of using arbitrary benchmarks. Oblique rotation enabled factor intercorrelation to be modelled, if present as theorised. Formation of the adjusted correlation matrix using linear mixed modelling accounted for the dependency introduced by repeated measures data as well as providing accurate estimates in the presence of unbalanced and missing data. Because the data were ordinal and not multivariate normally distributed, conservative interpretations of significance were used to better support inferences made in the presence of mild to moderate skewness and kurtosis. These methods therefore supported the robustness and credibility of conclusions by avoiding or minimising violations of assumptions inherent in the statistical modelling. However, similar results were obtained when the analysis was repeated using

methods that violated assumptions. It is useful to consider the results of these alternative analyses because they inform interpretation of the findings of this research in the light of previously conducted research that used different methods.

Comparison of results when assumptions were violated

Factor analysis method

Performing the analysis with principal components analysis rather than common factor analysis did not change the result with this data (analysis not shown). This occurs when reliabilities are high and the number of variables is low, as in this data (Velicer & Jackson, 1990). High reliabilities mean that communalities approach 1, thereby approximating a principal components analysis which uses 1 on the diagonals of the matrix for analysis instead of communality estimates. The fewer the variables to be analysed, the smaller the proportion of the whole is made up of the diagonals, and therefore the less influence imperfect reliability causes. Thus, for this data, the same conclusions would have been drawn had principal components analysis been used instead of common factor analysis. For the same reason, other published studies which used principal components analysis may also have had the same result had they performed a common factor analysis.

Rotation of factors

In contrast, performing the analysis with an orthogonal rotation (varimax), instead of oblique rotation as used in this research, reduced the clarity of the factor structure, with increased crossloading (analysis not shown). A clear 4-factor solution was not present and this would have therefore led to different conclusions being drawn about the factor structure. Reduced simplicity (and therefore clarity) of the factor solution is expected when factors that are actually correlated are constrained to be uncorrelated, as orthogonal rotations do (Fabrigar et al., 1999). Furthermore, orthogonal rotations do not allow the solution to be examined for higher-order factors and therefore may not provide a realistic representation of the data.

Dependency of data

Performing the analysis without accounting for the dependency in the data that resulted from repeated measures, increased the correlations between items a small amount, but did not change the factor solution (analysis not shown). The increase in correlations was expected because dependency is manifested as a degree of item intercorrelation, as repeated measures

on the same student are expected to be somewhat correlated with each other. The fact that the change was small and did not affect the factor solution suggests that, for this data, the within-student correlation over time was not a substantial contributor to the model. Similarly Cook et al. (2010) found that the factor solution was unaffected by dependency with repeated miniCEX evaluations, despite the fact that in their data, the reduction in item intercorrelation found when dependency was accounted for was substantial. Although these findings suggest that factor analysis is relatively robust to violations of the assumption of independence, researchers should not assume this is necessarily the case, because changes in the apparent strength of item intercorrelation could alter the apparent relationship between items and factors. This may suggest that more factors should be retained than are needed or that a variable's contribution to a factor is significantly stronger than is the case (Stapleton, 2006).

The presence (or absence) of dependency and how this has been accounted for in the analysis is frequently not mentioned in other factor analytic studies of ITEs. This may be because accounting for it increases the complexity of the analysis and methods to do so have only recently become available to applied researchers. The method used here, as described by Cook et al. (2010) is unusual in being able to be implemented in general statistical software. Other methods utilise multilevel factor models and are available through specialised structural equation modelling software such as MPlus, EQS and LISREL (T. A. Brown, 2006). Application of multilevel factor analysis to ordinal data is more complex and less frequently reported than analysis of linear data. Grilli and Rampichini (2007) demonstrate how to apply multilevel factor analysis to ordinal data using MPlus and Rabe-Hesketh, Skrondal, and Pickles (2004) provide details of the GLLAMM procedure implemented in Stata which brings together generalised linear mixed methods and structural equation modelling and can be applied to ordinal data. This latter method also has all of the advantages of generalised linear mixed models in handling missing and unbalanced data and therefore appears promising.

Ordinal data

As reviewed in Appendix D (page 275), factor analysis is a multivariate linear modelling technique that assumes variables are linear and normally distributed (Tabachnick & Fidell, 2013). However, the score data was ordinal in nature and therefore not necessarily linear, and potentially not normally distributed if significantly skewed or kurtotic. Polychoric correlations could have been used instead as input for the factor analysis and would have better modelled ordinal data (as reviewed in Appendix D, page 276), however would not have accounted for

the dependency, missingness, or unbalanced data. Given that the data were only mildly skewed (-0.43 to -0.88) and moderately kurtotic (0.75-2.84), the non-normality was thought less of a problem for the analysis than the dependency, missingness, and unbalanced data. The linear mixed modelling technique recommended by Cook et al. (2010) to account for repeated measures and derive the adjusted covariance matrix for factor analysis, was also a multivariate linear modelling technique. Another method of modelling the repeated measures in ordinal data, through generalised linear mixed modelling with PROC GLIMMIX (Stroup, 2013), could not output the residual matrix required for factor analysis according to the method used by Cook et al. (2010), and therefore was not an option.

Violations of normality assumptions can cause biased maximum likelihood estimates of parameters such as the factor pattern coefficients, and standard errors, which affects statistical tests for the number of factors and goodness of fit (Fabrigar et al., 1999; Schmitt, 2011). Bias is least, however, when sample size is high, correlation between variables and factors is high, and there are at least five ordinal response categories (Finney & DiStefano, 2006), all of which were features of the data for this research. Simulation studies have shown that mild to moderate degrees skewness and kurtosis, such as present in this data, do not affect the accuracy of pattern and structure coefficients but do bias standard errors, making them overly small (Muthén & Kaplan, 1985). Since confidence intervals would be therefore overly narrow, a more stringent standard of α=0.01 was applied in order to make decisions about the significance of factors. This had the result of widening confidence intervals, giving more conservative interpretations of the significance of factors. Because the factor structure was so clear in this study, the statistical significance of factors did not substantially influence interpretation and therefore the skewness and kurtosis in this data was not of practical significance. However, it may be significant in other datasets and ideally a method designed for use in ordinal data is preferable.

Missingness

The significant relationship between some-item missingness and overall grade (see Appendix C, page 265) had the potential to result in biased estimates of factor coefficients and standard errors because it suggested that the missingness of items was related to their value. This suggestion is consistent with the findings of Mazor, Clauser, Holtman, and Margolis (2007) who examined multisource feedback ratings of medical students and doctors. They found a significant relationship between missing (unable to observe) items and less positive ratings.

Maximum likelihood estimation is one method of addressing missingness, but only results in unbiased estimates if missingness in the dependent variable (in this case item scores) is unrelated to the value of the dependent variable and is either completely at random, or related to covariates that are included in the model (White & Carlin, 2010). When missingness is related to the value of the dependent variable, inclusion of covariates that are also strongly related to the missingness mitigates the bias when maximum likelihood measures are used because it facilitates prediction of the missing variable (Collins, Schafer, & Kam, 2001). Therefore the inclusion of the variables placement, academic status of the supervisor, the elective nature of the placement, and its interaction with academic status into the model used to derive the correlation matrix for factor analysis, as was done in this analysis, would have improved the quality of estimates.

Summary

In summary then, methods were used for this analysis that accounted as much as possible for data missingness, dependency, lack of balance, and expected factor correlation, and that also enabled statistical rather than arbitrary determination of the significance of factors. However, the results found were so clear that in many cases utilisation of other methods would not have changed the conclusions reached. An exception, however, is that if orthogonal rotation had been used, neither a 4-factor solution nor the higher-order structure would have been detected. Orthogonal methods have been most commonly reported in other studies of the internal structure of in-training evaluations as shown in Table 2.2 (page 33). However, based on the findings presented here and the literature reviewed, I recommend that future researchers investigate the internal structure of in-training evaluations using oblique rotations. In addition, reporting sensitivity analyses that model the robustness of conclusions drawn under the different assumptions implicit in factor analysis, is valuable and would enhance the credibility of published research.

Other limitations

Sampling

As discussed, many limitations of the data were managed by adjusting the analysis procedures to best account for them. However, a further limitation relates to the sampling. The research did not include all placements because some were not assessed using the in-training

evaluation and others used a slightly different form with differing items. Represented disciplines did, however span the breadth of species from small animal, equine and large animal, and included placements at the University and externally. Excluded disciplines were the specialised ones of anaesthesia, diagnostic imaging, pathology, public health, and some small animal surgery placements as this last discipline changed its form part way through the study period. Thus, the findings of factor analysis do not apply to all disciplines and further separate factor analyses would be needed to characterise the factor structure in these disciplines.

Method effects

Lastly, the possibility of method effects influencing the factor analysis findings needs to be considered. Method effects are correlations introduced because of similarities in the way data is collected (T. A. Brown, 2006). They therefore do not reflect correlations arising from an underlying common construct and, in the case of this analysis, represent a form of measurement error. Method effects could be present in this data because of the arrangement of items in groups according to the domains being assessed. This may increase the perceived similarity amongst items within the group and suggest a difference from items in other groups, producing a pattern of intercorrelations that are interpreted as factors on factor analysis. Tourangeau (2004) has shown experimentally that an effect of grouping on increasing item intercorrelation in surveys can occur. Podsakoff, MacKenzie, Lee, and Podsakoff (2003), however, point out that mixing items from different dimensions in groups on the survey form may give the opposite method effect of artifactually increasing correlation between different dimensions. Indeed the usual recommendation in survey design is to deliberately group related items because not doing so can make the survey incoherent and confusing (Dillman, Smyth, & Christian, 2014; Krosnick & Presser, 2010). Therefore, one has to consider whether, if method effects due to grouping are indeed present, they represent error or a helpful influence that directs supervisors towards the intentions of the instrument. Method effects may enhance construct-relevance and thereby increase validity rather than undermining it.

In addition, while research on surveys gives some insight into the potential method effects, in-training evaluation forms have one important difference from surveys, in that supervisors are already familiar with the items on the form. Method effects may be far less important in priming responses when the questions are known before completing the form. In in-training evaluation instruments items are usually grouped according to domain, but Silber et al. (2004)

randomised items in a specific attempt to minimise method effects. A two factor internal structure was found in which items loading on the factors were conceptually related but not related by position on the form. This suggests that grouping effects were not the main cause of item correlation. Further research would be necessary to determine the role of method effects in in-training evaluations and whether they improve, or are detrimental to, the validity of scores. Confirmatory factor analysis would be more helpful than exploratory factor analysis is separating method effects from relevant correlations (T. A. Brown, 2006).

Summary

In summary, the results of this phase of research suggest that a four dimensional, high-order structure, consistent with the design of the instrument, was present. However, there was a complexity to the dimensional structure such that solutions with fewer dimensions were also applicable and may be useful for different purposes. The three-dimensional graph was a valuable tool for visualising this complexity. The dimensional structure gives insight into the way supervisors score items, and it is of interest to consider how these dimensions contribute to the overall grade awarded, which was the subject of the next phase of the research.

Chapter 6:
Relationship of factors and overall grade

Findings of the regression analysis—Phase 3

Having established the complex factor structure of the in-training evaluation, in this phase of the research the aim was to examine the relationship between the factors and the overall grade awarded. For this purpose the 4-factor solution was used. This model was chosen in order that the effect of each of the multiple dimensions towards overall grade could be determined. Ordinal logistic regression was performed using generalised linear mixed modelling to examine this relationship and to determine the contribution of other variables such as prior knowledge, placement, academic status of the supervisor, and time. This phase of the research was an important step in determining whether the factors operate as constructs that have influence on scores. Without further analyses such as this, the explanatory nature of factors found on factor analysis remains a hypothesis (Comrey, 1978; Gorsuch, 1983).

Research method

Sample

The data used for analysis was the same as that used for the factor analysis phase to which the additional variables of factor scores were added, as detailed in the next section.

Deriving factor scores

As explained above, the 4-factor model from Phase 2 was used to calculate factor scores for each factor. Factor scoring coefficients were derived using the regression method in SAS PROC FACTOR. These are regression coefficients that reflect the weighting that each item contributes to a factor, taking into account the correlation between items and the correlation between factors (DiStefano, Zhu, & Mindrila, 2009). The standardised factor scores for each

evaluation were then determined using PROC SCORE. This procedure uses the standardised item value and the weighting reflected in the factor coefficient to calculate the standardised factor score using a regression equation.

Missingness analysis and management

The missingness analysis has already been described in relation to the factor analysis and is presented in full in Appendix C (page 257). The proportion of evaluations with missing factor scores ranged from 11.2% for the professional attitude factor to 30.2% for the knowledge factor. However, because these were distributed over evaluations, only 49.9% of evaluations had a complete set of four factor scores. Because this missingness was in the independent variable for the regression model, it could not be managed the same way as it had been for the factor analysis. In PROC GLIMMIX evaluations with missing independent variables are listwise deleted by default (Allison, 2012; SAS Institute Inc., 2014). Thus, the default analysis was complete case analysis and involved deletion of approximately half of the data and the potential for significant bias and loss of power.

Investigation of the possible biasing effects of the missingness revealed a relationship between missing factor scores and the value of the dependent variable (overall grade). Therefore bias might be expected even with the inclusion of other related covariates (Allison, 2000). An alternative analysis with multiple imputation was therefore performed as detailed in Appendix C (page 257). The results of analysis of the multiply imputed data were very similar to the results found with complete case analysis and did not alter any substantive conclusions. The complete case analysis data had the advantage of providing type three tests of fixed effects, and therefore it was reported here.

Ordinal logistic regression method

The relationship between overall grade and factor scores was examined using generalised linear mixed modelling with the SAS procedure GLIMMIX. The code is shown in Box F.1 (Appendix F, page 295). Covariates were chosen for inclusion in the model on the basis of a priori hypotheses or if they appeared to have a relationship with overall grade on examination of scatterplots or boxplots. Some variables could not be included even though they were of

interest. The identity of the supervisor, the site of the placement, and the subdiscipline could not be included because of missing data. The student's GPA at the end of placements could not be included because it was strongly predicted by their GPA at the beginning of placements. The resultant multicollinearity would have caused computational problems during logistic regression (Osborne, 2015; Tabachnick & Fidell, 2013). Covariates modelled were therefore the type of placement (placement), the student's GPA at the start of placements (GPA), and the academic status of the supervisor (academic), that is, whether the supervisor was an academic or from outside the University. GPA was calculated by the University on a scale ranging from 0-9, using the student's weighted average marks from the previous years' assessments. It was then standardised in the form of z scores as is recommended for continuous variables in logistic regression (Osborne, 2015). Factor scores were already in standardised form. Because inclusion of extraneous variables that do not have a relationship with the dependent variable can result in suppressor effects, a significant zero-order relationship for each variable was first verified by running the logistic regression model separately for each independent variable without any covariates (Osborne, 2015).

Laplace approximation was used for estimation of true likelihood which is more accurate than pseudo-likelihood methods with non-normally distributed data (Stroup, 2013). Bias correction was provided by the Morel Bokossa Neerchal empirical estimator as recommended by Stroup (2013) for non-normally distributed data. The denominator degrees of freedom were estimated using the between and within method. In order to determine the best covariance structure for the repeated measurements on each student over time, the main effects model was run with various covariance structures specified. Covariance structures modelled were selected from those that allowed evaluations closer in time to be more highly correlated than those further apart in time. Two heterogeneous covariance structures were modelled using first order antedependence (ANTE(1)) and heterogeneous first order autoregressive (ARH(1)) structures. These both allow the variances to change over time and are recommended for longitudinal studies (Fitzmaurice, Laird, & Ware, 2011) as being more realistic than an assumption of fixed variance. A spatial power and spatial exponential structure were also modelled. While the spatial structures assume fixed variances they have the advantage of allowing unequal spacing of time and accommodating unbalanced data (Fitzmaurice et al., 2011; Stroup, 2013), both of which were present in this data. The covariance structure chosen for use in the analysis was that which resulted in the lowest values of the Akaike information

criterion (AIC), the Bayesian information criterion (BIC), and the corrected Akaike information criterion (AICC), indicating the closest fit.

Model selection was performed by entering terms into the model in blocks in order to first establish whether curvilinear effects were present and then to determine if there were any significant interactions (Osborne, 2015). Firstly, a model containing the main effects and also quadratic terms of the continuous variables was examined, and quadratic terms retained in the model if statistically significant (p<0.05). If significant quadratic terms were found then the cubic term was also modelled and retained if significant. Next, all two way interactions between factor scores and covariates were added to the model. Non-significant interactions were then removed from the model. Lastly, interactions between any retained quadratic and cubic terms and other variables were added to the model, and then removed if they were non-significant or did not improve model fit, as determined by the information criteria. In interpreting results a p value of <0.05 was considered significant. To illustrate the effect size, predicted probabilities were calculated for certain values of factor scores and covariates using the regression equations.

Findings

Distribution of overall grade

Overall grades awarded ranged from fail (score of 0) to excellent (score of 10), with a mean score of 8.32 and a median of 8. The distribution of overall grades awarded is shown in Table 6.1, and the distribution of overall grades according to placement, academic status of the supervisor, and year is shown in Table F.1 (Appendix F, page 296).

Table 6.1: Distribution of overall grades across all evaluations.

Overall grade (score)	Frequency	Percent
Blank	149	4.3
Excellent (10)	1095	31.6
Good (8)	1747	50.4
Satisfactory (6)	421	12.2
Marginal (4)	42	1.2
Fail (0)	12	0.4

Note. Percentages are rounded to the nearest decimal place

Results of model selection procedures

Zero-order testing confirmed that each standardised factor score, as well the variables placement, GPA, and academic, were all significantly related to overall grade when present separately in a model (p<0.001) and thus each was suitable for inclusion in the model. There was no difference in model fit between covariance structures and therefore the spatial power covariance structure was used for further analysis. No significant quadratic curvilinear relationships between factor scores or GPA and overall grade were found. Significant interactions between Factor 3 (professional attitude) and placement (p=0.014) and Factor 2 (knowledge) and placement (p=0.009) were found. Interactions between Factor 1 (clinical skills) and placement and Factor 4 (communication) and placement were not significant when present in the model with the other factors and the removal of these interaction terms from the model improved model fit. However, when other factors were omitted from the model, the interactions of Factor 1 (clinical skills) with placement and Factor 4 (communication) with placement were significant (p<0.0002). The variables and interactions in the final model are shown in Table 6.2 and the parameter estimates are shown in Table F.3 (Appendix F, page 297). The main model, without interaction terms (Table F.2 in Appendix F, page 297) was used to calculate probabilities.

Regression findings

The results indicated that the overall grade awarded was significantly related to each factor and that for Factor 3 (professional attitude) and Factor 2 (knowledge) the relationship differed between placements (Table 6.2). Of the four factors, Factor 3 (professional attitude) had the

strongest effect, with the highest beta coefficient of 2.71 (Table F.3 in Appendix F, page 297). This indicates that every 1 standard deviation increase in the professional attitude score predicts an increase in overall grade of 2.71 standard deviations. Factor 1 (clinical skills) had the next strongest effect followed by Factor 4 (communication). Factor 2 (knowledge) had the least effect. The strong relationship between each factor and overall grade can be visualised by the distribution of the factor scores for each overall grade shown in Figure 6.1.

There was substantial variation in the number of excellent, good, and satisfactory grades awarded on each placement (Figure 6.2). The proportion of excellent grades awarded within a placement varied from 0% (PA6) to 56% (EXT). The proportion of good grades awarded varied between 0% (PA6) and 88% (SA2). The proportion of satisfactory grades awarded varied between 4% (EXT) and 99% (PA6).

There was no evidence that overall grade was related to GPA (p=0.489). Figure 6.3 shows the distribution of GPA for each overall grade awarded. While there is a small non-significant trend for increasing overall grade as student GPA increases, there is also a substantial range of GPAs seen for each overall grade category, indicating that students with a wide range of GPA were awarded satisfactory, good, and excellent grades.

There was no evidence that overall grade was related to the academic status of the supervisor (p=0.945). Figure 6.4 shows the distribution of overall grades between placements staffed by academic and non-academic supervisors. Academic supervisors awarded fewer excellent grades (21.6%) than non-academic supervisors (41.9%) and more good (63.0% vs 50.4%) and satisfactory (13.0% vs 6.9%) grades than non-academic supervisors but these differences were not statistically significant.

There was no significant relationship between evaluations for a student over time (p=1).

Table 6.2: Fixed effects of the independent variables on overall grade.

variable	degrees of freedom		F Value	*p* value
	numerator	denominator		
Factor1 (clinical skills)	1	1372	20.18	<.0001
Factor2 (knowledge)	1	1372	0.09	0.764
Factor3 (professional attitude)	1	1372	12.02	0.0005
Factor4 (communication)	1	1372	5.82	0.016
academic	1	196	0	0.945
placement	10	819	2.11	0.021
GPA	1	192	0.48	0.489
Factor2-placement interaction	10	1372	2.35	0.009
Factor3-placement interaction	10	1372	2.25	0.014

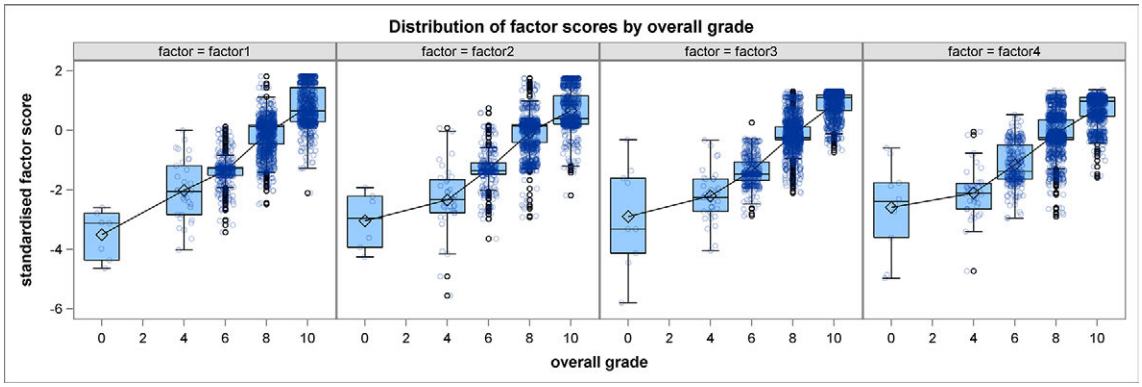


Figure 6.1: Distribution of standardised factor scores for each overall grade. Scatter plot and boxplots overlaid. Data points on the scatter plot have been moved slightly horizontally so that they do not lie exactly on top of one another. Each box represents the bounds of the upper and lower quartile and the whiskers indicate the range. The box is transected at the median and the diamond indicates the mean value. Key: factor1: clinical skills factor; factor2: knowledge factor; factor3: professional attitude factor; factor4: communication factor. Overall grade scores: 10=excellent, 8=good, 6=satisfactory, 4=marginal, 0=fail.

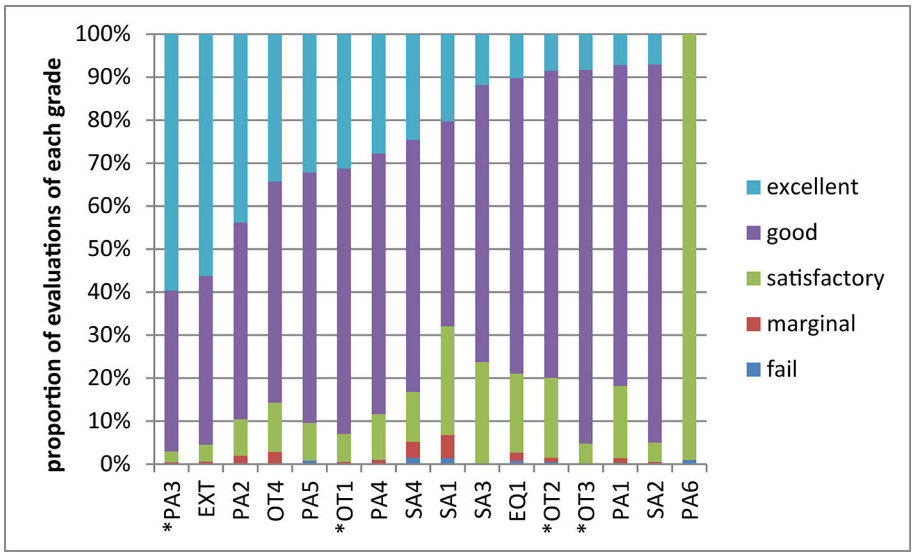


Figure 6.2: Distribution of the overall grades within each placement showing the variation between placements in the proportion of each level of overall grade awarded. Note. *indicates placements that were not included in the regression analysis because they were not evaluated using the common in-training evaluation form (OT1, OT2, OT3) or because of missingness in factor scores (PA3).

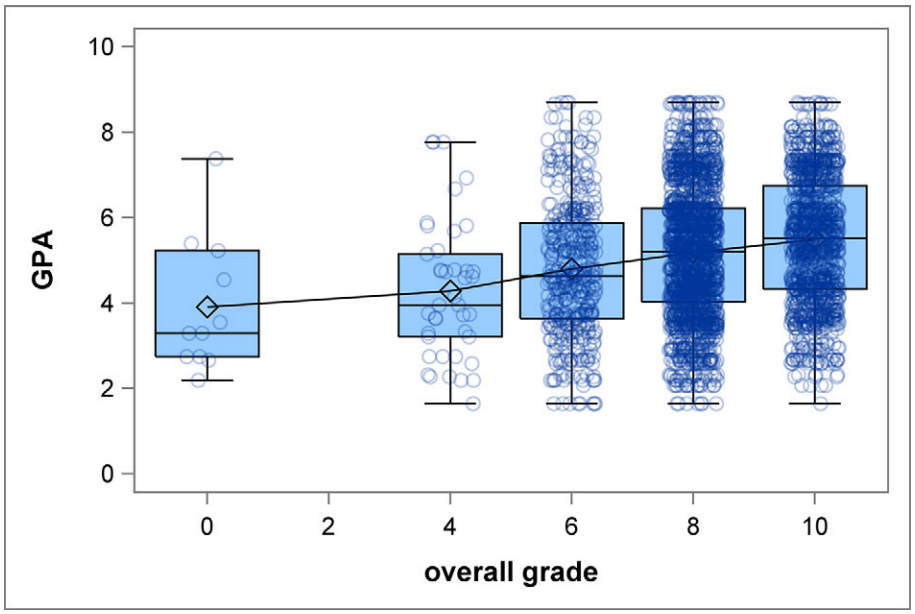


Figure 6.3: Distribution of GPA for each overall grade. Scatter plot and boxplots overlaid. Data points on the scatter plot have been moved slightly horizontally so that they do not lie exactly on top of one another. Each box represents the bounds of the upper and lower quartile and the whiskers indicate the range. The box is transected at the median and the diamond indicates the mean value. Overall grade scores: 10=excellent, 8=good, 6=satisfactory, 4=marginal, 0=fail. GPA was measured on a scale ranging from 0-9.

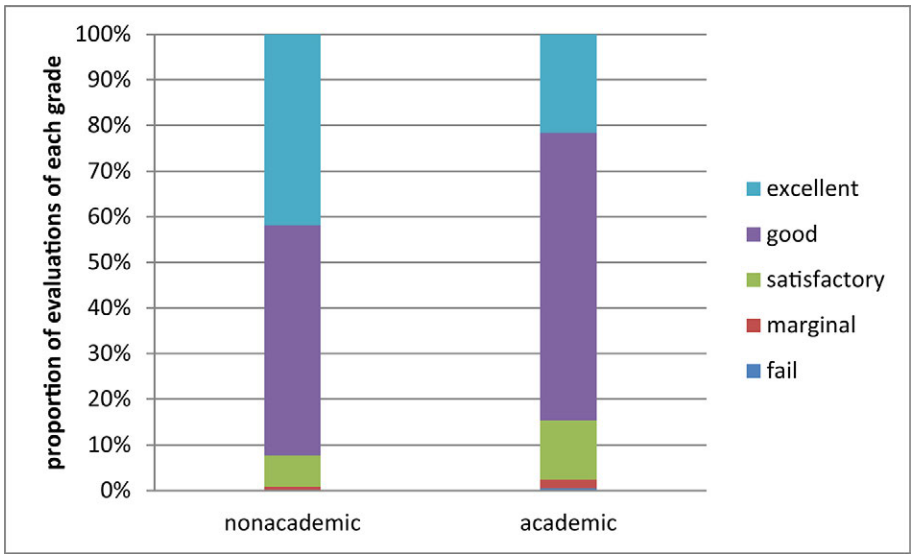


Figure 6.4: Distribution of the overall grades within placements based at the University (academic) and placements based in external veterinary practices (non-academic).

Illustrating the findings

The degree of influence of factors on overall grade can be illustrated with probabilities which gives a measure of effect size in addition to the beta coefficients. Table 6.3 and Table 6.4 show how changes in one factor score affect the probability of the overall grade changing, when other variables are held fixed. The tables use four placements to show the spectrum of results. Figure 6.5 illustrates the probabilities shown in the tables.

Several aspects of the results are illustrated in the tables and graphs. The first is that the effect of Factor 3 (professional attitude) predominated over the effect of the other factors in both raising and lowering overall grade. For example, in the PA1 placement, if the score for the professional attitude factor was typical of students who received an overall grade of good, but the other factor scores were typical of students who received overall grades of satisfactory, then the probability of the overall grade being at least good was 0.57. However, lower probabilities were seen when it was other factors that were good. For example if Factor 4 (communication) was good and the other three factors were satisfactory, the probability of an overall grade of at least good was much lower at 0.17. This illustrates the much greater effect of the professional attitude factor in raising overall grade compared to the communication factor.

A similar situation was seen when the opposite effect, the influence of a factor in lowering overall grade, was examined. On the PA1 placement, if Factor 3 (professional attitude) was satisfactory and the rest of the factors were good, the probability of receiving a grade of satisfactory or lower was 0.27. However if Factor 4 (communication) was satisfactory and the rest were good, the probability of receiving a grade of satisfactory or lower was much lower at 0.06.

Table 6.3: Effect of individual factor scores to raise overall grade for four representative placements – changes in probabilities (*p*).

overall grade that *p* is of	pattern of factor scores	*p*	pattern of factor scores	*p* when this factor has higher score			
				F1	F2	F3	F4
PA1 placement							
excellent	4 good	0.01	1 excellent, 3 good	0.04	0.03	0.08	0.02
good	4 satisfactory	0.09	1 good, 3 satisfactory	0.33	0.23	0.57	0.17
satisfactory	4 marginal	0.36	1 satisfactory, 3 marginal	0.64	0.55	0.79	0.48
marginal	4 fail	0.40	1 marginal, 3 fail	0.67	0.58	0.80	0.52
EXT placement							
excellent	4 good	0.14	1 excellent, 3 good	0.34	0.26	0.52	0.21
good	4 satisfactory	0.57	1 good, 3 satisfactory	0.86	0.79	0.94	0.72
satisfactory	4 marginal	0.88	1 satisfactory, 3 marginal	0.96	0.94	0.98	0.92
marginal	4 fail	0.90	1 marginal, 3 fail	0.96	0.95	0.98	0.93
PA5 placement							
excellent	4 good	0.41	1 excellent, 3 good	0.68	0.59	0.82	0.53
good	4 satisfactory	0.85	1 good, 3 satisfactory	0.96	0.94	0.99	0.91
satisfactory	4 marginal	0.97	1 satisfactory, 3 marginal	0.99	0.98	0.99	0.98
marginal	4 fail	0.97	1 marginal, 3 fail	0.99	0.99	1.00	0.98
SA1 placement							
excellent	4 good	0.03	1 excellent, 3 good	0.09	0.06	0.17	0.05
good	4 satisfactory	0.20	1 good, 3 satisfactory	0.54	0.41	0.76	0.33
satisfactory	4 marginal	0.58	1 satisfactory, 3 marginal	0.81	0.75	0.90	0.69
marginal	4 fail	0.62	1 marginal, 3 fail	0.83	0.77	0.91	0.72

Note. The table compares the probability of obtaining various overall grades or higher with all 4 factor scores at a particular level or 3 at that level and 1 higher.

For example, the first row shows that the probability of an excellent overall grade on the PA1 placement is 0.01 if all 4 factor scores are good, and 0.04 if F1 is excellent, but all the other three factors are good, and 0.03 if F2 is excellent and all of the other three factors are good, and so on.

The levels modelled are factor scores at the mean of the factor score for all students receiving that overall grade. GPA is held constant at its mean for all students.

Key: F1: clinical skills factor; F2: knowledge factor; F3: professional attitude factor; F4: communication factor.

Table 6.4: Effect of individual factor scores to lower overall grade for four representative placements – changes in probabilities (*p*).

overall grade that *p* is of	pattern of factor scores	*p*	pattern of factor scores	*p* when this factor has lower score			
				F1	F2	F3	F4
PA1 placement							
good	4 excellent	0.51	3 excellent, 1 good	0.77	0.69	0.87	0.63
satisfactory	4 good	0.03	3 good, 1 satisfactory	0.12	0.08	0.27	0.06
marginal	4 satisfactory	0.02	3 satisfactory, 1 marginal	0.07	0.05	0.14	0.04
fail	4 marginal	0.02	3 marginal, 1 fail	0.07	0.05	0.13	0.04
EXT placement							
good	4 excellent	0.08	3 excellent, 1 good	0.21	0.15	0.35	0.12
satisfactory	4 good	0.00	3 good, 1 satisfactory	0.01	0.01	0.03	0.00
marginal	4 satisfactory	0.00	3 satisfactory, 1 marginal	0.01	0.00	0.01	0.00
fail	4 marginal	0.00	3 marginal, 1 fail	0.01	0.00	0.01	0.00
PA5 placement							
good	4 excellent	0.02	3 excellent, 1 good	0.06	0.04	0.12	0.03
satisfactory	4 good	0.00	3 good, 1 satisfactory	0.00	0.00	0.01	0.00
marginal	4 satisfactory	0.00	3 satisfactory, 1 marginal	0.00	0.00	0.00	0.00
fail	4 marginal	0.00	3 marginal, 1 fail	0.00	0.00	0.00	0.00
SA1 placement							
good	4 excellent	0.31	3 excellent, 1 good	0.58	0.48	0.74	0.42
satisfactory	4 good	0.01	3 good, 1 satisfactory	0.05	0.03	0.14	0.02
marginal	4 satisfactory	0.01	3 satisfactory, 1 marginal	0.03	0.02	0.06	0.02
fail	4 marginal	0.01	3 marginal, 1 fail	0.03	0.02	0.06	0.02

Note. The table compares the probability of obtaining various overall grades or lower with all 4 factor scores at a particular level or 3 at that level and 1 lower.

For example, the first row shows that the probability of a good overall grade on the PA1 placement is 0.51 if all 4 factor scores are excellent, and 0.77 if F1 is good, but all the other three factors are excellent, and 0.69 if F2 is good and all of the other three factors are excellent, and so on.

The levels modelled are factor scores at the mean of the factor score for all students receiving that overall grade. GPA is held constant at its mean for all students.

Key: F1: clinical skills factor; F2: knowledge factor; F3: professional attitude factor; F4: communication factor.

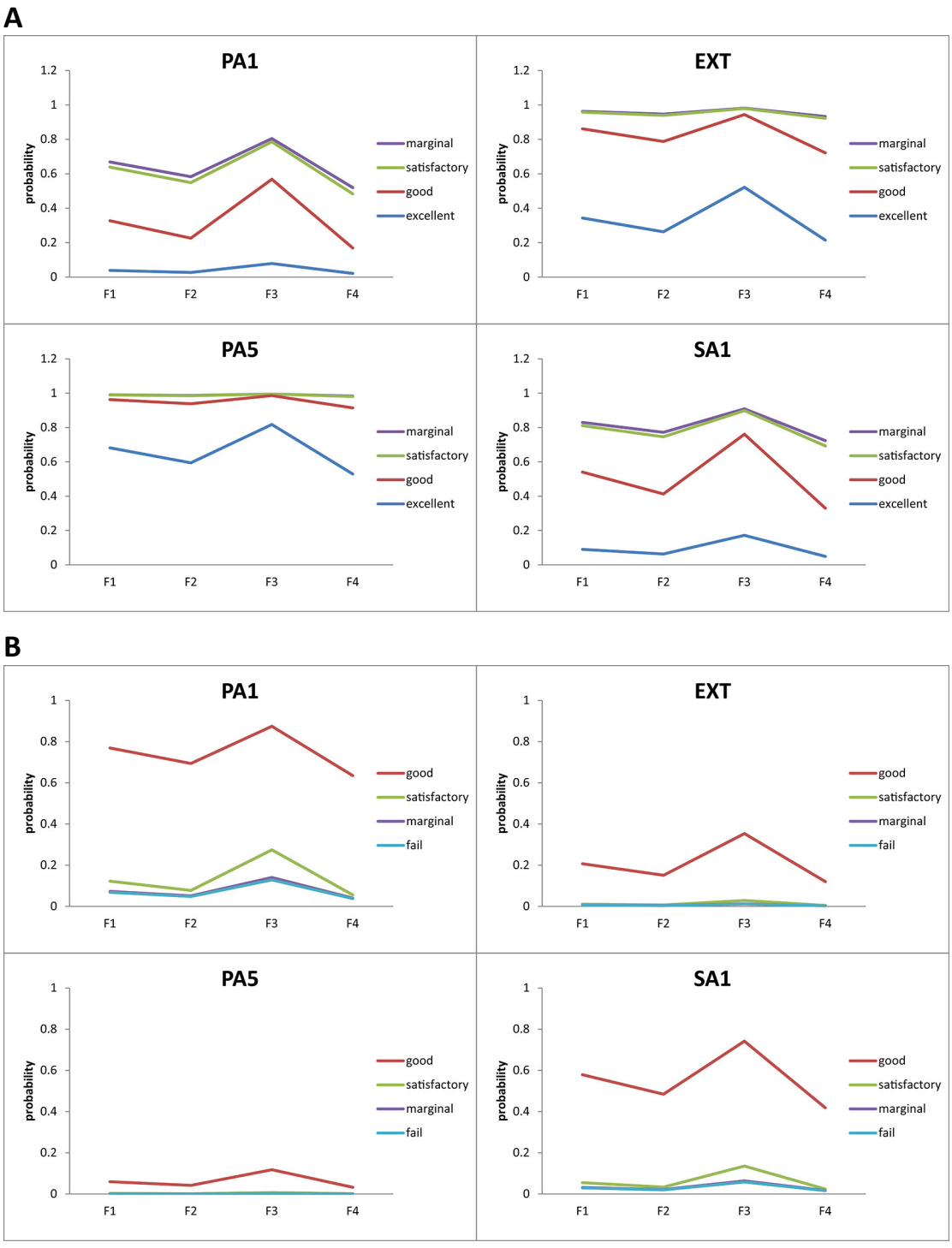


Figure 6.5: The effect of each factor score on raising (A) or lowering (B) overall grade for four representative placements. A shows the probability of a certain overall grade when one factor score is at a typical level and all three other factor scores are a level lower. **B** shows the probability of a certain overall grade when one factor score is at a typical level and all three other factor scores are a level higher. The levels modelled are factor scores at the mean of the factor score for all students receiving that overall grade. GPA is held constant at its mean for all students. Key: F1: clinical skills factor; F2: knowledge factor; F3: professional attitude factor; F4: communication factor.

The second aspect illustrated by the tables and graphs is that the degree of effect of changes in factor scores on the probability of obtaining a particular overall grade differed for high and low overall grades. This is best illustrated by the relative risks shown in Table 6.5 and Table 6.6. The influence of a better performance in one factor to raise overall grade was greater when the overall level of performance was high, than when overall level of performance was low. A different pattern was seen in the influence of a lower performance in one factor to lower overall grade, where it was greatest at middle levels of overall performance and lowest at very high levels of overall performance.

The third aspect illustrated by the tables and graphs is that the effect of factor scores on overall grade was different for different placements. Although across the different placements there was a consistently stronger relationship between Factor 3 (professional attitude) on overall grade than other factors, there was considerable difference between placements in the magnitude of this effect. To some extent this is related to the significant relationship between placement and overall grade. Thus, because there were marked differences between placements in the probability of certain grades, differences in the effect of particular factor scores to influence these were not unexpected.

Table 6.5: Effect of individual factor scores to raise overall grade for four representative placements – changes in relative risk (RR).

overall grade that RR is of	pattern of factor scores	RR when this factor has higher score			
		F1	F2	F3	F4
PA1 placement					
excellent	1 excellent, 3 good	3.05	2.11	6.11	1.63
good	1 good, 3 satisfactory	3.46	2.38	6.00	1.79
satisfactory	1 satisfactory, 3 marginal	1.75	1.51	2.16	1.33
marginal	1 marginal, 3 fail	1.66	1.45	2.00	1.29
EXT placement					
excellent	1 excellent, 3 good	2.40	1.84	3.65	1.50
good	1 good, 3 satisfactory	1.51	1.38	1.65	1.26
satisfactory	1 satisfactory, 3 marginal	1.09	1.07	1.11	1.05
marginal	1 marginal, 3 fail	1.07	1.06	1.10	1.04
PA5 placement					
excellent	1 excellent, 3 good	1.68	1.46	2.01	1.30
good	1 good, 3 satisfactory	1.14	1.11	1.17	1.08
satisfactory	1 satisfactory, 3 marginal	1.02	1.02	1.03	1.01
marginal	1 marginal, 3 fail	1.02	1.01	1.02	1.01
SA1 placement					
excellent	1 excellent, 3 good	2.94	2.07	5.60	1.61
good	1 good, 3 satisfactory	2.68	2.05	3.77	1.64
satisfactory	1 satisfactory, 3 marginal	1.40	1.28	1.55	1.19
marginal	1 marginal, 3 fail	1.34	1.25	1.47	1.17

Note. The table compares the relative risk of obtaining various overall grades or higher with 3 factor scores at a particular level and 1 higher.

For example, the first row shows that the relative risk of an excellent overall grade on the PA1 placement is 3.05 if F1 is excellent, but all the other three factors are good, and 2.11 if F2 is excellent and all of the other three factors are good, and so on.

The relative risk represents the ratio of the probability of the overall grade or higher with 3 factor scores at a particular level and 1 factor score higher, over the probability of all factor scores at the particular level. The levels modelled are factor scores at the mean of the factor score for all students receiving that overall grade. GPA is held constant at its mean for all students.

Key: F1: clinical skills factor; F2: knowledge factor; F3: professional attitude factor; F4: communication factor.

Table 6.6: Effect of individual factor scores to lower overall grade for four representative placements – changes in relative risk (RR).

overall grade that RR is of	pattern of factor scores	RR when this factor has lower score			
		F1	F2	F3	F4
PA1 placement					
good	3 excellent, 1 good	1.49	1.35	1.70	1.23
satisfactory	3 good, 1 satisfactory	4.21	2.65	9.41	1.90
marginal	3 satisfactory, 1 marginal	2.94	2.06	5.65	1.61
fail	3 marginal, 1 fail	2.86	2.03	5.44	1.59
EXT placement					
good	3 excellent, 1 good	2.69	1.97	4.59	1.56
satisfactory	3 good, 1 satisfactory	4.62	2.77	12.25	1.95
marginal	3 satisfactory, 1 marginal	3.08	2.12	6.34	1.63
fail	3 marginal, 1 fail	2.99	2.07	6.04	1.61
PA5 placement					
good	3 excellent, 1 good	3.00	2.09	5.90	1.62
satisfactory	3 good, 1 satisfactory	4.65	2.78	12.50	1.95
marginal	3 satisfactory, 1 marginal	3.09	2.12	6.39	1.63
fail	3 marginal, 1 fail	3.00	2.08	6.09	1.61
SA1 placement					
good	3 excellent, 1 good	1.90	1.59	2.43	1.37
satisfactory	3 good, 1 satisfactory	4.46	2.73	11.02	1.93
marginal	3 satisfactory, 1 marginal	3.02	2.10	6.07	1.62
fail	3 marginal, 1 fail	2.94	2.06	5.81	1.60

Note. The table compares the relative risk of obtaining various overall grades or lower with 3 factor scores at a particular level and 1 lower.

For example, the first row shows that the relative risk of a good overall grade on the PA1 placement is 1.49 if F1 is good, but all the other three factors are excellent, and 1.35 if F2 is good and all of the other three factors are excellent, and so on.

The relative risk represents the ratio of the probability of the overall grade or lower with 3 factor scores at a particular level and 1 factor score lower, over the probability of all factor scores at the particular level. The levels modelled are factor scores at the mean of the factor score for all students receiving that overall grade. GPA is held constant at its mean for all students.

Key: F1: clinical skills factor; F2: knowledge factor; F3: professional attitude factor; F4: communication factor.

Strengths and limitations of Phase 3

The aim of this phase of the research was to examine the relationship between the factors and the overall grade awarded, and determine the influence of other covariates. The factor analysis had established the factor structure of the in-training evaluation and given some indication of the nature of each factor and therefore the constructs underlying the items in the in-training evaluation. The ordinal logistic regression analysis was an important step in determining whether the overall grade, which represents the overall evaluation, captured the constructs underlying the items. The results indicated that the factors each were significantly related to the overall grade with this relationship moderated by the performance level (low scoring or high scoring) and the placement. The effect was strongest for Factor 3 (professional attitude) followed by (in decreasing order of effect) Factor 1 (clinical skills), Factor 4 (communication), and Factor 2 (knowledge). There was no evidence that overall grade was related to GPA, academic status of the supervisor, or the time of year of the evaluation. The implications of these findings will be discussed further in relation to the results of the other phases of research in the general discussion chapter. Here some limitations that need to be considered in interpreting the results are discussed.

A strength of this analysis was the use of generalised linear mixed modelling for ordinal logistic regression. This was an appropriate method to meet the aims of the research for a number of reasons. Regression methods are useful in being able to determine the relative influence of variables and to examine for interactions, while controlling type I error rates (Osborne, 2015). Although ordinal data could be converted to numerical scores and analysed with linear regression, this would have had a number of disadvantages. One is that the categories in ordinal data represent a range rather than a fixed point and converting categories to fixed points introduces measurement error. Also, analysing ordinal data with linear regression does not account for the fixed upper and lower limits of the ordinal scale, which contrasts with continuous scales. These issues, amongst others, can lead to misleading results if linear regression is conducted on ordinal data (Agresti, 2010). The generalised linear mixed model enables the ordinal nature of the data to be modelled at the same time as accounting for the dependency produced by repeated evaluations of each student over time, such as was present in the data for this phase of study.

The use of factor scores derived from common factor analysis is sometimes criticised because they are indeterminate (Steiger, 1979), meaning that different factor scores are possible (Gorsuch, 1983). The regression method used by SAS is a recommended method for calculating factor scores (Gorsuch, 1983) and accounts for the intercorrelations between variables, factors, and between variables and factors (DiStefano et al., 2009). The sufficiency of factor scores in representing the data was checked by evaluating the factor pattern obtained from performing common factor analysis on factor scores (analysis not shown) as suggested by Beauducel (2005). The same factor pattern was able to be reproduced, indicating that the factor scores were closely reproducing the original correlation matrix, and would lead to the same interpretation.

As discussed in detail in Appendix C (page 257), missingness in factor scores was a significant problem in this analysis as it led to removal of approximately half the data. Although the remaining sample was still large, the reduction in power may have affected the ability to detect relationships (Type II error), especially those involving interactions which may require greater power to detect (Marshall, 2007). However, the sample size was sufficient to detect several relationships including interactions. The comparative analysis with multiple imputation (Appendix C, page 271) suggested that any bias of the sample because of the missingness was not sufficient to alter the substantive conclusions of the analysis. Consistency of the results of this phase of research with other findings, as described in Phase 4, provides further support for the robustness of this analysis even in the presence of such a high degree of missingness.

A further limitation of the analysis was that one placement (EXT) was not specific to a discipline, but indicated work of any type of the student's choosing. It was not possible to determine more detail about the nature of the work on that placement, however it was known that this placement was always held at a site other than the University. The expected effect of this limitation would be to dilute the effect of placement as a variable in the model, because of the overlap of the EXT placement with other placements. Therefore, it may have increased the probability of type II errors, that is, of not detecting relationships that were present. However, since relationships were nevertheless detected this limitation did not appear to have a major effect.

Summary

In summary, the results of this analysis indicate that all factors were related to overall grade, with the professional attitude factor having the strongest effect, followed by the clinical skills factor, communication factor, and then the knowledge factor. The effects of the professional attitude factor and the knowledge factor on overall grade depended on the placement considered. There was substantial variation in the proportion of excellent overall grades awarded between placements, and a significant effect of placement on overall grade. There was no significant effect of academic status of the supervisor, or GPA on the overall grade. There was no significant trend for overall grades to differ over time.

The effect of the professional attitude factor on overall grade predominated over other factors in both raising and lowering overall grade. The effect was substantial and could lead to changes in probabilities of 20% in certain conditions. The degree of effect of changes in factor score on overall grade depended on the level of performance of the student (high performing vs low performing).

The three phases of research presented to date have shown what aspects of student performance supervisors value, the dimensions captured by the in-training evaluation, and the effect of those dimensions and other variables in influencing overall grade. In the next phase of research these findings will be integrated with a thematic analysis of the content of the mark scheme for the in-training evaluation and the competency frameworks for veterinary students in New Zealand. This will suggest whether the aspects of student performance that are of importance to supervisors are successfully captured in scores on the in-training evaluation and reflect what the in-training evaluation is intended to assess.

Chapter 7:
Alignment of practice with intentions

Thematic comparison and integration—Phase 4

Phase 1 of the research described what supervisors consider important in the performance of students in their placements. This gave us insight into supervisors' expectations of student performance and the picture of performance they develop. Phase 2 showed the complex higher-order dimensional structure of the in-training evaluation. This suggested that supervisors consider the performance holistically, as well as discriminating the four domains of competency. Phase 3 showed that the dimensions (factors) operate as constructs in explaining the overall grade awarded, and are influenced by contextual and/or rater factors related to the placement. This suggested that in translating their evaluation to a score, the complex higher-order dimensional structure of their evaluation is preserved, and a supervisor's judgement is both holistic and discriminating. The question remains, however, whether what they are discriminating about is what they also think is important. Supervisors are asked to evaluate certain things, but may think other things are important (Ginsburg et al., 2010; Rosenbluth et al., 2014). Investigating the alignment between what they think is important, what dimensions they do assess, what they are asked to assess, and the veterinary competency frameworks were therefore the subject of this fourth phase of the research. In short, this phase examined the alignment of practice with the intentions for the assessment.

Research method

Three documents were analysed in this part of the research. One was the programme level learning outcomes for the veterinary degree, abbreviated here as the BVSc learning outcomes (Massey University, 2012). The second was the Veterinary Council of New Zealand's Competency Standards and Performance Indicators for Veterinarians, abbreviated here as the VCNZ standards (Veterinary Council of New Zealand, n.d.). These are the statements of competency for a practising veterinarian in New Zealand. Both documents are a type of competency framework which details the expected attributes and skills of veterinarians that students at Massey University are working towards on placements, as specified by the

University and by the profession. The third was the mark scheme for the in-training evaluation. Each item on the in-training evaluation contained an overall descriptor and then level descriptors for each level of performance. Together, these formed the mark scheme for the instrument and represent the specified criteria for evaluation.

All documents were imported into NVivo for analysis. The thematic framework that emerged from the interviews with supervisors in the first phase of research was applied to the documents. Themes in the documents were compared to the coding definitions developed in Phase 1 and the words and phrases in the documents assigned codes from the original framework. Dual coding was permitted, as it had been in the original analysis. Although a deductive approach was taken at this stage (applying pre-defined codes to the data) I remained vigilant for ideas that did not fit with the coding framework. None were found other than headings used to structure the documents and instructions for users. These were not coded.

In order to compare the theme use in the learning outcomes and competency standards with those used in the interviews, the number of words devoted to each theme was quantified in NVivo. Although a limitation of this is that the number of words does not necessarily indicate the importance of a theme, two other possible methods of quantification were not suitable for this part of the analysis. Number of coding references could not be used because they were artificially inflated in the interview data as coding references were frequently interspersed with uncoded elements where the interviewer interrupted with encouragement to go on, or requests to clarify or elaborate. The presence or absence of themes could not be used, as it had been in the interview analysis, because there was only one of each document, so a relative count compared to the number of supervisors would be meaningless. This limitation was considered in drawing conclusions from the findings.

Word counts for the interviews, BVSc learning outcomes, VCNZ standards, and mark scheme were assembled in a spreadsheet by theme, by item of the in-training evaluation, and by dimension (factor). These were tabulated and graphed for comparison and analysis.

Findings

All content of the documents except the headings and instructions for document use, was able to be coded to one or more themes using the thematic framework generated in the analysis of interviews. No additional codes were generated during the coding process.

Themes in the mark scheme

Overall

The proportion of words in each theme was similar for the mark scheme and the interview data (Figure 7.1). In particular, engagement, knowledge, and social interactions had very similar proportions of words. The mark scheme had more focus on trustworthiness, technical and animal skills, and caring for animals than the interview data. The interview data had more focus on applied knowledge, communication, personal functioning (none) than the mark scheme. Personal functioning was notably absent from the mark scheme, but was a theme mentioned by supervisors in the interviews. Based on the proportion of words used, the most prominent theme in the mark scheme was engagement and the second most prominent was trustworthiness.

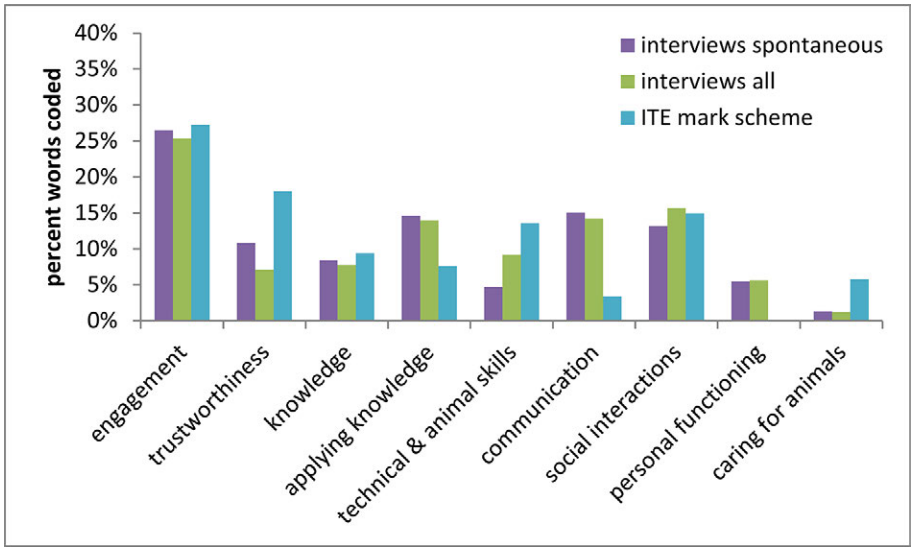


Figure 7.1: Proportion of words devoted to each theme in the in-training evaluation (ITE) mark scheme, the spontaneous descriptions by supervisors in the interviews and all (both spontaneous and clarified and prompted) descriptions in the interviews.

By dimension (factor)

The descriptors for items on the in-training evaluation were not purely composed of single themes from the thematic framework, but contained a mixture of themes (Figure 7.2). The dimensions (factors) of the in-training evaluation were therefore also comprised of a mixture of themes but each formed a unique combination (Figure 7.3). Each factor differed in what was most prominent, based on proportion of words in the mark scheme for the items in each factor. The most prominent themes in the clinical skills factor were technical and animal skills and engagement, with caring for animals, application of knowledge, and communication also important elements. The most prominent themes in the knowledge factor were application of knowledge, engagement, and knowledge. The most prominent themes in the professional attitude factor were engagement and trustworthiness. The most prominent themes in the communication factor were social interactions, trustworthiness, and communication.

Most themes from the thematic framework were not exclusive to any one factor (dimension), but occurred in more than one factor, showing that factors overlapped in thematic content. Only the knowledge theme and the technical and animal theme were confined to one factor. Overlap of the knowledge factor with the professional attitude factor was entirely due to the self-directed learning factor being aligned with the theme of engagement, which was a large component of the professional attitude factor.

The most important themes in the pictures of student performance derived from the interviews (engagement and trustworthiness) were the themes that made up most of the mark scheme for the professional attitude factor.

The themes in each factor were conceptually consistent with the aspect of performance each dimension was intended to assess, as indicated by the mark scheme. For example, the clinical skills dimension mapped to a mixture of themes of applying knowledge, technical and animal handling skills, communication, caring for animals as well as engagement and trustworthiness. This is conceptually consistent with the type of work involved in running a veterinary consultation with a client and animal. Table 7.1 illustrates how the themes within each dimension correspond to the item descriptors on the mark scheme for each dimension.

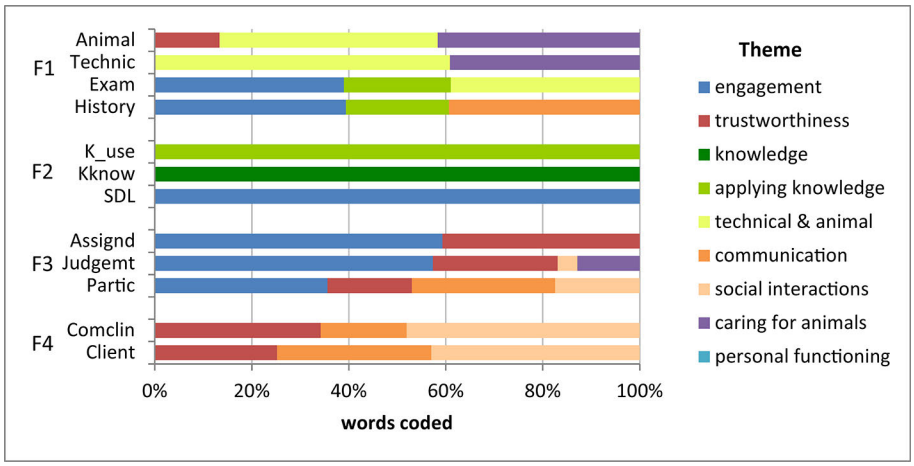


Figure 7.2: Proportion of words in the mark scheme represented by each theme, for each item in the in-training evaluation. In-training evaluation items are indicated on the vertical axis and arranged in groups that correspond to the four dimensions found in the factor analysis. Themes were those derived from the interviews with supervisors. Key: F1: clinical skills factor; F2: knowledge factor; F3: professional attitude factor; F4: communication factor.

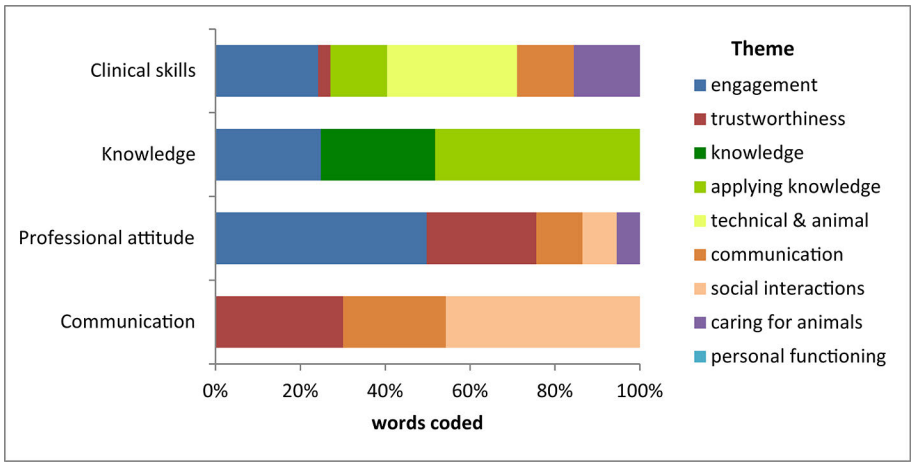


Figure 7.3: Proportion of words in the mark scheme represented by each theme, for each factor in the in-training evaluation. Dimensions (factors) are indicated on the vertical axis. Themes were those derived from the interviews with supervisors.

Table 7.1: Side by side comparison of the item descriptors for the in-training evaluation and the themes from the interviews represented in each dimension. This table continues on the next page.

Dimensions and items on the in-training evaluation			Themes from the interviews with supervisors	
Dimension	**Item**	**Description**	**Themes**	**Theme description**
Clinical skills	Animal	Animal handling and patient care, including empathy and manual skill.	engagement	Enthusiasm for and participation in the activities of the placement.
			technical & animal	Discipline-specific technical and animal handling skills including safe handling of animals
	Technic	Technical skills, proficiency, including care of patient	applying knowledge	Problem solving, planning, decision making, report writing, giving presentations
			communication	Character and effectiveness of communication and willingness to communicate, listening, body language
	Exam	Physical examination skills including accuracy, thoroughness, organisation	caring for animals	Showing concern for animals under their care, paying attention to the animals, concern for animal welfare
			trustworthiness	Honesty, reliability, taking responsibility, and ability to discern own limits and act within them, including perceptions about own abilities
	History	History taking from clients, including thoroughness, organisation and completeness		
Knowledge	K_use	Use of knowledge including ability to integrate findings and theory, determine differential diagnoses and make plans	applying knowledge	Problem solving, planning, decision making, report writing, giving presentations
	Kknow	Knowledge of the discipline	knowledge	Discipline-specific knowledge
	SDL	Self-directed learning including use of resources and effort in sourcing information.	engagement	Enthusiasm for and participation in the activities of the placement.

Note. This table presents dimensions found on factor analysis side by side with the corresponding domain descriptions from the mark scheme for the in-training evaluation for ease of comparison. Comparison should be made by domain only. Individual rows within domains are not specifically aligned.

Table 7.1: Continued.

Dimensions and items on the in-training evaluation			Themes from the interviews with supervisors	
Dimension	**Item**	**Description**	**Themes**	**Theme description**
Professional attitude	Assignd	Performance of assigned tasks including completeness, effort, and initiative in going beyond the bare minimum required but not beyond own limitations.	engagement	Enthusiasm for and participation in the activities of the placement.
	Judgemt	Professional attitude and judgement, including maturity of demeanour, responsibility, reliability, dependability, interest, motivation, knowledge of own limitations, punctuality, personal presentation, care for animal welfare.	trustworthiness	Honesty, reliability, taking responsibility, and ability to discern own limits and act within them, including perceptions about own abilities
			communication	Character and effectiveness of communication and willingness to communicate, listening, body language
			social interaction	Social awareness, relationships with others, ability to work with others
	Partic	Participation including team work, interest, dependability, availability, contribution to discussion.	caring for animals	Showing concern for animals under their care, paying attention to the animals, concern for animal welfare
Communica-tion	Comclin	Communication and interactions with clinical staff, including respectfulness, reliability, responsibility, teamwork, sensitivity to other's concerns, maturity, tidiness and professionalism.	social interactions	Social awareness, relationships with others, ability to work with others
			trustworthiness	Honesty, reliability, taking responsibility, and ability to discern own limits and act within them, including perceptions about own abilities
	Client	Communication and rapport with clients, including respect, empathy, integrity and ability to convey medical information and establish trust.	communication	Character and effectiveness of communication and willingness to communicate, listening, body language

Themes in the two competency frameworks: the BVSc learning outcomes and the VCNZ standards

Based on proportion of words, the relative prominence of themes in the BVSc learning outcomes differed from the relative prominence of themes in the pictures supervisors painted in the interviews (Figure 7.4). The learning outcomes had more focus on the discipline-specific skills of knowledge, applying knowledge, and technical and animal skills than the supervisors' pictures. Technical and animal skills made up high proportion of the learning outcomes because a long list of specific skills was listed. The learning outcomes had less focus on the nontechnical skills of engagement, trustworthiness, communication, social interaction, and personal functioning (none) than the supervisors' pictures.

Similarly, the relative prominence of themes in the VCNZ competency standards differed from the relative prominence of themes in the pictures supervisors painted based on the proportion of words (Figure 7.4). Like the learning outcomes, there was more focus on the discipline-specific skills than the supervisors' pictures, but in contrast to the learning outcomes there was also more focus on trustworthiness. The standards had less focus on the nontechnical aspects of engagement, communication, social interactions, and personal functioning (none) than the supervisors' pictures.

The BVSc learning outcomes and the VCNZ standards were similar to each other in the relative prominence of themes based on the proportion of words. However, the learning outcomes had more focus on technical and animal skills than the standards. The standards had more focus on trustworthiness and knowledge than the learning outcomes. There was also a qualitative difference in the aspects of trustworthiness discussed. The learning outcomes focused only on knowing one's own limitations. The standards also discussed knowing one's own limitations, but also included aspects of practising legally and ethically, and upholding public trust and the integrity of the profession. The concept of trust is specifically mentioned: "[The veterinarian] practises ethically and upholds the public's trust in, and integrity of the profession" (Veterinary Council of New Zealand, n.d., p. 4). This contrasts with supervisors' descriptions, which included responsibility, dependability, as well as knowing one's own limitations.

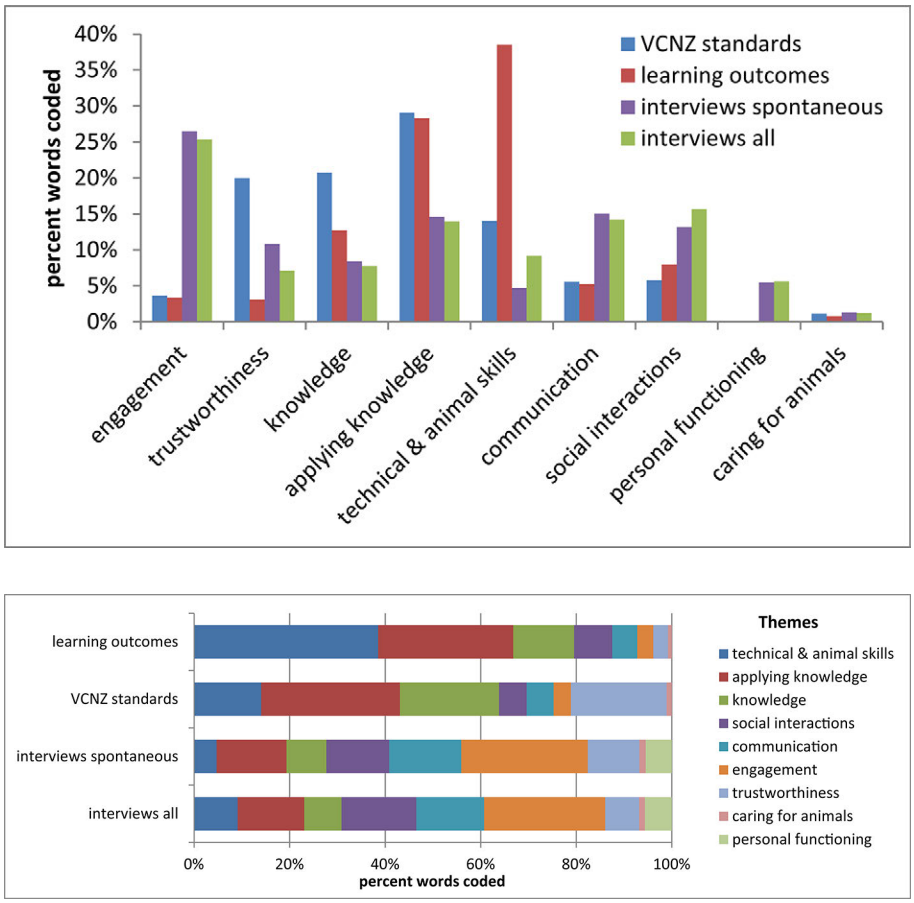


Figure 7.4: Proportion of words devoted to each theme in the VCNZ standards, BVSc learning outcomes, spontaneous descriptions by supervisors in the interviews and all (both spontaneous and clarified and prompted) descriptions in the interviews. The same data is presented in two different formats.

Strengths and limitations of Phase 4

The aim of this phase of the research was to investigate the alignment between what characterises excellence and weakness in supervisors' minds, the dimensions that are assessed in the in-training evaluation, the content of the mark scheme, and the veterinary competency frameworks set forth in the BVSc learning outcomes and the VCNZ standards. In order to do so, the same coding framework that arose from the interview phase was applied to the mark scheme and competency frameworks. Comparisons were then made based on the number and type of words used. In interpreting the findings, certain limitations of the analysis need to be

considered, as well as the differing purposes of the documents. The interpretations are therefore discussed here to inform the discussions in the next chapter.

Using the same coding framework provided an appropriate basis for comparison of the different forms of qualitative data in this phase of the research. It enabled comparison of the themes arising in each. Remaining open to the presence of themes that did not fit the coding framework ensured that the data was not forced to fit the coding framework. The thematic framework proved to be useful in the different contexts of mark schemes and competency frameworks, because all content was able to be captured in one or more themes without redefining themes or adding to them.

A limitation of the analysis, however, is the use of word counts to compare the relative importance of themes in each data source. The proportion of words used to describe a theme does not necessarily indicate its importance. Counting the number of times a theme was coded is another method that has been used to compare importance by other researchers (for example Ginsburg et al., 2010). However, because of the way the data was set up for coding in the software in this research, the number of times a theme was coded would not have been representative of the time or space devoted to a theme, and therefore this was not used. This limitation has been considered when interpreting the findings and other information has been integrated in drawing conclusions, such as the emphasis implied by the type of words used and the structure and purpose of the document.

Another limitation is that the data sources being compared represent a somewhat different set of veterinary subdisciplines. For the factor analysis and regression only a subset of placements that used a common form for in-training evaluation were used. While this subset represented the great majority of placements and included all species subdisciplines, it omitted some specialised subdisciplines in anaesthesia, diagnostic imaging, pathology, and was not fully representative of small animal surgery. These specialised subdisciplines used adjusted in-training evaluation forms, which included specific items relevant to those subdisciplines, and thus could not be included in the same factor analysis. The descriptors for the mark scheme that was analysed were from the common in-training evaluation form used on the major subset of placements, and therefore corresponded to the factor analysis and regression data. In contrast, the interviews were conducted with supervisors from the range of subdisciplines

and, likewise, the BVSc learning outcomes and the VCNZ standards apply to all subdisciplines. Thus, the dimensions and mark scheme may not fully align with the interview descriptions and competency frameworks. However, the in-training evaluations used in the specialised subdisciplines were based on the same four domains of competency and contained many items that overlapped or were only slightly reworded from those on the common form. Therefore, it is likely that there were not substantially different themes omitted from comparison with the competency frameworks. Further research to confirm this would be valuable.

Interpreting the findings

The findings suggest that the balance of themes considered by supervisors in the interviews was mostly well aligned with the mark scheme for the in-training evaluation. Where differences in the proportion of words used occurred they were not large (less than 12% difference) and given that proportion of words used is a crude measure, likely do not indicate a substantive difference. Qualitatively the themes in each were similar with the coding definitions derived from the interviews well capturing the themes used in the mark scheme. Therefore, the relative prominence of nontechnical aspects, and engagement and trustworthiness in supervisors' descriptions was replicated in the mark scheme. An exception was personal functioning, which was notable by its complete absence from the mark scheme, despite being a theme considered by supervisors in the interviews.

Examining the themes coded in the mark scheme by item and factor (as determined from the factor analysis phase) revealed complex structure of the dimensions. Items were mixtures of themes but the mixtures were conceptually appropriate rather than random. They represented the mixture of aspects of competency that would be needed to be brought to bear for the skills represented by the dimension. The mixture of themes within items also meant that dimensions were mixtures of themes. The clinical skills dimension contained many contributing themes. The other dimensions had emphasis on particular themes. The knowledge factor largely comprised knowledge and its application, and the communication factor comprised social interactions, trustworthiness, and communication. The professional attitude factor, which was the factor that was found to have the greatest influence on overall grade in the regression analysis of Phase 3, largely comprised engagement and

trustworthiness. Therefore, the most important themes arising from the interviews in Phase 1 (engagement and trustworthiness) were also the most important themes in influencing the overall grade in the in-training evaluation.

Themes also spanned across items and dimensions. This meant that dimensions were related to each other by having themes in common, and none were exclusive in thematic content. This is consistent with the qualitative findings from Phase 1 which indicated that themes were highly interrelated, and with the high factor intercorrelation found in Phase 2. It also explains the substantial crossloading seen in the self-directed learning item in the factor analysis. This item can be seen to be thematically composed of engagement (Figure 7.2), but to load mostly with the knowledge factor, as well as substantially with the professional attitude factor (Table 5.4, page 130). This suggests that both the student's engagement and knowledge contributed to this factor. Evidence for the engagement that self-directed learning involves may come from the knowledge demonstrated to the supervisor on the placement.

The relative proportion of words for each theme suggested that the BVSc learning outcomes and VCNZ standards differed somewhat in emphasis from the mark scheme and supervisor depictions. Both were more focused on discipline-specific skills than the mark scheme and interview data, and the VCNZ standards were also more focussed on trustworthiness. However, these differences in emphasis may reflect the different purposes of the documents rather than a difference in the conceptualisation of competency. The audience for the BVSc learning outcomes is the not-yet-competent student and the teacher who teaches and assesses the student. The document therefore incorporates a long list of specific technical and animal handling skills to aid student preparation, which substantially increases the proportion of words devoted to that theme. The theme of communication was represented by few words, but occupied an entire learning outcome, suggesting more importance than the proportion of words would indicate. Rather than a different conceptualisation of the balance of knowledge, skills, and attitudes that encompass veterinary competency, the apparent difference in emphasis likely reflects the different purpose of the document.

The purpose of the VCNZ standards is partly informative for a broad veterinary audience and partly legal. The specifications are made in such a way as to be able to be used as defensible criteria for judgement of incompetency in the case of a disciplinary hearing, as well as

providing accessible information and reassurance for both the public and veterinarians. It is not surprising, therefore, that a high proportion of its content was formed by the specifications of the discipline-specific functions of veterinarians. Trustworthiness, too, was a relatively prominent theme in the VCNZ standards where it encompassed concepts not prominent in the learning outcomes, mark scheme, or supervisor interviews, of compliance with statutory requirements, practising in a professional and ethical manner, and upholding public trust.

The BVSc learning outcomes and VCNZ standards were consistent with the mark scheme in containing no mention of themes relating to personal functioning, or prospects of the student, impact on the supervisor, or difficulties in judging competency. While the latter would not be expected in mark schemes or competency frameworks, prospects of the student is sometimes included in mark schemes for in-training evaluations in other veterinary schools (Matthew et al., 2010; Walsh et al., 2012), and impact of the student may, like trustworthiness, prove a useful addition to rating scales. The absence of personal functioning raises issues for validity and the definition of veterinary competency. These aspects will be discussed in the next chapter.

Summary

The comparisons in this phase of the research showed that the themes in the supervisor interviews were reflected in the mark scheme with a similar relative prominence, suggesting good alignment between what supervisors thought important and what the instrument was designed to assess. Items and dimensions contained mixtures of themes but these were conceptually appropriate combinations, expected to be applied together in the clinical situation. The professional attitude factor from the 4-factor solution in Phase 2, was largely comprised of the themes of engagement and trustworthiness. This was the factor that had the most influence on overall grade found in the regression analysis of Phase 3, and the themes comprising it were the themes of most importance to supervisors found in Phase 1. The three phases of research therefore support each other to suggest that what is most important to supervisors is what influences the overall grade the most.

The analysis also showed that themes spanned items and dimensions, providing further evidence for the interrelatedness of items and dimensions and explaining the intercorrelations

found on factor analysis in Phase 2. The interrelatedness was also consistent with that found in the interviews with supervisors in Phase 1.

A somewhat different emphasis on discipline-specific skills and trustworthiness was found in the BVSc learning outcomes and VCNZ standards, which contrasts with the emphasis on nontechnical skills found in the interviews and mark scheme. This difference in emphasis, however, most likely reflects the differing purposes of the competency frameworks, rather than a different conceptualisation of what veterinary competency encompasses. An exception that may mark a difference in conceptions of competency or an issue for the validity of scores was the absence of personal functioning from competency frameworks or the mark scheme. Other themes absent from the mark scheme and competency frameworks were prospects of the student, impact on the supervisor and the difficulty in judging competency.

The results of this analysis show that the coding framework derived from the interviews could usefully be applied to different types of data. The findings deepen our understanding of the linkage between the supervisor's view of performance of the student, the scores they award, and what the in-training evaluation is designed to assess. Their implications will be discussed in the next chapter.

Chapter 8:
General discussion

Critics of the in-training evaluation have suggested that supervisors may base their judgement on insufficient or irrelevant aspects of student performance. This research explored aspects of the supervisor judgement in veterinary contexts in order to deepen our understanding of what supervisors are basing their judgement on and how that relates to what they are expected to assess in in-training evaluations. The preceding four chapters have presented the findings from the interview phase of research, which explored the aspects of student performance that supervisors value; two quantitative phases, which characterised the dimensions of the in-training evaluation that inform overall grade; and a final qualitative phase, which drew comparisons between the quantitative and qualitative findings and the competency frameworks relevant to New Zealand veterinary students. In this chapter, the findings are synthesised across all four phases and situated with respect to the literatures on veterinary competency and in-training evaluation. Seven aspects of the findings are discussed: (1) supervisor judgement as holistic and discriminating; (2) the importance of engagement and trustworthiness; (3) the alignment of what supervisors value with assessment criteria and competency frameworks; (4) issues for the meaning of veterinary competency and its assessment; (5) the influence of veterinary subdiscipline on in-training evaluation; (6) insights into the process of judgement; and (7) methodological considerations in determining the factor structure.

Supervisor judgement as holistic and discriminating

The findings of this research suggest that supervisor judgement is a complex, holistic evaluation of multiple, interrelated—but distinct—aspects, and moderated by contextual factors. Consideration of the way things are related and how they are influenced by the surroundings is fundamental to holistic thinking and is contrasted with analytical thinking which treats aspects as distinct and separable (Nisbett, Peng, Choi, & Norenzayan, 2001). Therefore, it may seem a contradiction to conclude that supervisor judgements are both holistic and discriminating, yet considering both the detail and the overall picture is consistent

with Sadler's (2009b) definition of holistic grading as involving attention to both the particular aspects and to the quality of the work as a whole. Hoy-Mack (2005) similarly concluded that workplace-based performance assessment involved a combination of holistic and analytical judgement, which she described as involving a constant "flicking back and forth between the whole picture and competencies" (p. 87). In my research, several findings across phases of the study point to the multiple aspects balanced by supervisors, their interlinked and overlapping nature, and their differential weighting and influence which indicates their independence. In this section, I will first discuss how the qualitative findings lead to this conclusion and then how it is supported from the quantitative findings.

The interview phase of research revealed multiple themes that supervisors used when describing excellent, weak, and marginal students. Each theme, while representing a cluster of related ideas, spanned a range of separate abilities, behaviours, attitudes, and personal characteristics. Similar studies that investigated supervisor descriptions of excellent and weak medical and social work students have also found multiple themes described (Bogo et al., 2006; Ginsburg et al., 2010; Lavine et al., 2004; Rosenbluth et al., 2014). The interrelatedness of aspects described by supervisors in this research was apparent because they often described different aspects together, even in the same sentence, indicating their association. As a result, dual coding was required to successfully capture the themes in the transcripts. Likewise Ginsburg et al. (2011) found that written comments on in-training evaluation forms often spanned more than one aspect of performance at once. In this research the aspects described together were sometimes similar, such as aspects of communication and social interaction, but often conceptually distinct, such as animal handling and confidence, or knowledge and communication. Yet it will be apparent to readers that both the similar and distinct concepts described together are aspects that would be brought to bear at the same time in performing certain types of veterinary work, and therefore represent interrelated parts of veterinary competency. Recognition that any work performance requires a combination of aspects of competency has led to the conceptualisation of competency as a multidimensional but holistic and unitary concept (Le Deist & Winterton, 2005). The relational nature of competency was a key part of its definition by Gonczi (1994) as the "complex structuring of attributes needed for intelligent performance in specific situations" (p. 29).

As well as supervisors speaking of different themes together, patterns of relatedness were seen in the co-occurrence of particular positive and negative themes within descriptions of

particular students. For example, positive engagement was associated with positive application of knowledge, and negative engagement with negative trustworthiness. The co-occurrence is likely a result of their influence on each other, as engagement would help supervisors determine a student was able to apply knowledge and a lack of engagement would make entrustment difficult. However, exceptions showed that the aspects were independent, even though they were closely related. For example one student described was well able to apply knowledge when required (positive application of knowledge) but with an unwilling, don't care attitude (negative engagement). In addition, the aspects considered were found to assume different importance depending on the context of the overall performance of the student, the type of work of the placement, and the inferences supervisors made about the reasons for student behaviour. This tends to indicate that although related, the aspects considered by supervisors were distinct enough to contribute independently to the overall picture. Evidence that suggests aspects of performance are differentiated has also been found in medical education. Similar to the findings in this research, Ginsburg et al. (2010) noted that medical supervisors differentially weighted aspects of performance depending on the other aspects of the student's performance. In a hierarchical cluster analysis Sebok and Syer (2015) showed how the aspects of performance that were linked could differ when raters considered performances involving different cases.

The mark scheme for the in-training evaluation also indicated that aspects of performance were interrelated, as found in Phase 4 of this research. Very few items were thematically pure. Most involved combinations of themes but with greater or lesser emphasis on different aspects. This meant that the dimensions (factors) found on factor analysis were also not pure themes, but they included combinations of themes. Themes were combined in ways that were conceptually consistent with the combination of skills needed for competency. For example the combination of application of knowledge, technical and animal handling skills, communication, and caring for animals in the clinical skills factor, along with engagement and trustworthiness, reflects what is needed to consult with a client and examine their animal. This indicates that the separate components contributed meaningfully to the whole dimension and their combination was not just a jumbled association. This internal structure therefore suggests that supervisors were evaluating how distinct components of competency were being applied together—an overarching metacompetency inherent in the definition of competency (Fernandez et al., 2012).

The way themes were clustered within items and factors often illustrated the interlinking of themes that was seen qualitatively in the interview phase. For example, the link between engagement and trustworthiness, seen in supervisors' descriptions, was also apparent in the thematic phase (Phase 4) findings with both being large components of the professional attitude dimension. The linking of engagement and trustworthiness with many other themes was apparent with them each being present in three of the four dimensions. As previously discussed, overlap and interconnectedness of aspects of competency is appreciable in frameworks such as the taxonomy of competencies for health professionals constructed by Englander et al. (2013). Domains of competency are not pure themes.

Thus, in the interview phase, both interrelatedness and independence of the themes described by supervisors were apparent, and are in keeping with previously published findings. The ability to consider and discriminate several aspects of performance while still taking a holistic view, is, as Govaerts et al. (2013) concluded, directly contradictory with suggestions from previous factor analysis studies that only one or two dimensions are considered. This conclusion assumes that distinctions between different aspects are maintained during scoring and successfully translated to scores. This is a point I will return to shortly.

The interlinked thematic content of items on the evaluation would be expected to manifest as intercorrelation between item scores on the in-training evaluation. This was what was found in the factor analysis in Phase 2. It suggests that the item and factor intercorrelation is an expected part of the constructs being assessed. Since halo error is defined as more than expected intercorrelation between dimensions (Murphy, 1982), it is apparent that item and factor intercorrelation should not be entirely attributed to halo error. In other words, concluding that supervisors are mistakenly linking aspects of performance when making an overall evaluation is not in keeping with relational and holistic definitions of competency and with the interlinked nature of item descriptors and frameworks that guide their judgement. Without further investigation, it is impossible to tell how much item and factor correlation might be considered error.

Evidence that suggests the distinctions between different aspects of performance are successfully translated to item scores is provided by the findings of the factor analysis and regression phases of this research. The 4-factor higher-order factor model suggests supervisors

do discriminate the four dimensions the instrument was designed to assess, under the umbrella of an overall holistic judgement. Although not the only plausible factor solution, it provides a conceptual model of supervisors considering both the whole and the individual parts, and their judgement being both holistic and discriminating. It is thus consistent with the findings of the qualitative phases of this research. The finding that four factors were discriminated is in contrast to the findings of most previous studies, which, as previously reviewed, suggested that only one or two dimensions were discriminated in in-training evaluations (Table 2.2, page 33). Most factor analysis studies of in-training evaluations have not considered the possibility of a higher-order structure and indeed, there would be no point in studies in which 1- or 2-factor solutions were found. However, two studies in which three or more factors were found reported the factor intercorrelations (Brasel et al., 2004; Gough et al., 1964). In both studies, these were substantial, suggesting the likelihood of an underlying higher-order factor structure if this had been investigated.

Further support for the idea that the factors, although related, represent distinct dimensions comes from the regression analysis phase of the research. The results of the regression analysis indicated that each dimension (factor) was significantly related to the overall grade awarded. Each had a different strength of relationship with overall grade, with professional attitude factor the strongest, followed by (in decreasing order of effect) the clinical skills factor, the communication factor, and the knowledge factor. The fact that each factor had a different influence on overall grade suggests that the factors are separate to each other and operate independently, despite their interrelationships. It also suggests that supervisors did not just average the dimension scores to obtain an overall grade, in contrast to the practice of some supervisors reported by Kogan et al. (2011).

In sum, evidence from the interviews confirms that multiple aspects of performance are considered, and that these are highly related but also separately distinguished. Evidence from the factor analysis and regression phases confirms that the concurrently holistic and discriminated structure of evaluations is translated to scores. Evidence from the thematic analysis of the instrument confirms that interrelated aspects are an expected part of the evaluation. This is also consistent with conceptions of competency. The findings therefore not only suggest that in-training evaluations are both holistic and discriminating but that item and factor intercorrelation is not all attributable to halo error.

The importance of engagement and trustworthiness

Another finding of this research was the relative importance of student engagement and trustworthiness in determining the overall evaluation made by supervisors. In the interview phase, the importance of engagement was highlighted by the fact that it was the theme discussed more than any other in the spontaneous comments of supervisors and they often discussed it first or said it was important. Trustworthiness was important because so many supervisors discussed elements of trustworthiness in reference to weak students, and because it was never traded off. A sense of its importance was also conveyed by the type of words used. The prominence of both engagement and trustworthiness was mirrored in the mark scheme for the in-training evaluation, as found in Phase 4 of the research. Although there were no specific items relating to engagement or trustworthiness, these themes were still emphasised by being present in the level descriptors for many of the items. Their presence suggested good alignment between the mark scheme and the most important aspects supervisors considered when discussing excellent, weak, and marginal students.

Engagement and trustworthiness were also prominent in the thematic content of the professional attitude factor on the in-training evaluation, which was the dimension with the most influence on the overall grade awarded, as found in the regression analysis of Phase 3. Therefore the most important aspects supervisors considered when discussing excellent, weak, and marginal students, were also the most important aspects of the mark scheme for in-training evaluation, and had the most influence in determining the overall grade awarded. Thus, the findings of the four phases of this research support the conclusion that engagement and trustworthiness are the most important aspects determining overall grade in the veterinary in-training evaluation in use at Massey University. Furthermore, this finding supports the alignment of the scoring processes with the mark scheme for the evaluation, and suggests that, in respect of these two aspects, the in-training evaluation is assessing what it was designed to assess.

The importance of engagement to supervisors has been noted before by other researchers. Several studies, which similarly investigated descriptions of excellent and weak students on placements, found engagement as a prominent theme. Ginsburg et al. (2010) found that work ethic was the most frequently mentioned theme in medical students. Bogo et al. (2006), working with supervisors of social work students, found that initiative, energy, commitment,

and approaches to learning were important aspects to supervisors, who used terms like motivated and eager to learn. Energy was so frequently mentioned by medical student supervisors, in a study by Rosenbluth et al. (2014), that they felt it warranted a separate theme they called the "zing factor". It included energy, passion, enthusiasm, excitement, interest, motivation (including motivation to learn), and inquisitiveness. Being energetic and enthusiastic were also features commented on by doctors, nurses, nursing students, and patients as characterising excellent nurses (Medigovich, 2012). In regards to veterinary students on placements, two recent studies of the motivations of supervisors indicated that engagement and interest of the student was important in maintaining supervisor interest in teaching (Hashizume, Myhre, Hecker, Bailey, & Lockyer, 2016; Scholz et al., 2015). A commissioned report on the opinions of 110 supervisors of veterinary students in Australia emphasised the frustration and resentment felt by supervisors when students were unmotivated and "just going through the motions" (Scarlett, 2013). In addition, Hashizume et al. (2016) found that supervisors were concerned that uninterested and unenthusiastic students could negatively impact their own professional image with clients. Trust has also been found to be important to supervisors in other research. Ginsburg et al. (2010) identified trust, comprising believability and discernment, as a theme in descriptions of excellent and weak medical students and in the feedback comments supervisors provide on in-training evaluations (Ginsburg et al., 2011).

A relationship between engagement of the student and the overall grade awarded on in-training evaluation has also been identified in some previous research. In veterinary education, Matthew et al. (2010) found that higher in-training evaluation scores were significantly related to an engaged approach by students on placements. In medical education, Quarrick and Sloop (1972) and Benyamini, Kedar, and Raveh (1987) both found student engagement to be more highly correlated with overall grade than other aspects of performance. The broader domain of professionalism as a whole was found more highly correlated with overall grade than other domains of performance in medical students by Pulito et al. (2007), but others have found it to be not as highly related to overall grade as discipline-specific skills (Chahine, Holmes, & Kowalewski, 2016; Verhulst et al., 1986).

The relative prominence of engagement and trustworthiness in the descriptors of the mark scheme reflected the importance that the assessment designers placed on these aspects. In-training evaluations in use in other veterinary schools have also referred to these themes with

items that explicitly asked about engagement and trustworthiness (Fuentealba & Hecker, 2008), or elements of them (Root Kustritz et al., 2011; Walsh et al., 2012). This suggests that engagement and trustworthiness are also considered important by other assessment designers in other veterinary schools.

Neither engagement nor trustworthiness, however, featured strongly in the BVSc learning outcomes analysed in Phase 4 of this research, although both were present. This relative lack of prominence may simply reflect the balance of words devoted to specific guidance for students about the other aspects of competency they are to achieve, which are more discipline-specific. In contrast, the VCNZ standards prominently incorporated elements of trustworthiness. Knowing limitations is an aspect of trustworthiness that has been identified as the most important new graduate competency in the *RCVS Day One Competences* for veterinarians (Royal College of Veterinary Surgeons, 2014). Although engagement was not an important feature of the VCNZ standards, it is important to the profession in New Zealand in terms of continued professional development and is the focus of a large amount of other documentation and guidance for veterinarians. Elements of trustworthiness and engagement were also present in a taxonomy of health professional competency developed by Englander et al. (2013) from a review of published competency frameworks. Thus, engagement and trustworthiness are important to the profession as a whole and are present in competency frameworks, even though not prominent in the BVSc learning outcomes.

Of interest is whether engagement and trustworthiness are themselves competencies or are reflections of the degree of competency. As reviewed in the literature review, the use of trustworthiness as a basis for assessment criteria, in the form of entrustable professional activities, is founded on the idea that entrustment requires the breadth of competency to be demonstrated (ten Cate et al., 2015), and therefore trustworthiness is an indicator of competency. Engagement may act similarly. Although theories of employee engagement are not yet well developed, current conceptualisations link engagement with aspects that include feeling valued, and physically, emotionally, and psychologically capable of performing the roles required of the job (Saks & Gruman, 2014). Dornan, Boshuizen, King, and Scherpbier (2007) found that medical student's own level of confidence and competency contributed to their motivation to participate in the clinical workplace. Thus feeling competent is likely to be at least one requirement for engagement in the workplace, and engagement may therefore be an indirect measure of competency, rather than reflecting a competency in itself. There may,

however, be a great deal of difference between feeling competent as a learner in a supervised environment, to being competent as a newly graduated veterinarian.

Both engagement and trustworthiness may therefore reflect the degree of competency of a student rather than being components of competency themselves. In addition, their importance may derive from their role as contributors to the achievement of competency. Students need to be engaged in order to learn (Finn & Zimmer, 2012) and participation underpins learning in clinical workplaces (Dornan et al., 2007). Bok et al. (2015) found that supervisors gave more clinical responsibilities to highly engaged students. One may argue, on this basis, that engagement is not an indicator of attainment, but rather a process through which attainment occurs, and should not be part of assessment criteria (Sadler, 2009a). A similar argument could be made for trustworthiness, which may also be part of the process of becoming competent. Students need to be trusted in order to be given opportunities to participate, and therefore learn, and must participate in order to be trusted, as discussed by Hauer et al. (2014). This may point to a key reason for the importance of engagement and trustworthiness to supervisors, and their close interrelatedness, as found in this research. Indicators of the student's interest and motivation to participate and learn may be even earlier starting points for development of trust by the supervisor than the actions described by Hauer et al. (2015), of checking of a student's work and remaining in close proximity.

There is some evidence that engagement and trustworthiness may be important in clinical practice, which would provide support for extrapolation from performance as students to performance in later practice. Matthew et al. (2011) have shown the importance of engagement amongst the types of behaviours that promote success in clinical practice as new veterinary graduates. Engagement of newly graduated veterinarians was associated with personal resources that can minimise burnout and other negative effects (Mastenbroek, Jaarsma, Demerouti, et al., 2014). A link between both engagement and trustworthiness with success in medical practice was also noted by Teherani, Hodgson, Banach, and Papadakis (2005). They found that negative feedback on in-training evaluation relating to poor reliability and responsibility, lack of self-improvement, poor initiative, and poor motivation was predictive of future disciplinary proceedings against practitioners. While these findings suggest that engagement and trustworthiness as a student predict graduate performance, they provide very limited evidence and more research is required.

A difficulty with basing assessment of students on engagement and trustworthiness is that both are dependent on the environment and others in it. For example, Mastenbroek, Jaarsma, Scherpbier, van Beukelen, and Demerouti (2014) found that work conditions that provided autonomy, social support, and feedback promoted feelings of self-efficacy and therefore work engagement in recently graduated veterinarians. A variety of factors are thought to influence student engagement including provision of a safe and supportive environment and encouragement of student interaction (Finn & Zimmer, 2012). In a study of junior medical student's experiences in the workplace, Dornan et al. (2007) found that the supervisor's enthusiasm, support, and provision of the right level of challenge, were important in motivating students. Decisions to trust medical students have been found to be influenced by qualities of the supervisor, such as their experience, and their attitude towards clinical training; of the circumstances, such as the time of day, urgency of the case; and of the environment, such as the institutional culture, and the functioning of the caregiving team (Choo, Arora, Barach, Johnson, & Farnan, 2014; Sterkenburg, Barach, Kalkman, Gielen, & ten Cate, 2010). Engagement and trustworthiness are, however, not the only aspects of competency that are influenced by the environment and others in it, and it may be argued that this extends to all aspects of competency. Communication is an obvious example of an aspect of competency that necessarily depends on the competency of others. Lingard (2012) describes a way of seeing competency as distributed amongst teams, systems, and contexts, rather than as a property of an individual. The challenge of assessing students based on aspects that are not entirely of their own making is therefore not limited to engagement and trustworthiness.

In summary then the finding of the importance of engagement and trustworthiness to supervisors and their influence on the overall evaluation is in keeping with their importance to the profession and demonstrates the alignment of the instrument. However further research would be useful to clarify the relationship of engagement and trustworthiness with student and graduate competency, and their usefulness as indicators for assessment.

Alignment of what supervisors value with the assessment criteria and competency frameworks

For inferences based on assessment scores to be valid, scores must reflect the construct being assessed, as dictated by the assessment criteria, and not reflect other, construct-irrelevant aspects (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In addition, the assessment criteria must themselves be relevant for the situation that scores will be extrapolated to, such as future performance in the workplace. Without this alignment, the validity of scores can be questioned.

The findings of this research suggest generally good alignment of what supervisors consider with the mark scheme and competency frameworks. However, some differences of note were seen. There were areas where what supervisors valued differed from the apparent emphasis in the mark scheme and competency frameworks, and supervisors generally focussed more on nontechnical aspects of competency than discipline-specific aspects. Furthermore, supervisors may have been applying a varying standard that accounted for the student's stage of training.

General alignment

The findings of this research showed that what supervisors valued, which influenced their overall opinion of a student's performance, was generally well aligned with the mark scheme for the in-training evaluation. The balance of themes considered, and the importance of engagement and trustworthiness, was similar in both the interview findings and the mark scheme as shown in Phase 4. These findings suggest that supervisors' expectations align with what the instrument is intended to assess, with some exceptions that will I will return to shortly. Furthermore, the finding that the aspects of most prominence in the mark scheme (engagement and trustworthiness) had the most influence on overall grade provided evidence for the successful translation of supervisor judgement into scores on the instrument. These findings provide initial evidence that as a group, supervisors are assessing what they are asked to assess and that the scoring process enables successful translation of judgement to scores. However, it does not provide evidence at an individual level and, as seen in the literature review, there is considerable reason to expect a great deal of variation between supervisors in

the aspects they pay attention to and the scores they award (Govaerts et al., 2013; Mazor, Zanetti, et al., 2007; St-Onge et al., 2016; Yeates et al., 2013).

There were four aspects of student performance that were described by supervisors in the interviews but not present at all in the mark scheme for the in-training evaluation. These were the personal functioning, impact on the supervisor, prospects of the student, and the difficulty in judging competency. The last of these—the difficulty in judging competency— would not be expected in mark schemes, because it concerns the instrument rather than the performance of the student. As a theme, however it raised important issues that are discussed later. Both personal functioning and impact seemed to influence the supervisor's opinion of student performance, either positively and negatively. These are therefore areas where there was a lack of alignment between what supervisors valued and the constructs set down for assessment in the in-training evaluation. While this may indicate that they are construct-irrelevant aspects that jeopardise the validity of scores, they may also point to areas where our conceptions of veterinary competency are not well articulated in mark schemes. This will be discussed further in a subsequent section on what the findings reveal about the meaning of competency. The student's prospects served more to illustrate the qualities of students being described than to influence supervisor's judgement of student performance, however this theme is also discussed further in a subsequent section as it is sometimes included in mark schemes for in-training evaluations in other institutions.

The alignment of the themes from the interviews with the competency frameworks was not as clear as its alignment with the mark scheme. On the whole, there was good representation of themes from the interviews in both frameworks, but there was a greater apparent emphasis on discipline-specific competencies in the frameworks than in the supervisor interviews or mark schemes which will be discussed in the next section. The emphasis on trustworthiness in the VCNZ standards has already been discussed.

Balance of nontechnical and discipline-specific aspects

In this research, discipline-specific aspects of what supervisors valued were captured thematically as knowledge, application of knowledge, and technical and animal handling skills. All other themes were nontechnical, and therefore not skills and attributes specific to

veterinary medicine, and these were what supervisors concentrated on in their descriptions. Less than a third of the words supervisors used in their spontaneous descriptions of students concerned discipline-specific aspects. This balance of focus was also seen in the emphasis used in discussing the various aspects of performance.

Of the discipline-specific aspects, application of knowledge was the most important to supervisors, especially in relation to excellent and marginal students. Animal handling skills were important when they made students dangerous and untrustworthy. Technical skills and knowledge were less important and could be compensated for if other aspects of performance were excellent or weak. The relative lack of importance of knowledge was also seen in the regression analysis, in which there was only a small and non-significant relationship between the overall grade achieved on in-training evaluation and the preclinical GPA (primarily derived from knowledge-based assessments). In addition, the knowledge factor had the least effect on overall grade of any of the factors, and some supervisors indicated in the interviews that it could be difficult to assess knowledge during the clinical rotations. These findings in relation to knowledge are consistent with the lack of emphasis on knowledge seen in supervisor depictions of excellent and weak medical students (Ginsburg et al., 2010; Rosenbluth et al., 2014). They are also consistent with the findings of Kan Ma et al. (2013) who studied the perceptions of medical student supervisors. Those researchers found that exceptional knowledge was not as much valued as exceptional performance in other aspects of practice, and also that weak knowledge was not viewed as negatively as weak performance in other aspects.

The mark scheme has a similar lack of emphasis on discipline-specific skills. In contrast, the emphasis was completely reversed in the competency frameworks represented by the BVSc learning outcomes and VCNZ standards, for which the greater proportion of the word count concerned discipline-specific aspects of competency. These differences in emphasis may reflect the fact that these competency frameworks are primarily concerned with the aspects that differentiate veterinarians from other people, as their purpose is to communicate what it is to be a veterinarian to both students and a wider public audience. Little description may be needed of aspects of competency that are more generally applied and understood, and this need not indicate that these are less important aspects of performance. Other competency frameworks, such as that produced by the North American Veterinary Medical Education Consortium (2011), tend to sum up discipline-specific knowledge and skills in one or a few

components, and therefore the nontechnical competencies can appear more numerous. The taxonomy produced by Englander et al. (2013) which was based on a review of competency frameworks across the health professions and including veterinary medicine, similarly collapses discipline-specific knowledge and skills into a few components. These findings serve to illustrate that the proportion of words used is not necessarily a good measure of the relative emphasis given and therefore, despite the differences in apparent emphasis in the supervisor interviews and mark scheme, the general confluence of themes indicates alignment.

Nevertheless, the findings of this research in relation to the lack of emphasis on knowledge, and technical skill, tend to indicate that the in-training evaluation is not sufficient as a sole assessment of all aspects veterinary competency, and that it should continue to be combined with a variety of other assessments that focus on discipline-specific aspects more specifically. They also underscore the inappropriateness of using GPA or other knowledge-based tests as a criterion measure of validity (that is, a benchmark for comparison) for in-training evaluation scores.

Alignment of standards

Alignment of scores with the mark scheme pertains not only to what is considered. The standard that is applied should also reflect the standards set down by the mark scheme. Consistent application of these standards would be expected to manifest as an increase in a student's scores over time, as their level of competency increased through the year. However, no significant change in scores over time was found during the regression phase of this research. This could indicate that the study was underpowered to detect such a difference, or that supervisors account for the stage of training when making judgements. Supervisors may use a relative level of performance (that is, the meaning of satisfactory is satisfactory for that stage of training) rather than a fixed and absolute level. Walsh et al. (2012) reported this was the case in another in-training evaluation in veterinary medicine. The finding could also indicate that students do not detectably improve over the year of placements. The main contributors to overall grade may be fixed traits and attributes, not learned ones, or aspects that do not show a great deal of change within the period of the year. This is perhaps consistent with the finding of this research, that the main aspects considered by supervisors were the nontechnical ones, which may already be well developed or potentially fixed, in

veterinary students before they enter the final year of training. Alternatively, as Boerebach, Arah, Heineman, and Lombarts (2016) discussed, the five-point Likert-type scale used may be too crude to detect small changes in performance and the skewed distribution of scores, with most students already performing at high levels, may also inhibit detection of improvement. Bok et al. (2013) found that rater leniency and a tendency for scores to be high limited the ability to detect improvement over the year in performance assessments of veterinary competency. However van Lohuizen et al. (2010) did detect significant improvement over a 14-week period of evaluations in medical students using the mini-clinical evaluation exercise (miniCEX) which is also a rater-based performance assessment. Thus it is not clear whether supervisors apply the same standard over the time course of assessment or not, and further research would be needed to determine whether any or all of these factors contributed to the lack of change over time found in this research.

Summary

In summary then, what supervisors value was generally well aligned with the constructs set out for assessment in the marking criteria for the in-training evaluation and the competency frameworks but there was a tendency to concentrate on nontechnical aspects of competency and some aspects considered which were not part of the mark scheme or competency frameworks. In addition, supervisors may vary the standard applied to account for the student's stage of training. These differences raise issues in regards to the meaning of veterinary competency that will be discussed in the next section.

Issues for the meaning of veterinary competency and its assessment

The findings of this research give some insights into what the concept of veterinary student competency means to supervisors. Overall, the findings suggest that competency is a multidimensional concept in which aspects take on varying importance depending on the situation and overall level of performance of the student. Nontechnical aspects, and especially engagement and trustworthiness, are prominent determiners of the overall impression of competency. Some additional aspects that also contribute to the overall impression are not

present in the mark scheme or competency frameworks, including personal functioning, impact on the supervisor, and prospects of the student. Another aspect that supervisors spoke of, caring for animals, was present in the mark scheme and competency frameworks but showed some differences in the way it was expressed. These differences raise questions about the meaning of veterinary competency and how it should be assessed.

Personal functioning

The student's management of their own performance in the face of personal issues seemed to influence supervisor's opinions, either positively or negatively. Personal functioning was commented on by about half of supervisors in relation to marginal students because whether to make allowances for personal functioning in their evaluation presented a dilemma. This theme was completely absent from the mark scheme and competency frameworks. However elements of personal functioning, such as confidence, were present, but infrequent, in veterinary competency frameworks from other institutions (Cake et al., 2016). Confidence was one of the attributes that facilitated balance in the model of veterinary professionalism developed by Mossop (2012). No similar theme was found in descriptions of excellent and weak students by medical educators (Ginsburg et al., 2010), but confidence was mentioned in feedback comments on in-training evaluations of medical students (Ginsburg et al., 2011), and coping well with stress was incorporated into the taxonomy of competency domains for health professionals by Englander et al. (2013). Confidence and ability to cope with pressure were thought very important attributes by over 60% of veterinary student and new graduate respondents (Rhind et al., 2011), and similarly confidence was an area identified by veterinary employers that new graduates needed most help with (Heath & Mills, 1999). There is some indication, therefore, that this may be an aspect of veterinary competency that should be more clearly articulated in frameworks. More research would be helpful to determine how the impact of personal functioning should be managed in assessment criteria.

Impact on the supervisor

Another area not present in the mark scheme or competency frameworks but which was discussed by supervisors in the interviews was the student's impact on them or other staff. Supervisors commented on changes in workload and enjoyment of teaching. Impact was

closely linked to trust. Although impact was mentioned regarding fewer students than trust, where it was mentioned, positive impact was always associated with positive trust and negative impact always associated with lack of trust. Hauer et al. (2015) noted that entrustment of senior medical students by supervisors lead to a shift in the supervisor's role, with less time spent checking the student's work, less direct work with patients and more consulting and teaching work. These changes were also accompanied by less anxiety and more sleep. Because veterinary students are at an earlier stage of training and cannot carry as much responsibility, they are not likely to have as much impact on a supervisor as the senior medical students that Hauer et al. (2015) describes, but nevertheless still could have a noticeable impact. Impact has also been found as a prominent theme in studies of supervisors' descriptions of excellent and weak students in medical and social work education (Bogo et al., 2006; Ginsburg et al., 2010). Like trust, impact relates to how a supervisor feels about a student rather than being a characteristic of the student themselves. Therefore, while it is not part of competency frameworks it is an aspect that contributes to a supervisor's opinions of students, and, like trust, may be an indirect indicator of competency, since for a student to have a positive impact likely requires a certain degree of competency. Impact may therefore be a useful additional criterion in rating scales, in the same way that entrustment is currently proving useful for assessment of competency (ten Cate et al., 2015).

Prospects of the student

The prospects of students was a distinct theme, though not frequently used in spontaneous comments. Rather, comments about potential career pathways and employment prospects were used to elaborate and provide examples of performance. However, its use shows that supervisors were thinking forward, and that their views were shaped by a concept of competency in the workforce. The finding is in keeping with research showing that written feedback on in-training evaluations in medical education frequently contained mention of a resident's future performance or success (Ginsburg et al., 2011). The researchers interpreted this as evidence that assessors provide an integrated holistic assessment of competencies rather than thinking in terms of isolated individual competencies. In the past, some medical in-training evaluations have incorporated this theme deliberately in the instrument as an item, for example as reported by Quarrick and Sloop (1972), Gough et al. (1964), and Davis, Inamdar, and Stone (1986). No such item was present in three veterinary in-training evaluation instruments reported (Fuentealba & Hecker, 2008; Root Kustritz et al., 2011; Roush et al.,

2014), but two others from veterinary medicine contained an item relating to the student's trajectory towards meeting the graduating competencies by the time of graduation (Matthew et al., 2010; Walsh et al., 2012). Such an item also calls upon supervisors to predict future potential. Walsh et al. (2012) felt it was useful for flagging poor performance because supervisors seemed to score it less leniently than the overall grade. Its potential use as a holistic way of conveying the standard achieved may be a worthwhile area of further investigation.

Caring for animals

In the interview phase, caring for animals was found to be a theme distinct from animal handling ability because amongst supervisors' descriptions were students who did handle animals well, but still did not seem to care for them or like them. Such behaviour was important to some supervisors and contributed to them thinking of a student's performance as weak. This theme seemed to have a differential importance as positive caring could not balance other weaknesses, implying that it was expected that veterinary students would care for animals. The mark scheme referred to aspects of patient care and empathy which mapped to caring for animals, whereas in the BVSc learning outcomes and the VCNZ standards, caring for animals was framed in terms of protecting animal welfare and relief of pain and suffering, rather than caring and liking. There are subtle differences here, centering on liking animals, which raise questions about whether veterinarians need to like animals to be competent. Part of the veterinary image is of a compassionate, empathetic carer of animals (Walsh, Osburn, & Christopher, 2001) and caring defender of animal welfare (Willis et al., 2007). Showing affection towards small animal pets appears to be expected by clients. Case (1988) found that 81.6% of small animal clients surveyed agreed or strongly agreed that their veterinarian should talk affectionately to their pets. Love of animals, as well as concern for and interest in animals, was ranked highly as a reason for choosing veterinary medicine as a career by first year students (Heath, Lynch-Blosse, & Lanyon, 1996). This may explain why supervisors seemed to expect it, but does not indicate whether it is necessary for competency. In agreement with the findings of this research, Mossop (2012) found that veterinarians considered a lack of caring for animals to be a sign of unprofessionalism. One reason that students may appear uncaring is that they are fearful or uncomfortable around animals, which may indicate a lack of competency in handling animals. Another reason, however, might be cultural practices and preferences. A person may like some types of animals and not others. Most of the comments

by supervisors in relation to caring were clearly about pet animals, but some students are most interested in farm animal veterinary work where the relationship with the animal can be different. However, one would still expect them to care for animals, as concern for animal welfare is an important part of competency frameworks. A question for the profession then is whether caring for animals is important for veterinarians and if it is, is it something that can be learned, or something that we should select for on entry to the programme?

Summary

Some aspects of what supervisors discussed in the interviews were not well aligned in emphasis or apparent meaning with the mark scheme or competency frameworks. These could be viewed as rater error, or may indicate areas where further consideration by the profession is needed about what it means to be a competent veterinarian and how we can assess that. The findings pointed to issues regarding whether the way a person manages their personal functioning, or aspects of it, is a competency; whether consideration of a student's impact on the supervisor and others could usefully be incorporated in assessment criteria; whether predictions of a student's prospects are a helpful way to convey the standard achieved; and whether caring for animals, in the sense of actually liking them, is necessary for veterinarians.

The influence of veterinary subdiscipline on in-training evaluation

The interview phase of this research captured a range of themes that were common to supervisors even though they were from a range of different subdisciplines. Thus, there was a similarity in the overall aspects considered by supervisors irrespective of subdiscipline. However, the findings from the regression analysis in Phase 3 indicated that the overall evaluation of students depended to a degree on the veterinary subdiscipline the students were working in at the time. The variable placement indicated the subdiscipline, and was found to be an important influence on overall grade. Different subdisciplines differed markedly in their propensity to award grades of various levels. Placement also interacted with the knowledge factor and the professional attitude factor so that the effect of knowledge and professional

attitude on overall grade depended on which placement was considered. Although professional attitude always had the strongest effect compared to other factors, the magnitude of the effects of the professional attitude and knowledge factors on overall grade varied across placements. This indicated that there were differences in the emphasis given to these dimensions by supervisors in different subdisciplines, although professional attitude remained the most important.

It was possible that at least some of this effect of placement reflected an effect of different supervisors—a rater effect. This is because, in this research, placement and supervisor were confounded, in that supervisors mostly were confined to one particular placement, and therefore when students went to different placements they also had different supervisors. It was not possible to obtain data about individual supervisors from the databases in order to account for this. However, each type of placement involved multiple different supervisors. Therefore the effect of placement indicated that there was some degree of similarity between the judgements of different supervisors on the same placements and suggested that the context of the placement itself had some effect on their overall judgement. This means that not all of the similarity and variation between placements was due their involving different supervisors. Some of it was an effect of placement irrespective of the supervisor. In contrast to the effect of placement, the academic status of the supervisor (indicating whether they were university-based or from external workplaces) was not a significant predictor of overall grade. Placement types awarding higher grades included both academic and external workplaces, and, likewise, placements awarding lower grades included both academic and external workplaces.

The effect of placement but not of academic status, is an interesting finding that could be interpreted in several ways. Firstly, of course, the finding of no difference in the scores awarded by academic and non-academic supervisors does not mean that no difference exists, and the research may have been insufficiently powered to detect a difference between the grades awarded by academics and non-academics. Secondly, the findings may indicate that there were real differences in the performance of students on different placements, irrespective of whether the placement is located in an academic or non-academic location. The finding of a significant interaction with placement and the professional attitude factor and the knowledge factor suggests that these might be key areas of difference in the performance of students on different placements. In regards to knowledge, placements may have differed in

difficulty or range of skills required according to the subdiscipline involved, or the curriculum may have prepared students better for some placements than others. This would not be surprising as curricula often have an emphasis that depends on the school's mission, location (for example rural or urban), the expertise of staff, and the local professional, political, and economic drivers for training of veterinarians in particular discipline areas (for example Abbott, 2009). Also students may have more experience with some types of veterinary subdiscipline work prior to entry to the programme, for example, depending on whether their background is primarily rural or urban (Heath, Hyams, Baguley, & Abbott, 2006). This could mean that some types of subdiscipline work are more difficult than others. In regards to professional attitude, it may be that different placements require different types of professional attitude so that this too shows differences in difficulty or in the preparation of students.

Thirdly, these findings may indicate that while academic and non-academic supervisors do not differ in their expectations of students or apply different standards for grading, supervisors from different subdisciplines have different expectations and standards. Evidence that veterinarians from different disciplines value different personal characteristics tends to support this interpretation (Conlon et al., 2012). It suggests that the expectations and standards are inherent in the subdiscipline knowledge or that subdiscipline members effectively form a community of practice through which they generate a shared understanding of the standards. It also suggests that being co-located is not the most important aspect of the development of the community of practice, as placements are located all around the country. Thus, the subdiscipline community may be a professional one rather than a location-based one. Such an interpretation would be surprising in some ways and not in others. Since in-training evaluation is designed to assess competency and the standards of competency are set by the profession rather than the University, it might be expected that academic status of the supervisor is less important than their subdiscipline. On the other hand, Shay (2003) found that although examiners had a basis for understanding the expected level of performance that was particular to the field from which they came, they developed a "feel" for the standard to apply through their experience within a co-located community of practice. Thus, a difference between the marks awarded by academics and non-academics might have been expected. Further research would be needed to investigate these interpretations further. In particular, it would be of interest to characterise the competencies required of different subdisciplines and the expectations of supervisors in those subdisciplines.

Insights into the process of judgement

The findings of this research offer some insights into aspects of the process of judging veterinary competency. The conceptual framework for the process of rating, developed from review of the literature in Chapter 2, provided a useful model of the process that situates the findings in relation to each other and shows how the aspects of the process investigated in this research fit with aspects that were not investigated. This was illustrated in Figure 2.1 on page 39.

The interview phase illuminated the picture supervisors develop of the student from the combination of their observations and inferences. The picture that supervisors form of the student's performance is highly influenced by the student's engagement, and the trustworthiness they convey. This means that the student's enthusiasm for and participation in the activities of the placement, and their honesty, reliability, acceptance of responsibility, and ability to discern their own limits and act within them, are the most important elements of the picture developed. However other aspects of student performance that influence the overall picture are the student's knowledge, skills in applying knowledge, technical and animal handling skills, communication and social interaction skills, personal functioning and management of their own emotions, and their care for animals. Together the student's actions have an impact on supervisors that also influences the picture developed. These findings echo those of researchers in other disciplines who have found a similarly wide range of characteristics of students that influence supervisor's opinion of them, and a prominence of themes such as engagement (Bogo et al., 2006; Ginsburg et al., 2010; Lavine et al., 2004; Rosenbluth et al., 2014).

In developing their picture, supervisors appear to consider the stage of training of the student, and the context of the placement as discussed in the previous sections. In keeping with findings from the similar study performed by Ginsburg et al. (2010), the overall level of performance of the student influences the aspects of competency considered. What was important in excellent students differed from what was important in weak students in the Phase 1 findings. As reported by several authors, the findings from the interview phase also suggest that supervisors supplement observations with inferences and consider these in forming their opinion (Govaerts et al., 2013; Kogan et al., 2011; Mazor et al., 2008; Pulito et al.,

2006; St-Onge et al., 2016; Stroud et al., 2015). The inferences concern both unobserved aspects of performance, and reasons for student actions.

The research also gave some insights into supervisor expectations of student performance. The mark scheme and competency frameworks studied in Phase 4, were well aligned with what supervisors think is most important, and therefore may have acted as social and institutional influences informing supervisor expectations of performance. However, whether this was the case, or simply reflected a correspondence of thinking, was not examined. Some differences were highlighted, which have been discussed, and which may point to areas of other influence on supervisor expectations. As reviewed in Chapter 2, a supervisor's own experiences as a student and as a practitioner, and the performance of other students they have supervised has been found to have more influence than assessment criteria on supervisor expectations for student performance in clinical practice (Kogan et al., 2011; St-Onge et al., 2016; Yeates et al., 2013).

My research also indicated that supervisors share a view of the general aspects of performance that are important, even across disciplines and academic contexts, that may indicate a social influence of participation in a community of professionals (Shay, 2004; St-Onge et al., 2016). However, this influence may be at a subdiscipline level. The subdiscipline context may operate on either (or both) the expectations of the supervisor or the student's level of performance and therefore the picture that supervisors develop of it (Kogan et al., 2011). The potential that expectations are adjusted by supervisors according to the training stage of the student was also highlighted by the findings of the regression phase, similar to findings in other veterinary reports (Bok et al., 2013; Dawson et al., 2013; Walsh et al., 2012). Individual supervisor influences on the picture formed or on the expectations of performance were not examined in this research.

Both the qualitative and quantitative phases of this research suggested how intertwined aspects of performance were weighed and balanced against expectations to form a holistic evaluation, and showed the contribution and differential influence of the aspects considered. This extends the observations of Govaerts et al. (2013) that supervisors considered a number of separate but interlinked dimensions of performance, by showing that the dimensions operate independently to influence the overall evaluation. Insight into the success of

information translation from the holistic evaluation to item and overall scores was obtained from the quantitative phases of research. However, this research provides only a crude view of the translation process at a group level across all supervisors. Therefore, at an individual level, there may be problems in translation, as suggested by G. Regehr et al. (2011), and further research to define this is warranted.

An unexpected outcome of this research was the insight it gave into the difficulties supervisors have in making judgements. All but one supervisor made comments related to this theme without prompting, so it was clearly an important concern. It echoes that of many researchers who have reported the lack of observation of students in medical hospital-based training (Burdick & Schoffstall, 1995; Chisholm et al., 2004; Daelmans et al., 2004; Scott et al., 1993; Stillman, 1986, 1991; Stillman et al., 1990), and the challenges of deciding a grade (Berendonk et al., 2013; Hashizume et al., 2016; McQueen et al., 2016; Tweed & Ingham, 2010). These are undoubtedly areas that need further evaluation and addressing. The conceptual framework (Figure 2.1, page 39) provides a way of situating these findings in relation to the rest of the research and provides a useful organising framework for further research. One set of difficulties mentioned by supervisors can be seen to involve the process of developing a sufficiently detailed picture of the student's performance. These included aspects of insufficient observation arising from limited contact, short placements, and because of lack of engagement on the part of students. A lack of distinctiveness of individual student performances amongst the group also contributed to difficulties in forming a detailed picture. Other concerns relating to the criteria to apply, and what should count, relate to the supervisor expectations of student performance and how these are influenced and developed. Concerns about how to weight aspects of performance and trade-off performance deficits point to the difficulties of comparing the picture developed to expectations to arrive at an overall judgement. Influences on the translation of information to scores were also apparent in supervisor concerns regarding the consequences of their decision and the differing opinions of others.

Thus, the conceptual framework for the process of rater judgement, derived from the literature on rater judgement in medicine and other disciplines, has proved useful in considering the results of this research and integrating its unexpected findings. Although this research considers only very limited aspects of the framework, the results are consistent with

it, and suggest that this framework is a useful basis for research on rater judgement in veterinary medicine.

Determining the factor structure

In this final section of the discussion, I make some observations on the novel graphical method used to aid determination of the factor structure in Phase 2.

Given that the in-training evaluation is widely criticised on the basis of the number of factors, it was important to develop a well-founded conclusion about the number of factors present. However, determining the number of factors is considered a difficult step in factor analysis (Fabrigar & Wegener, 2012). A number of methods are available and their accuracy depends on characteristics of the data. These are reviewed in Appendix D (page 275). Not all are available in standard statistical software. Those that are easy to apply, such as the eigenvalue greater than one rule, are not considered accurate (Baglin, 2014; Costello & Osborne, 2005; Fabrigar et al., 1999; Gaskin & Happell, 2014; Gorsuch, 1983; Preacher & MacCallum, 2003; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). As seen in my data, different methods of analysis can produce conflicting solutions for the number of factors, making a final decision difficult.

Many general recommendations for procedures are based on findings in normally distributed linear data. The expected accuracy of methods for determining the number of factors in ordinal data is an area of more recent study and software development. There are, as yet, limited numbers of naturalistic and simulation studies that might indicate the accuracy of different methods for establishing the number of factors in ordinal data. These were reviewed (as reported in Appendix D, page 275) for information that would indicate the accuracy in data with characteristics such as mine, which combined ordinal data with a large sample size, mild-moderately skewed and kurtotic distribution, few variables per factor, strong factor loadings, and strong interfactor correlation. Since no study exactly mirrored this combination of data characteristics, the literature provided only general guidance, although some methods were clearly likely to be less accurate than others.

In this research, the results of the various procedures indicated conflicting solutions, even from methods that my review suggested should be the most accurate in this data. It seemed very subjective to settle on one or another as "the answer", as there seemed to be no factor solution that could be better justified than another. However, from a constructivist perspective, a single "right answer" need not be expected. Rather than the solution objectively representing reality, it represents the meaning we make of reality and therefore can have different meaning for different perspectives and purposes (Crotty, 1998). This is not inconsistent with a view of the factor solution as a model. As MacCallum (2003) reminds us "it is safe to say that these models all have one thing in common: *They are all wrong*" (p. 114, emphasis in original). There is no correct answer, only the best model for the purpose. From this perspective one can see that, in this research, all the models of the factor structure, with their different numbers of factors, were good summaries of the data, giving solutions that gave clear simple structures that were statistically interpretable, and that grouped items in a way that was conceptually interpretable. All were therefore ways of seeing, looking from different perspectives. The three-dimensional graph helped me to understand these different perspectives, and how the complexity of the data meant it could be interpreted in different ways.

A three-dimensional graphical representation of a factor analysis has not, to my knowledge, been previously reported, but graphs are often used during data analysis for other statistical methods as they provide insights which can illuminate findings, check assumptions, and suggest new hypotheses (Armitage, Berry, & Matthews, 2002). Graphing the correlations to visualise their clustering applies the principles of factor analysis in a graphical rather than numerical way. A group of variables that are highly correlated with each other have something in common which is the underlying factor (McDonald, 1985). In small correlation matrices (those with few variables), clusters of correlations may be directly discernible from studying the matrix (Furr & Bacharach, 2014; Gorsuch, 1983). A disadvantage of the three-dimensional analysis is that, in being a type of intuitive approach similar to those described by Gorsuch (1983), it could potentially lead to the development of more factors than necessary to adequately represent the data. Also, the method would be difficult to apply when there were large numbers of variables, because the rearranging of the matrix so that the most correlated variables were closest together would be difficult. Further, it gives no indication of the strength of relationship between the variables and the factors. It cannot, therefore, replace the statistical analysis.

A potential criticism of the three-dimensional analysis is that in rearranging the correlation matrix, the researcher may be imposing an a priori structure to the matrix, placing variables together according to hypothesised relationships. At one extreme, if the matrix is completely unsorted (that is, variables are in a random order) then the number of peaks can be as many as there are variables. The aim of sorting is to do the opposite, and to form the fewest peaks possible by bringing values with the highest correlation together and values with the lowest correlation together. However, because sorting a matrix row or column based on one correlation also changes the order of all the other correlations in the row or column, only some high correlations and some low correlations can be brought together. They remain separated no matter how the matrix is ordered. Trial and error is needed to determine how individual peaks can be combined and this is aided visually by colour coding high and low correlations with contrasting colours. The process stops when rearrangement cannot bring individual peaks any closer together. As long as sufficient time is spent exploring options for rearrangement then the process should be objective. The process should also be amenable to development of a computer programme that would determine the best order automatically.

In this research, the three-dimensional analysis was invaluable for helping understand the apparently conflicting results of the factor analysis. By showing the complex structure in an easily interpretable way it enabled me to see that, depending on the perspective taken, 1-, 2-, 3-, or 4-factor solutions were all valid ways of representing the data. It suggested that settling on any one factor solution may have been an over-simplification. It also provided a clear way to share my interpretation with readers who may be less familiar with interpreting factor analysis results. Many authors of studies on in-training evaluations have presented the analysis methods and results of factor analysis as if there was only one possible solution and determining the number of factors was unproblematic. Others have acknowledged the difficulties and explained their interpretation and the models they considered plausible, providing enough detail for readers to make their own interpretation (Cook et al., 2010; McGill et al., 2013). Although it may have been a peculiarity of this data that enabled it to offer such a clear visualisation, it would be useful to study other natural or simulated data to determine the utility of the method more generally, and define its limitations.

Summary

Seven main findings of the research have been examined in this discussion. These included insights into the nature of supervisor judgement and what is assessed, the process of judgement, and what competency means to supervisors. Some methodological considerations were also discussed. In the next—and final—chapter I will draw conclusions from the findings and discuss their implications.

Chapter 9:
Conclusion

> Of those evaluation methods currently available for the assessment of clinical competence, in-training evaluation comes closest to measuring true performance. Unfortunately, current methods are neither reliable nor valid, leaving evaluative decisions made on this basis indefensible. (Turnbull & van Barneveld, 2002, p. 793)

Despite unwavering criticism such as this, the in-training evaluation remains an indispensable part of assessment of competency in many disciplines. Rather than abandoning this form of assessment, my research sought to increase our understanding of it. The findings discussed in the previous chapter provide the basis for conclusions that inform recommendations for further research and use of the instrument.

In this chapter, the findings that address each of the research questions are first summarised, and then the conclusions reached through synthesis across the phases of research are presented. The strengths and limitations of the research as a whole are considered and future research directions identified. The implications of the research for the validity of scores and use of the in-training evaluation at Massey University, and more widely in veterinary medicine, are then explored.

Summary of the research findings

This research set out to deepen our understanding of what is actually being captured in the scores on in-training evaluation and how that aligns with our intentions. A mixed method approach entailing four phases was used to elaborate and enhance our understanding of the judgement of veterinary competency. The phases involved a qualitative interview phase, two quantitative phases which used common factor analysis and then ordinal logistic regression, and a final qualitative phase which extended the thematic framework developed from the interview phase to new data, and then integrated the themes from each phase together. Four research questions guided the research.

The first question concerned the aspects of student performance that supervisors valued. I found that there were many interlinked aspects considered, but that overall, engagement was most important. Trustworthiness was also important as a differentiator of excellent and weak students. The emphasis of what was considered was on nontechnical skills. Supervisors weighed up positive and negative aspects of performance and made inferences about the student's reasons and attitudes, and the aspects of performance they did not observe, to support their observations.

The second question concerned the nature of the dimensions captured by the in-training evaluation. The answer was not straightforward. The scores reflected the multidimensional higher-order structure the instrument was designed to assess, but also reflected one or two strong dimensions. I concluded that the instrument captures both the separate dimensions it was intended to assess and an overarching holistic construct, but that other interpretations were also possible and might suit different purposes.

The third question asked how the dimensions influenced the overall grade supervisors awarded. What I found was that all the dimensions influenced the overall grade to different degrees, but the professional attitude factor had, by far, the greatest influence on the overall grade, both positively and negatively. Placement also had a strong influence on overall grade but GPA, academic status of the supervisor, and time, did not.

The last question concerned the alignment between what supervisors considered important, what their scores revealed had influenced the grade awarded, and what they were asked to assess. I found that, overall, the alignment was good and that what was considered most important to supervisors was what influenced the overall grade the most. I concluded that this meant their evaluation was generally well aligned with what they were asked to assess and was transferred well to scores.

In investigating the specific research questions, the research also revealed other findings. These included the difficult and troubling nature of the judgement process for supervisors, and that supervisors considered aspects of the student's personal functioning, and their prospects, as well as their impact on the supervisor and other staff in making their evaluation. These are aspects not represented in the mark scheme or competency frameworks.

Research contribution

This research provides evidence that supervisor judgement involves consideration of multiple aspects of performance and is at the same time holistic and discriminating. Aspects of performance show considerable interlinkage but are able to act independently to influence the overall evaluation. The judgement process involves weighing and balancing aspects in consideration of the whole picture. Both the holistic and discriminating nature is successfully translated to item scores giving a multidimensional higher-order structure to the evaluation. This appropriately reflects the multidimensional, holistic, and unitary conception of competency, and should not be entirely attributed to halo error.

This research also suggests that student engagement and trustworthiness are the most important aspects that supervisors consider in making their evaluation, and have the most influence on the overall grade awarded. Since these are also the most important aspects in the mark scheme it demonstrates that, in respect of these aspects, the in-training evaluation assesses what it was designed to assess. Engagement and trustworthiness are also aspects of importance to the profession and may determine success in practice. Therefore, there is some evidence that in-training evaluation scores can be credibly extrapolated to indicate success in practice after graduation.

This research identifies that, while generally well aligned with mark schemes and competency frameworks, the aspects that inform supervisor judgement differ somewhat in emphasis by focusing most on nontechnical aspects of performance and include aspects which are not present in the mark scheme or competency frameworks. Also supervisors may not apply a fixed standard, but vary it according to the stage of training of the student. These findings raise questions for the profession about the nature of veterinary competency and whether aspects considered by supervisors, such as personal functioning, are rightly part of the construct. They also raise questions about the future design of instruments and whether aspects such as a student's impact and prospects should form part of the evaluation.

This research indicated that the veterinary subdiscipline influenced the overall grade awarded and that the effect of knowledge and professional attitude on determining overall grade varied according to the subdiscipline. This means that there was a different emphasis, particularly on these two dimensions of performance, across different subdisciplines. The effect did not

depend on whether the placement was at the University or elsewhere, indicating that the nature of veterinary work, rather than the academic environment was the greatest influence. The findings indicate that student performance or supervisor expectations, or both, differ across the subdisciplines and not just in terms of the discipline-specific knowledge and skills. It may be that the idea of one unitary competency across all the disparate subdisciplines of veterinary medicine is inappropriate. Students, and the profession, may be better served by characterising veterinary competency according to subdisciplines. This may facilitate consideration of the best way to manage subdiscipline specialisation within the profession and within curricula.

Arising from this research is a conceptual framework for the process of judgement (Figure 2.1, page 39) that usefully informed this research. The framework conceptualises the judgement process as involving the formation of a picture of student performance, informed by observation and inference, which is compared to expectations to form a holistic narrative judgement. This judgement is then translated to an overall score and scores for the various aspects of performance. The judgement process is influenced by a variety of factors at all levels, including personal, social, institutional, and contextual factors. Although a gross simplification of a complex process that is not yet fully understood, this framework provided a useful way of thinking about supervisor judgement during this research, and could inform future research, for example by locating the difficulties supervisors face within a framework for investigation and improvement.

Finally, the research has contributed another perspective on exploratory common factor analysis and a method for visualising the complexity of the data. It reminds us that determination of the number of factors needs to be an interpretive judgement that considers the purposes of the factor analysis, the theoretical expectations, and the “shape” of the data. More than one solution may represent the data well and it may be an oversimplification to consider only one solution as “the answer”. The three-dimensional analysis of the data proved to be a good way to see the relationships in the correlation matrix and may prove useful in other datasets.

Limitations of the research

Some limitations particular to each phase of research have been discussed in the results chapters so that their implications could be considered in drawing conclusions. These included the small sample size for interviews, the difference in the placements sampled in the interviews and quantitative phases, and that the findings do not characterise the decisions of individual supervisors, but the group of supervisors as a whole. In reflecting on the body of work overall, there are other limitations that also need consideration.

One limitation is that this research provides only a glimpse of the complexity of rater judgement processes during veterinary in-training evaluation. As discussed in Chapter 2, the conceptual map of the process of judgement involves a series of steps and multiple influences. The research presented is such a small part that it is bound to be incomplete. It serves as a useful first step in what needs to be a body of work in order to understand how rater judgement processes operate in the context of assessment of veterinary competency.

Secondly, the glimpse is seen through the lens of my interpretations. While I have interpreted the results with reference to a theoretical framework built on previous research, and followed a careful and systematic pathway, accounting for as many sources of bias as possible, it is still only one possible interpretation of the data. The real test of the validity of the conclusions I have drawn will be in the resonance they have with readers and with other future research. I hope that my lens, coming from within the discipline, gives a view that makes sense to our profession.

Thirdly, the glimpse provided by this research reflects only one point in time, a time that has now been and gone. The in-training evaluation instrument is constantly being re-evaluated and changed as teachers recognise issues and seek to improve the way it works. I hope that the meaning derived from this research gives us a theoretical basis from which to go forward with these revisions at Massey University and conduct further research.

Along those same lines, this research is a glimpse of the instrument at one veterinary school. While the instrument has many similarities to the instruments used for in-training evaluation at other veterinary schools, it was purpose designed for use here and therefore will differ from

others in use. The contexts within which the supervisors in this research work will be different in some ways to those supervisors at other veterinary schools. Therefore, the findings are not necessarily generalisable. However, the great similarity in the way veterinary education is practised across veterinary schools within many regions of the world, as well as intermixing of supervisors across nations, means that some of the findings may be applicable elsewhere. I have provided detailed descriptions that will facilitate readers determining the applicability of the findings to their own situation.

Future research directions

As an initial study into the in-training evaluation at one institution, the findings of this research suggest areas that can be built on to enhance our understanding of competency and in-training evaluation. I have commented on some of these throughout the thesis, and now present some additional directions for further research.

Firstly, further insights into the process of judging student performance might be gained from examining the written feedback supervisors provide on in-training evaluations. As a window into the narrative evaluation it might further our understanding of the pictures supervisors form and their holistic evaluation. Such work has given interesting insights into supervisor thinking in medical education (Ginsburg et al., 2011). It may be possible to use themes arising from feedback comments to more clearly characterise how rater judgements are translated to scores at an individual rater level if databases that contain both are used for analysis.

Secondly, the student perspective is entirely lacking from this body of research. How the evaluation process impacts students and influences their feelings, actions, and performance would help us to further understand what the in-training evaluation is capturing and how relevant that is to future performance in the workplace. It would also help us to determine how best to utilise evaluations for formative purposes. Some studies have indicated significant issues with student perceptions of the validity of in-training evaluation scores in veterinary education that limit their formative use and warrant further investigation (Dawson et al., 2013; Weijs et al., 2016).

Leading on from this, the finding of the key importance of engagement in contributing to evaluations of performance suggests a third area of research. Based on work with medical students, Dornan et al. (2007) found a pivotal influence of supervisors and other staff on the engagement or disengagement of students, as well as roles for contextual factors such as the curriculum and staff to student ratio, student factors such as confidence and practical skills. Investigating whether and how these and other aspects contribute to veterinary student engagement would support curriculum development and optimal functioning of the veterinary teaching environment for development of competency in our students.

Fourthly, research is needed that furthers our understanding of how veterinary supervisors come to understand the meaning of competency themselves and develop expectations of student performance. The role of professional and institutional factors and the potential for rater training building on the theoretical framework established in this research would be valuable areas of research applicable to other types of competency assessment.

Lastly, the construction of an argument for the validity of scores awarded requires gathering a body of evidence in support of all the inferences involved in interpreting the score (Cook, Brydges, Ginsburg, & Hatala, 2015). Very little of this has been done for any veterinary assessment instruments, and more of this work is needed to support the inferences derived from scores as I discuss in the next section.

Implications and recommendations

The conclusions drawn from this research have implications for the validity of interpretations of in-training evaluation scores at Massey University, and enable some recommendations to be made that may inform future development and use of the instrument. Many of the recommendations will have relevance for similar in-training evaluations in other programmes.

Implications for validity of interpretations of in-training evaluation scores

The validity of in-training evaluation scores is often criticised because of issues with interrater reliability, including leniency and subjectivity. A lack of discrimination of dimensions of performance and a lack of relationship with other measures of performance are also issues raised. These aspects are of concern because they suggest that conclusions we may draw from the scores are not credible. They lead to uncertainty about what is being assessed and whether the appropriate things are being assessed each time.

This research did not consider rater subjectivity or leniency. However, it gave some insight into the dimensions being assessed by the instrument in use at Massey University, and the relationship between the aspects being assessed and the constructs the evaluation is designed to assess. The findings suggest that the dimensionality is complex, but that this complexity appropriately reflects the complexity of the construct. The findings also suggest alignment between supervisors' judgements and the scoring criteria, both in conception and as enacted. This, however, is on a global level across the year for all students, rather than at an individual evaluation level, which the research did not investigate. The findings also suggest that the constructs assessed by the in-training evaluation involve more than knowledge, and that therefore knowledge-based examinations would not be expected to be closely related to performance on the in-training evaluation. Therefore, the in-training evaluation at Massey University should not be criticised because of low correlation of scores with knowledge-based examinations.

These findings do not complete the validity argument. Constructing a validity argument for a high stakes assessment requires many different types of study (Cook et al., 2015). Since the validity of scores depends on the entire chain of validity evidence (Kane, 2013), conclusions cannot be drawn at this point. However, the research gives some insight that suggests that not all aspects of the validity of in-training evaluation scores warrant the criticisms they have been subject to. This therefore suggests that more work on addressing issues that could be improved, such as the number of evaluations made and the delay in recording of scores, would be worthwhile pursuing.

Recommendations for future development and use of in-training evaluation

The findings of this research provide a more informed theoretical perspective from which to consider improvements to in-training evaluation in veterinary medicine at Massey University. Firstly, they provide insight into the current alignment of the instrument with what supervisors think is important and how they assign scores. Crossley and Jolly (2012) have previously discussed the importance of aligning instruments with the way raters think. Considerable gains in score reliability and reductions in leniency have been demonstrated from thoughtful instrument redesign that aligned scales with supervisors' judgements of the trustworthiness of medical students (Crossley et al., 2011; Weller et al., 2014). The finding that supervisors in this research referred to elements of trustworthiness, and that it is was an important component of their overall judgement, suggests that similarly aligning performance benchmarks with levels of independence and trustworthiness, or basing assessment around entrustable professional activities (ten Cate & Scheele, 2007) might also be viable for veterinary medicine. However Dawson et al. (2013) found significant issues when implementing an independence-based rating scale in an in-training evaluation in veterinary medicine. Supervisors found it difficult to account for the stage of training in grading students. If students were beginners, assigning low scores because they were not yet independent caused supervisors concerns about affecting student self-esteem and evaluations of their teaching. This demonstrates the importance of choosing activities for assessment using an entrustment scale that students could reasonably be expected to be at or nearing competency in, at the time they are assessed, especially if the assessment is used summatively.

Another way of improving alignment between rater thinking and the instrument would be to consider the content of the items and dimensions supervisors are asked to assess. Working in the field of social work, C. Regehr, Bogo, and Regehr (2011) were able to improve the ability of in-training evaluations to discriminate levels of student performance by redesigning items to accommodate the dimensions and language supervisors used in describing excellent and weak social work students and eliminating numerical scales. Similar work for the veterinary in-training evaluation at Massey University might reduce the mixing of concepts within an item and the overlap in concepts across items and thereby reduce rater subjectivity arising from distributing an overall judgement amongst items that don't quite fit. However, the findings of this research also signal a warning that simply adopting the thematic framework arising from

supervisors' depictions as a new assessment instrument is unlikely to be successful without further research, because even the themes themselves were not distinct and easily separable. Furthermore, the findings of Sebok and Syer (2015), in demonstrating both interrater and intercontextual differences in the way raters link ideas during assessment, suggest that themes that seem distinct in one context or for one supervisor may not be for another context or another supervisor. Thus, it may be impossible for an instrument to adequately reflect the interlinked holistic weighting of aspects of a whole performance within the individual item scores. This may need to be left to the overall judgement of the supervisor.

Following on from this, this research indicates that the interrelatedness of dimensions has a basis in the construct being assessed and thus, though it may obscure the discrimination of strengths and weaknesses for the purposes of formative assessment, is not entirely error. It is also unlikely to be able to be completely removed without distorting the assessment such that it no longer evaluates the holistic and unitary concept of competency taking into account the context of the situation. Thus, it may be best to focus efforts on other ways of providing good formative feedback to students. The studies by Bok et al. (2013) demonstrate the difficulties in combining formative and summative assessment, in that students take a different approach to even low-weighted summative assessments. We should seek additional ways to provide formative feedback, as Bok et al. (2015) suggest, rather than relying solely on in-training evaluations to provide all of the formative feedback students need.

Another finding of this research that informs in-training evaluation development was the differential weighting supervisors gave to aspects of performance depending on the level of the student, consistent with findings in medical education (Ginsburg et al., 2010). This implies that the supervisor's freedom to weight the aspects considered in determining the overall grade is important in enabling the score to best represent their evaluation. Applying fixed weighting to various dimensions or simply averaging item or dimension scores would not enable the considerations that supervisors make to be reflected in the overall score and therefore should be avoided.

The findings from this research also provide some perspectives on how difficult supervisors find the process of making a judgement to be. Many of these suggest areas that could be targets for improvement to administrative processes to facilitate evaluation. Littlefield et al.

(2005) demonstrated the substantial gains in reliability of scores, specific feedback provided, and prompt return of evaluations achievable through changes in the way in-training evaluations were administered and managed. The changes included improvements in data storage and systems for distribution and collection of evaluation forms; monitoring and reminder systems for form completion with reporting of compliance; training supervisors in the use of systems and forms; provision of good feedback to supervisors; and sharing of good and poor evaluation practice. Facilitating notetaking and ensuring workloads enable plenty of time for direct observation of student performance are also areas likely to improve the quality of evaluations. Supervisor training, mentorship, and any moves that encourage the sharing of practice by supervisors with each other are likely to help develop and maintain shared understanding of the assessment criteria. Even simple reassurance of the value of the evaluation for both students and the profession, based on evidence such as provided by this research, is likely also to help supervisors feel more confident in the difficult task of evaluation.

Finally, the findings of this research suggest that the in-training evaluation at Massey University is assessing aspects of performance that may not be well assessed in other types of assessment. This means that not only, as previously discussed, would it be inappropriate to validate scores using another knowledge-based assessment, but that if the in-training evaluation is to be replaced with other assessments, careful consideration needs to be given to what is assessed in those other instruments. In both medical and veterinary education there is increased use of direct observation tools for workplace-based assessment. These have obvious advantages in being authentic, providing immediate formative feedback and contemporaneous recording of the evaluation. However, research on score validity in relation to rater judgement processes is still lacking from direct observational tools (Kogan et al., 2009; Sandilands & Zumbo, 2014) and further work is needed on these assessments.

While these recommendations are made in respect of the in-training evaluation instrument used at Massey University, to the extent that in-training evaluation instruments are similar, and the learning environment and supervisor values are consistent with those discussed here, they may also be applicable to other veterinary schools.

Final thoughts

> By far the most maligned and scrutinized method of evaluating residents' clinical skills has been the ward [in-training] evaluation. (Reddy & Vijayakumar, 2000, p. 4)

A perfect assessment instrument has not yet been devised and likely never will be. Although there was a time when the aim was to specify the assessment criteria so clearly that even a computer or an unqualified observer could reliably score performance, it is now more widely recognised that judging competency is a complex task requiring the judgement of expert raters. An exciting wave of new research has its basis in a more open view, embracing subjectivity and seeking to better understand the process of judgement and the influences on it. I hope that this research adds to that body of research by illuminating aspects of the judgement of competency in a veterinary context. The in-training evaluation will never be perfect, but understanding its limitations and strengths will allow its use to be tailored in a programme of assessment that, as a whole, provides valid inferences on the performance of veterinary students.

References

Abbott, K. A. (2009). Innovations in veterinary education: The Charles Sturt University programme (Wagga Wagga, Australia). *OIE Revue Scientifique et Technique, 28*(2), 763-770. doi:10.20506/rst.28.2.1912

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ, USA: Wiley.

Albanese, M. A. (2001). Challenges in using rater judgements in medical education. *Journal of Evaluation in Clinical Practice, 6*(3), 305-319. doi:10.1046/j.1365-2753.2000.00253.x

Allan, F., & Parkinson, T. (2010). Regional branch roadshow: Curriculum change and the needs of the profession in 2020. *Vetscript, 23*(10), 4-8.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research, 28*(3), 301-309. doi:10.1177/0049124100028003003

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557. doi:10.1037/0021-843x.112.4.545

Allison, P. D. (2008, March). *Convergence failures in logistic regression*. Paper presented at the SAS Global Forum, San Antonio, TX, USA. Retrieved from http://www2.sas.com/proceedings/forum2008/360-2008.pdf

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 72-89). Thousand Oaks, CA, USA: Sage.

Allison, P. D. (2012, April). *Handling missing data by maximum likelihood*. Paper presented at the SAS Global Forum, Orlando, FL, USA. Retrieved from http://support.sas.com/resources/papers/proceedings12/312-2012.pdf

Alves de Lima, A. (2013). *Assessment of clinical competence: Reliability, validity, feasibility and educational impact of the mini-CEX.* (Doctoral thesis, Maastricht University, Netherlands). Retrieved from http://www.icba.com.ar/profesionales/pdf/aal/Thesis_Alberto_Alves_de_Lima_170x240_v10.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC, USA: American Educational Research Association.

Armitage, P., Berry, G., & Matthews, J. N. S. (2002). *Statistical methods in medical research* (4th ed.). Malden, MA, USA: Blackwell Science.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397-438. doi:10.1080/10705510903008204

Association of American Medical Colleges. (2015). *Medical school graduation questionnaire. 2015 all schools summary report*. Washington, DC, USA: Author. Retrieved from https://www.aamc.org/download/440552/data/2015gqallschoolssummaryreport.pdf

Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. *Medical Education, 39*(3), 276-283. doi:10.1111/j.1365-2929.2005.02090.x

Babakus, E., Ferguson, C. E., Jr., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*(2), 222-228. doi:10.2307/3151512

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: A demonstration using FACTOR. *Practical Assessment, Research and Evaluation, 19*(5), 1-14. Retrieved from http://pareonline.net/getvn.asp?v=19&n=5

Baker, K. (2011). Determining resident clinical performance. Getting beyond the noise. *Anesthesiology, 115*(4), 862-878. doi:10.1097/ALN.0b013e318229a27d

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975-985. doi:10.1037//0021-9010.77.6.975

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis. Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). New York, NY, USA: Routledge.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2014). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling, 22*(1), 87-101. doi:10.1080/10705511.2014.934850

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data* (2nd ed.). Boca Raton, FL, USA: CRC Press.

Bartlett, J. W., Carpenter, J. R., Tilling, K., & Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics, 15*(4), 719-730. doi:10.1093/biostatistics/kxu023

Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software, 46*(4), 1-29. doi:10.18637/jss.v046.i04

Bateman, K., Menzies, P., Sandals, D., Duffield, T., LeBlanc, S., Leslie, K., . . . Swackhammer, R. (2008). Objective structured clinical examinations (OSCEs) as a summative evaluation tool in a ruminant health management rotation for final-year DVM students. *Journal of Veterinary Medical Education, 35*(3), 382-388. doi:10.3138/jvme.35.3.382

Beauducel, A. (2005). How to describe the difference between factors and corresponding factor-score estimates. *Methodology, 1*(4), 143-158. doi:10.1027/1614-1881.1.4.143

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation, 18*(6), 1-13. Retrieved from http://pareonline.net/getvn.asp?v=18&n=6

Bell, M. L., & Fairclough, D. L. (2014). Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Statistical Methods in Medical Research, 23*(5), 440-459. doi:10.1177/0962280213476378

Benyamini, K., Kedar, H. S., & Raveh, I. (1987). How do supervising doctors construe the medical student in clinical training? *Medical Education, 21*(5), 410-418. doi:10.1111/j.1365-2923.1987.tb00389.x

Berendonk, C., Stalmeijer, R., & Schuwirth, L. T. (2013). Expertise in performance assessment: Assessors' perspectives. *Advances in Health Sciences Education, 18*(4), 559-571. doi:10.1007/s10459-012-9392-x

Biesta, G. (2010). Pragmatism and the philosophical foundations of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (2nd ed., pp. 95-117). Thousand Oaks, CA, USA: Sage.

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods, 7*(4), 400-417. doi:10.1177/1094428104268542

Boerebach, B. C., Arah, O. A., Heineman, M. J., & Lombarts, K. M. (2016). Embracing the complexity of valid assessments of clinicians' performance: A call for in-depth examination of methodological and statistical contexts that affect the measurement of change. *Academic Medicine, 91*(2), 215-220. doi:10.1097/ACM.0000000000000840

Bogo, M., Hughes, J., Regehr, C., Power, R., Woodford, M., & Regehr, G. (2006). Beyond competencies: Field instructors' descriptions of student performance. *Journal of Social Work Education, 42*(3), 579-593. doi:10.5175/JSWE.2006.200404145

Bogo, M., Regehr, C., Hughes, J., Power, R., & Gioberman, J. (2002). Evaluating a measure of student field performance in direct service: Testing reliability and validity of explicit criteria. *Journal of Social Work Education, 38*(3), 385-401. doi:10.5175/jswe.2006.200404145

Bok, H. G. J., Jaarsma, D. A. D. C., Spruijt, A., van Beukelen, P., van Der Vleuten, C. P. M., & Teunissen, P. W. (2015). Feedback-giving behaviour in performance evaluations during clinical clerkships. *Medical Teacher, 38*(1), 88-95. doi:10.3109/0142159X.2015.1017448

Bok, H. G. J., Jaarsma, D. A. D. C., Teunissen, P. W., van der Vleuten, C. P. M., & van Beukelen, P. (2011). Development and validation of a competency framework for veterinarians. *Journal of Veterinary Medical Education, 38*(3), 262-269. doi:10.3138/jvme.38.3.262

Bok, H. G. J., Teunissen, P. W., Boerboom, T. B. B., Rhind, S. M., Baillie, S., Tegzes, J., . . . van Beukelen, P. (2014). International survey of veterinarians to assess the importance of competencies in professional practice and education. *Journal of the American Veterinary Medical Association, 245*(8), 906-913. doi:10.2460/javma.245.8.906

Bok, H. G. J., Teunissen, P. W. F., Robert P., Rietbroek, N. J., Theyse, L. F. H., Brommer, H., Haarhuis, J. C. M., . . . Jaarsma, D. A. D. C. (2013). Programmatic assessment of competency-based workplace learning: When theory meets practice. *BMC Medical Education, 13*, 123. doi:10.1186/1472-6920-13-123

Borman, K. R., Augustine, R., Leibrandt, T., Pezzi, C. M., & Kukora, J. S. (2013). Initial performance of a modified milestones global evaluation tool for semiannual evaluation of residents by faculty. *Journal of Surgical Education, 70*(6), 739-749. doi:10.1016/j.jsurg.2013.08.004

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219. doi:10.1037/0033-295x.110.2.203

Brasel, K. J., Bragg, D., Simpson, D. E., & Weigelt, J. A. (2004). Meeting the Accreditation Council for Graduate Medical Education competencies using established residency training program assessment tools. *The American Journal of Surgery, 188*(1), 9-12. doi:10.1016/j.amjsurg.2003.11.036

Brennan, R. L. (2005). (Mis) conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5-10. doi:10.1111/j.1745-3992.2000.tb00017.x

Brooks, M. A. (2009). Medical education and the tyranny of competency. *Perspectives in biology and medicine, 52*(1), 90-102. doi:10.1353/pbm.0.0068

Brown, J. P., & Silverman, J. D. (1999). The current and future market for veterinarians and veterinary medical services in the United States - executive summary. *Journal of the American Veterinary Medical Association, 215*(2), 161-183.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY, USA: Guilford Press.

Browne, M. W. (1992). FITMOD [Computer software]. Columbus, OH, USA: The Ohio State University. Retrieved from http://faculty.psy.ohio-state.edu/browne/software.php

Buchtel, E. E., & Norenzayan, A. (2009). Thinking across cultures: Implications for dual processes. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 217–238). Oxford, England: Oxford University Press.

Burdick, W. P., & Schoffstall, J. (1995). Observation of emergency medicine residents at the bedside: How often does it happen? *Academic Emergency Medicine, 2*(10), 909-913. doi:10.1111/j.1553-2712.1995.tb03108.x

Byrne, A., Tweed, N., & Halligan, C. (2014). A pilot study of the mental workload of objective structured clinical examination examiners. *Medical Education, 48*(3), 262-267. doi:10.1111/medu.12387

Cake, M. A., Bell, M. A., Williams, J., Brown, F., Dozier, M., Rhind, S. M., & Baillie, S. (2016). Which professional (non-technical) competencies are most important to the success of graduate veterinarians? A best evidence medical education (BEME) systematic review: BEME guide no. 38. *Medical Teacher, 38*(6), 550-563. doi:10.3109/0142159X.2016.1173662

Cake, M. A., Rhind, S. M., & Baillie, S. (2014). The need for business skills in veterinary education: Perceptions versus evidence. In C. Henry (Ed.), *Veterinary business and enterprise: Theoretical foundations and practical cases* (pp. 9-22). Edinburgh, UK: Saunders Elsevier.

Campbell, C., Crebbin, W., Hickey, K., Stokes, M.-L., & Watters, D. (2014). Work-based assessment: A practical guide. Building an assessment system around work. Tri-Partite Alliance (Royal College of Physicians and Surgeons of Canada, Royal Australasian College of Physicians, Royal Australasian College of Surgeons). Retrieved from http://www.surgeons.org/media/20786937/bkt_tripartite_wba__march_7__2_.pdf

Campion, M. C., Campion, E. D., & Campion, M. A. (2015). Improvements in performance management through the use of 360 feedback. *Industrial and Organizational Psychology, 8*(01), 85-93. doi:10.1017/iop.2015.3

Carraccio, C. L., Benson, B. J., Nixon, J., & Derstine, P. L. (2008). From the educational bench to the clinical bedside: Translating the Dreyfus developmental model to the learning of clinical skills. *Academic Medicine, 83*(8), 761-767. doi:10.1097/ACM.0b013e31817eb632

Carraccio, C. L., Wolfsthal, S. D., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting paradigms: From Flexner to competencies. *Academic Medicine, 77*(5), 361-367. doi:10.1097/00001888-200205000-00003

Case, D. (1988). Survey of expectations among clients of three small animal clinics. *Journal of the American Veterinary Medical Association, 192*(4), 498-502.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. doi:10.1207/s15327906mbr0102_10

Chahine, S., Holmes, B., & Kowalewski, Z. (2016). In the minds of OSCE examiners: Uncovering hidden assumptions. *Advances in Health Sciences Education, 21*(3), 609-625. doi:10.1007/s10459-015-9655-4

Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher, 21*(6), 13-17. doi:10.2307/1176502

Chisholm, C. D., Whenmouth, L. F., Daly, E. A., Cordell, W. H., Giles, B. K., & Brizendine, E. J. (2004). An evaluation of emergency medicine resident interaction time with faculty in different teaching venues. *Academic Emergency Medicine, 11*(2), 149-155. doi:10.1197/j.aem.2003.08.017

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement, 69*(5), 748-759. doi:10.1177/0013164409332229

Choo, K. J., Arora, V. M., Barach, P., Johnson, J. K., & Farnan, J. M. (2014). How do supervising physicians decide to entrust residents with unsupervised tasks? A qualitative analysis. *Journal of Hospital Medicine, 9*(3), 169-175. doi:10.1002/jhm.2150

Cleland, J. A., Knight, L. V., Rees, C. E., Tracey, S., & Bond, C. M. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education, 42*(8), 800-809. doi:10.1111/j.1365-2923.2008.03113.x

Cohen, G. S., Blumberg, P., Ryan, N. C., & Sullivan, P. L. (1993). Do final grades reflect written qualitative evaluations of student performance? *Teaching and Learning in Medicine, 5*(1), 10-15. doi:10.1080/10401339309539580

Coleman, G. T., Salter, L. K., & Thornton, J. R. (2000). What skills should veterinarians possess on graduation? *Australian Veterinary Practitioner, 30*(3), 124-131.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi:10.1037/1082-989X.6.4.330

Competence. (2012). In *OED Online*. Retrieved from http://oed.com/view/Entry/37567?redirectedFrom=competence&

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46*(4), 648-659. doi:10.1037/0022-006x.46.4.648

Conlon, P., Hecker, K., & Sabatini, S. (2012). What should we be selecting for? A systematic approach for determining which personal characteristics to assess for during admissions. *BMC Medical Education, 12*(1), 105. doi:10.1186/1472-6920-12-105

Cook, D. A., Beckman, T. J., Mandrekar, J. N., & Pankratz, V. S. (2010). Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. *Advances in Health Sciences Education, 15*(5), 633-645. doi:10.1007/s10459-010-9224-9

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*(6), 560-575. doi:10.1111/medu.12678

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine, 24*(1), 74-79. doi:10.1007/s11606-008-0842-3

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218-244. doi:10.1037/0033-2909.90.2.218

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation, 10*(7), 1-9. Retrieved from http://pareonline.net/getvn.asp?v=10&n=7

Craven, J. (2004). Review of veterinary science education and registration requirements. Melbourne, Australia: Australasian Veterinary Boards Council Inc. Retrieved from http://www.ava.com.au/sites/default/files/documents/Other/Craven%20ReviewVetSciEducation.pdf

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement, 70*(6), 885-901. doi:10.1177/0013164410379332

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Edinburgh Gate, UK: Pearson.

Cron, W. L., Slocum, J. V., Goodnight, D. B., & Volk, J. O. (2000). Executive summary of the Brakke management and behavior study. *Journal of the American Veterinary Medical Association, 217*(3), 332-338. doi:10.2460/javma.2000.217.332

Crossley, J., Johnson, G., Booth, J., & Wade, W. (2011). Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Medical Education, 45*(6), 560-569. doi:10.1111/j.1365-2923.2010.03913.x

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. *Medical Education, 46*(1), 28-37. doi:10.1111/j.1365-2923.2011.04166.x

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. St Leonards, Australia: Allen & Unwin.

Cruess, S. R., & Cruess, R. L. (2005). The medical profession and self-regulation: A current challenge. *Virtual Mentor, 7*(4). doi:10.1001/virtualmentor.2005.7.4.oped1-0504

Daelmans, H. E. M., Hoogenboom, R. J. I., Donker, A. J. M., Scherpbier, A. J. J. A., Stehouwer, C. D. A., & van der Vleuten, C. P. M. (2004). Effectiveness of clinical rotations as a learning environment for achieving competences. *Medical Teacher, 26*(4), 305-312. doi:10.1080/01421590410001683195

Daelmans, H. E. M., van der Hem-Stokroos, H. H., Hoogenboom, R. J. I., Scherpbier, A. J. J. A., Stehouwer, C. D. A., & van der Vleuten, C. P. M. (2005). Global clinical performance rating, reliability and validity in an undergraduate clerkship. *Netherlands Journal of Medicine, 63*(7), 279-284. Retrieved from http://www.njmonline.nl/article.php?a=332&d=212&i=71

Davis, J. K., Inamdar, S., & Stone, R. K. (1986). Interrater agreement and predictive validity of faculty ratings of pediatric residents. *Journal of Medical Education, 61*(11), 901-905. doi:10.1097/00001888-198611000-00006

Dawson, S. D., Miller, T., Goddard, S. F., & Miller, L. M. (2013). Impact of outcome-based assessment on student learning and faculty instructional practices. *Journal of Veterinary Medical Education, 40*(2), 128-138. doi:10.3138/jvme.1112-100R

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14-24. doi:10.3102/0013189x 027002014

Dellinger, A. B., & Leech, N. L. (2007). Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research, 1*(4), 309-332. doi:10.1177/1558689807306147

Demirtas, H. (2004). Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods, 3*(2), Article 5. Retrieved from http://digitalcommons.wayne.edu/jmasm/ vol3/iss2/5

DeNisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology, 81*(6), 717-737. doi:10.1037/0021-9010.81.6.717

Dielman, T. E., Hull, A. L., & Davis, W. K. (1980). Psychometric properties of clinical performance ratings. *Evaluation and the Health Professions, 3*(1), 103-117. doi:10.1177/ 016327878000300106

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ, USA: Wiley.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research and Evaluation, 14*(20), 1-11. Retrieved from http://pareonline.net/getvn.asp?v=14&n=20

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 222. doi:10.1186/2193-1801-2-222

Donnon, T., Al Ansari, A., Al Alawi, S., & Violato, C. (2014). The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review. *Academic Medicine, 89*(3), 511-516. doi:10.1097/ACM.0000000000000147

Dornan, T., Boshuizen, H., King, N., & Scherpbier, A. (2007). Experience-based learning: A model linking the processes and outcomes of medical students' workplace learning. *Medical Education, 41*(1), 84-91. doi:10.1111/j.1365-2929.2006.02652.x

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology and Society, 24*(3), 177-181. doi:10.1177/0270467604264992

Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: The perspectives of clinical supervisors. *Academic Medicine, 80*(10), S84-S87. doi:10.1097/00001888-200510001-00023

Durning, S. J., Artino, A. R., Boulet, J. R., Dorrance, K., van der Vleuten, C., & Schuwirth, L. (2012). The impact of selected contextual factors on experts' clinical reasoning performance (does context impact clinical reasoning performance in experts?). *Advances in Health Sciences Education, 17*(1), 65-79. doi:10.1007/s10459-011-9294-3

Durning, S. J., Artino, A. R., Jr., Pangaro, L., van der Vleuten, C. P. M., & Schuwirth, L. (2011). Context and clinical reasoning: Understanding the perspective of the expert's voice. *Medical Education, 45*(9), 927-938. doi:10.1111/j.1365-2923.2011.04053.x

Durning, S. J., Hanson, J., Gilliland, W., McManigle, J. M., Waechter, D., & Pangaro, L. N. (2010). Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. *Military Medicine, 175*(6), 448-452. doi:10.7205/milmed-d-09-00044

Durning, S. J., Pangaro, L. N., Lawrence, L. L., Waechter, D., McManigle, J., & Jackson, J. L. (2005). The feasibility, reliability, and validity of a program director's (supervisor's) evaluation form for medical school graduates. *Academic Medicine, 80*(10), 964-968. doi:10.1097/00001888-200510000-00018

Edwards, A., & Knight, P. (1995). The assessment of competence in higher education. In A. Edwards & P. Knight (Eds.), *Assessing competence in higher education* (pp. 10-24). London, UK: Kogan Page.

Elwood, J. M. (2007). *Critical appraisal of epidemiological studies and clinical trials* (3rd ed.). Oxford, UK: Oxford University Press.

Englander, R., Cameron, T., Ballard, A. J., Dodge, J., Bull, J., & Aschenbrener, C. A. (2013). Toward a common taxonomy of competency domains for the health professions and competencies for physicians. *Academic Medicine, 88*(8), 1088-1094. doi:10.1097/ACM.0b013e31829a3b2b

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387-396. doi:10.1056/NEJMra054784

Essers, G., Dielissen, P., van Weel, C., van der Vleuten, C., van Dulmen, S., & Kramer, A. (2015). How do trained raters take context factors into account when assessing GP trainee communication performance? An exploratory, qualitative study. *Advances in Health Sciences Education, 20*(1), 131-147. doi:10.1007/s10459-014-9511-y

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 Education and Counseling, 93*(3), 567-572. doi:10.1016/j.pec.2013.08.024

Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E., & Regehr, G. (2015). Towards a program of assessment for health professionals: From training into practice. *Advances in Health Sciences Education*, Advance online publication. doi:10.1007/s10459-015-9653-6

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278. doi:10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*(2-3), 86-102. doi:10.1016/j.dr.2011.07.007

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223-241. doi:10.1177/1745691612460685

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford, UK: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299. doi:10.1037/1082-989X.4.3.272

Faculty of Veterinary Science University of Sydney. (2004). Extramural supervisor survey. *Partners in Veterinary Education Newsletter,* (2), 3-4. Retrieved from http://sydney.edu.au/vetscience/partners/educational/newsletters/Oct2004.pdf

Ferguson, K. J., & Kreiter, C. D. (2004). Using a longitudinal database to assess the validity of preceptors' ratings of clerkship performance. *Advances in Health Sciences Education, 9*(1), 39-46. doi:10.1023/B:AHSE.0000012215.41645.7b

Fernandez, N., Dory, V., Ste-Marie, L. G., Chaput, M., Charlin, B., & Boucher, A. (2012). Varying conceptions of competence: An analysis of how health sciences educators define competence. *Medical Education, 46*(4), 357-365. doi:10.1111/j.1365-2923.2011.04183.x

Finch, W. H. (2013). Exploratory factor analysis. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 167-186). Rotterdam, Netherlands: Sense.

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 97-131). New York, NY, USA: Springer.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling. A Second Course.* (pp. 269-314). Greenwich, CT, USA: Information Age.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(4), 419-429. doi:10.1177/014662169001400407

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, NJ, USA: Wiley.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299. doi:10.1037/1040-3590.7.3.286

Forsythe, G. B., Mcgaghie, W. C., & Friedman, C. P. (1985). Factor structure of the resident evaluation form. *Educational and Psychological Measurement, 45*(2), 259-264. doi:10.1177/001316448504500208

Frohna, A., & Stern, D. (2005). The nature of qualitative comments in evaluating professionalism. *Medical Education, 39*(8), 763-768. doi:10.1111/j.1365-2929.2005.02234.x

Fuentealba, I. C., & Hecker, K. G. (2008). Clinical preceptor evaluation of veterinary students in a distributed model of clinical education. *Journal of Veterinary Medical Education, 35*(3), 389-396. doi:10.3138/jvme.35.3.389

Fuentealba, I. C., Mason, R. V., & Johnston, S. D. (2008). Community-based clinical veterinary education at Western University of Health Sciences. *Journal of Veterinary Medical Education, 35*(1), 34-42. doi:10.3138/jvme.35.1.034

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics : An introduction* (2nd ed.). Thousand Oaks, CA, USA: Sage.

Garrido, L. E. (2012). *Dimensionality assessment of ordinal variables: An evaluation of classic and modern methods.* (Doctoral thesis, Universidad Autónoma de Madrid, Madrid, Spain). Retrieved from https://repositorio.uam.es/handle/10486/11283

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's minimum average partial factor retention method with categorical variables. *Educational and Psychological Measurement, 71*(3), 551-570. doi:10.1177/0013164410389489

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods, 18*(4), 454-474. doi:10.1037/a0030005

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies, 51*(3), 511-521. doi:10.1016/j.ijnurstu.2013.10.005

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107-143. doi:10.1111/j.1756-8765.2008.01006.x

Gingerich, A. (2015). What if the 'trust' in entrustable were a social judgement? *Medical Education, 49*(8), 750-752. doi:10.1111/medu.12772

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055-1068. doi:10.1111/medu.12546

Gingerich, A., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2014). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in mini-CEX ratings. *Academic Medicine, 89*(11), 1510-1519. doi:10.1097/ACM.0000000000000486

Ginsburg, S., Eva, K. W., & Regehr, G. (2013). Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Academic Medicine, 88*(10), 1539-1544. doi:10.1097/acm.0b013e3182a36c3d

Ginsburg, S., Gold, W., Cavalcanti, R. B., Kurabi, B., & McDonald-Blumer, H. (2011). Competencies "plus": The nature of written comments on internal medicine residents' evaluation forms. *Academic Medicine, 86*(10 Suppl.), S30-S34. doi:10.1097/ACM.0b013e 31822a6d92

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K. W., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine, 85*(5), 780-786. doi:10.1097/Acm.0b013e3181d73fb6

Ginsburg, S., Regehr, G., Hatala, R., Mcnaughton, N., Frohna, A., Hodges, B., . . . Stern, D. (2000). Context, conflict, and resolution: A new conceptual framework for evaluating professionalism. *Academic Medicine, 75*(10 Suppl.), S6-S11. doi:10.1097/00001888-200010001-00003

Ginsburg, S., Regehr, G., & Lingard, L. (2004). Basing the evaluation of professionalism on observable behaviors: A cautionary tale. *Academic Medicine, 79*(10 Suppl.), S1-S4. doi:10.1097/00001888-200410001-00001

Ginsburg, S., Regehr, G., & Mylopoulos, M. (2009). From behaviours to attributions: Further concerns regarding the evaluation of professionalism. *Medical Education, 43*(5), 414-425. doi:10.1111/j.1365-2923.2009.03335.x

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi:10.2307/1169991

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy and Practice, 1*(1), 27-44. doi:10.1080/0969594940010103

Gonczi, A., & Hager, P. (2010). The competency model. In P. Penelope, B. Eva, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 8, pp. 403-410). Oxford, UK: Elsevier.

Gonnella, J. S., Hojat, M., Erdmann, J. B., & Veloski, J. J. (1993). A case of mistaken identity. *Academic Medicine, 68*(2), S9-S16. doi:10.1097/00001888-199302000-00023

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ, USA: Lawrence Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*(3), 532-560. doi:10.1207/s15327752jpa6803_5

Gough, H. G., Hall, W. B., & Harris, R. E. (1964). Evaluation of performance in medical training. *Journal of Medical Education, 39*(7), 679-692. Retrieved from http://journals.lww.com/academicmedicine/Fulltext/1964/07000/Evaluation_of_Performance_in_Medical_Training_.3.aspx

Govaerts, M. J. B. (2008). Educational competencies or education for professional competence? *Medical Education, 42*(3), 234-236. doi:10.1111/j.1365-2923.2007.03001.x

Govaerts, M. J. B., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education, 16*(2), 151-165. doi:10.1007/s10459-010-9250-7

Govaerts, M. J. B., van de Wiel, M. W. J., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375-396. doi:10.1007/s10459-012-9376-x

Govaerts, M. J. B., & van der Vleuten, C. P. M. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education, 47*(12), 1164-1174. doi:10.1111/medu.12289

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education, 12*(2), 239-260. doi:10.1007/s10459-006-9043-1

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*(3), 206-213. doi:10.1007/s11121-007-0070-9

Green, M. L., Aagaard, E. M., Caverzagie, K. J., Chick, D. A., Holmboe, E., Kane, G., . . . Iobst, W. (2009). Charting the road to competence: Developmental milestones for internal medicine residency training. *Journal of Graduate Medical Education, 1*(1), 5-20. doi:10.4300/01.01.0003

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274. doi:10.2307/1163620

Grilli, L., & Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling, 14*(1), 1-25. doi:10.1080/10705510709336734

Guerrasio, J., Cumbler, E., Trosterman, A., Wald, H., Brandenburg, S., & Aagaard, E. (2012). Determining need for remediation through postrotation evaluations. *Journal of Graduate Medical Education, 4*(1), 47-51. doi:10.4300/JGME-D-11-00145.1

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT, USA: Praeger.

Hamdy, H. (2009). AMEE guide supplements: Workplace-based assessment as an educational tool. Guide supplement 31.1-viewpoint. *Medical Teacher, 31*(1), 59-60. doi:10.1080/01421590802298215

Hamdy, H., Prasad, K., Anderson, M. B., Scherpbier, A., Williams, R., Zwierstra, R., & Cuddihy, H. (2006). BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher, 28*(2), 103-116. doi:10.1080/01421590600622723

Hanson, J. L., Rosenberg, A. A., & Lane, J. L. (2013). Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Frontiers in Psychology, 4*(668), 1-10. doi:10.3389/fpsyg.2013.00668

Hardie, E. M. (2008). Current methods in use for assessing clinical competencies: What works? *Journal of Veterinary Medical Education, 35*(3), 359-368. doi:10.3138/jvme.35.3.359

Hashizume, C. T., Myhre, D. L., Hecker, K. G., Bailey, J. V., & Lockyer, J. M. (2016). Exploring the teaching motivations, satisfaction, and challenges of veterinary preceptors: A qualitative study. *Journal of Veterinary Medical Education, 43*(1), 1-9. doi:10.3138/jvme.0715-120R

Hauer, K. E., Oza, S. K., Kogan, J. R., Stankiewicz, C. A., Stenfors-Hayes, T., ten Cate, O., . . . O'Sullivan, P. S. (2015). How clinical supervisors develop trust in their trainees: A qualitative study. *Medical Education, 49*(8), 783-795. doi:10.1111/medu.12745

Hauer, K. E., ten Cate, O., Boscardin, C., Irby, D. M., Iobst, W., & O'Sullivan, P. S. (2014). Understanding trust as an essential element of trainee supervision and learning in the workplace. *Advances in Health Sciences Education, 19*(3), 435-456. doi:10.1007/s10459-013-9474-4

Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-clinical evaluation exercise: A review of the research. *Academic Medicine, 85*(9), 1453-1461. doi:10.1097/ACM.0b013e3181eac3e6

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205. doi:10.1177/1094428104263675

Heath, T. J., Hyams, J., Baguley, J., & Abbott, K. A. (2006). Effect of different methods of selection on the background, attitudes and career plans of first year veterinary students. *Australian Veterinary Journal, 84*(6), 217-222. doi:10.1111/j.1751-0813.2006.tb12804.x

Heath, T. J., Lynch-Blosse, M., & Lanyon, A. (1996). A longitudinal study of veterinary students and recent graduates 1. Backgrounds, plans and subsequent employment. *Australian Veterinary Journal, 74*(4), 291-296. doi:10.1111/j.1751-0813.1996.tb13778.x

Heath, T. J., & Mills, J. N. (1999). Starting work in veterinary practice: An employers' viewpoint. *Australian Veterinary Practitioner, 29*(4), 146-152.

Hecker, K. G., Norris, J., & Coe, J. B. (2012). Workplace-based assessment in a primary-care setting. *Journal of Veterinary Medical Education, 39*(3), 229-240. doi:10.3138/jvme.0612.054R

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416. doi:10.1177/0013164405282485

Hodges, B. D. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher, 35*(7), 564-568. doi:10.3109/0142159X.2013.789134

Hodges, B. D., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Academic Medicine, 74*(10), 1129-1134. doi:10.1097/00001888-199910000-00017

Hodgson, J. L., Pelzer, J. M., & Inzana, K. D. (2013). Beyond NAVMEC: Competency-based veterinary education and assessment of the professional competencies. *Journal of Veterinary Medical Education, 40*(2), 102-118. doi:10.3138/jvme.1012-092R

Hojat, M., Veloski, J. J., & Borenstein, B. D. (1986). Components of clinical competence ratings of physicians: An empirical approach. *Educational and Psychological Measurement, 46*(3), 761-769. doi:10.1177/0013164486463034

Holmboe, E. S., & Hawkins, R. E. (1998). Methods for evaluating the clinical competence of residents in internal medicine: A review. *Annals of Internal Medicine, 129*(1), 42-48. doi:10.7326/0003-4819-129-1-199807010-00011

Holmboe, E. S., Yamazaki, K., Edgar, L., Conforti, L., Yaghmour, N., Miller, R. S., & Hamstra, S. J. (2015). Reflections on the first 2 years of milestone implementation. *Journal of Graduate Medical Education, 7*(3), 506-511. doi:10.4300/JGME-07-03-43

Hoy-Mack, P. (2005). Workplace assessment in New Zealand : Stated intentions and realisations. *International Journal of Training Research, 3*(1), 79-95. doi:10.5172/ijtr.3.1.79

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi:10.1080/10705519909540118

Huddle, T. S., & Heudebert, G. R. (2007). Viewpoint: Taking apart the art: The risk of anatomizing clinical competence. *Academic Medicine, 82*(6), 536-541. doi:10.1097/ACM.0b013e3180555935

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253-263. doi:10.3102/0013189x14542154

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling, 5*(4), 344-364. doi:10.1080/10705519809540111

Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. *Medical Education, 47*(7), 650-673. doi:10.1111/medu.12220

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26. doi:10.3102/0013189X033007014

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267-293). New York, NY, USA: Cambridge University Press.

Kan Ma, H., Min, C., Neville, A., & Eva, K. (2013). How good is good? Students and assessors' perceptions of qualitative markers of performance. *Teaching and Learning in Medicine, 25*(1), 15-23. doi:10.1080/10401334.2012.741545

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000

Kassam, A., Donnon, T., & Rigby, I. (2014). Validity and reliability of an in-training evaluation report to measure the CanMEDS roles in emergency medicine residents. *Canadian Journal of Emergency Medicine, 16*(2), 144-150. doi:10.2310/8000.2013.130958

Kenward, M. G., & Carpenter, J. (2007). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research, 16*(3), 199-218. doi:10.1177/0962280206075304

Khan, K., & Ramachandran, S. (2012). Conceptual framework for performance assessment: Competency, competence and performance in the context of assessments in healthcare--deciphering the terminology. *Medical Teacher, 34*(11), 920-928. doi:10.3109/0142159X.2012.722707

Kline, P. (1994). *An easy guide to factor analysis*. London, UK: Routledge.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY, USA: Guilford Press.

Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. M. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 171-207). New York, NY, USA: Routledge.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education, 45*(10), 1048-1060. doi:10.1111/j.1365-2923.2011.04025.x

Kogan, J. R., Conforti, L. N., Bernabeo, E., Iobst, W., & Holmboe, E. (2015). How faculty members experience workplace-based assessment rater training: A qualitative study. *Medical Education, 49*(7), 692-708. doi:10.1111/medu.12733

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. *Journal of the American Medical Association, 302*(12), 1316-1326. doi:10.1001/jama.2009.1365

Kreiter, C. D., Ferguson, K., Lee, W.-C., Brennan, R. L., & Densen, P. (1998). A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Academic Medicine, 73*(12), 1294-1298. doi:10.1097/00001888-199812000-00021

Kreiter, C. D., & Ferguson, K. J. (2001). Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Evaluation and the Health Professions, 24*(1), 36-46. doi:10.1177/01632780122034768

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263-313). Bingley, UK: Emerald Group.

Kwolek, C. J., Donnelly, M. B., Sloan, D. A., Birrell, S. N., Strodel, W. E., & Schwartz, R. W. (1997). Ward evaluations: Should they be abandoned? *Journal of Surgical Research, 69*(1), 1-6. doi:10.1006/jsre.1997.5001

Lai, E. R., Wolfe, E. W., & Vickers, D. (2014). Differentiation of illusory and true halo in writing scores. *Educational and Psychological Measurement, 75*(1), 102-125. doi:10.1177/0013164414530990

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*(4), 223-232. doi:10.1016/j.hrmr.2008.03.002

Lance, C. E., LaPointe, J. A., & Fisicaro, S. A. (1994). Tests of three causal models of halo rater error. *Organizational Behavior and Human Decision Processes, 57*(1), 83-96. doi:10.1006/obhd.1994.1005

Lane, I. F., & Bogue, E. G. (2010). Faculty perspectives regarding the importance and place of nontechnical competencies in veterinary medical education at five North American colleges of veterinary medicine. *Journal of the American Veterinary Medical Association, 237*(1), 53-64. doi:10.2460/javma.237.1.53

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods, 42*(3), 871-876. doi:10.3758/BRM.42.3.871

Lavine, E., Regehr, G., Garwood, K., & Ginsburg, S. (2004). The role of attribution to clerk factors and contextual factors in supervisors' perceptions of clerks' behaviors. *Teaching and Learning in Medicine, 16*(4), 317-322. doi:10.1207/s15328015tlm1604_3

Le Deist, F. D., & Winterton, J. (2005). What is competence? *Human Resource Development International, 8*(1), 27-46. doi:10.1080/1367886042000338227

Lee, M., & Wimmers, P. F. (2010, April). *Construct validity of three clerkship performance assessments*. Paper presented at the American Educational Research Association Annual Meeting, Denver, CO, USA. Retrieved from http://files.eric.ed.gov/fulltext/ED509953.pdf

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School Psychology Quarterly, 22*(4), 557-584. doi:10.1037/1045-3830.22.4.557

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology, 66*, 799-823. doi:10.1146/annurev-psych-010213-115043

Levine, H. G., & McGuire, C. H. (1971). Rating habitual performance in graduate medical education. *Journal of Medical Education, 46*(4), 306-311. doi:10.1097/00001888-197104000-00007

Lievens, F., Sanchez, J. I., Bartram, D., & Brown, A. (2010). Lack of consensus among competency ratings of the same occupation: Noise or substance? *Journal of Applied Psychology, 95*(3), 562-571. doi:10.1037/a0018035

Lingard, L. (2012). Rethinking competence in the context of teamwork. In B. D. Hodges & L. Lingard (Eds.), *The question of competence: Reconsidering medical education in the twenty-first century* (pp. 42-69). Ithaca, NY, USA: Cornell University Press.

Littlefield, J. H., DaRosa, D. A., Paukert, J., Williams, R. G., Klamen, D. L., & Schoolfield, J. D. (2005). Improving resident performance assessment data: Numeric precision and narrative specificity. *Academic Medicine, 80*(5), 489-495. doi:10.1097/00001888-200505000-00018

Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2009). Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: A systematic review. *Academic Medicine, 84*(3), 301-309. doi:10.1097/ACM.0b013e3181971f08

Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2011). Pitfalls in assessment of competency-based educational objectives. *Academic Medicine, 86*(4), 412-414. doi:10.1097/ACM.0b013e31820cdb28

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113-139. doi:10.1207/S15327906MBR3801_5

MacCallum, R. C. (2009). Factor analysis. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 123-147). Los Angeles, CA, USA: Sage.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. doi:10.1037/1082-989x.1.2.130

Magnier, K. M., Dale, V. H. M., & Pead, M. J. (2012). Workplace-based assessment instruments in the health sciences. *Journal of Veterinary Medical Education, 39*(4), 389-395. doi:10.3138/jvme.1211-118r

Magnier, K. M., Wang, R., Dale, V. H. M., Murphy, R., Hammond, R. A., Mossop, L., . . . Pead, M. J. (2011). Enhancing clinical learning in the workplace: A qualitative study. *Veterinary Record, 169*(26), 682. doi:10.1136/vr.100297

Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P., & Hawkins, R. E. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine, 81*(10 Suppl.), S56-S60. doi:10.1097/01.ACM.0000236514.53194.f4

Marshall, S. W. (2007). Power for tests of interaction: Effect of raising the type I error rate. *Epidemiologic Perspectives and Innovations, 4*(4), 1-7. doi:10.1186/1742-5573-4-4

Massey University. (2012). *Bachelor of Veterinary Science programme guide*. Palmerston North, NZ: Institute of Veterinary, Animal and Biomedical Sciences, Massey University.

Massie, J., & Ali, J. M. (2016). Workplace-based assessment: A review of user perceptions and strategies to address the identified shortcomings. *Advances in Health Sciences Education, 21*(2), 455-473. doi:10.1007/s10459-015-9614-0

Mastenbroek, N. J. J. M., Jaarsma, A. D. C., Demerouti, E., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & van Beukelen, P. (2014). Burnout and engagement, and its predictors in young veterinary professionals: The influence of gender. *Veterinary Record, 174*(6), 144. doi:10.1136/vr.101762

Mastenbroek, N. J. J. M., Jaarsma, A. D. C., Scherpbier, A. J. J. A., van Beukelen, P., & Demerouti, E. (2014). The role of personal resources in explaining well-being and performance: A study among young veterinary professionals. *European Journal of Work and Organizational Psychology, 23*(2), 190-202. doi:10.1080/1359432x.2012.728040

Matthew, S. M., Ellis, R. A., & Taylor, R. M. (2011). New graduates' conceptions of and approaches to veterinary professional practice, and relationships to achievement during an undergraduate internship programme. *Advances in Health Sciences Education, 16*(2), 167-182. doi:10.1007/s10459-010-9252-5

Matthew, S. M., Taylor, R. M., & Ellis, R. A. (2010). Students' experiences of clinic-based learning during a final year veterinary internship programme. *Higher Education Research and Development, 29*(4), 389-404. doi:10.1080/07294361003717903

Maxim, B. R., & Dielman, T. E. (1987). Dimensionality, internal consistency and interrater reliability of clinical performance ratings. *Medical Education, 21*(2), 130-137. doi:10.1111/j.1365-2923.1987.tb00679.x

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11. doi:10.3102/0013189x033002003

Maxwell, J. A. (2012). *A realist approach for qualitative research*. London, UK: Sage.

Maxwell, J. A. (2013). *Qualitative research design : An interactive approach* (3rd ed.). Thousand Oaks, CA, USA: Sage.

Maxwell, J. A., & Mittapalli, K. (2010). Realism as a stance for mixed method research. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (2nd ed., pp. 145-167). Los Angeles, CA: Sage.

Mazor, K. M., Canavan, C., Farrell, M., Margolis, M. J., & Clauser, B. E. (2008). Collecting validity evidence for an assessment of professionalism: Findings from think-aloud interviews. *Academic Medicine, 83*(10 Suppl.), S9-S12. doi:10.1097/ACM.0b013e318183e329

Mazor, K. M., Clauser, B. E., Holtman, M., & Margolis, M. J. (2007). Evaluation of missing data in an assessment of professional behaviors. *Academic Medicine, 82*(10 Suppl.), S44-S47. doi:10.1097/ACM.0b013e3181404fc6

Mazor, K. M., Zanetti, M. L., Alper, E. J., Hatem, D., Barrett, S. V., Meterko, V., . . . Pugnaire, M. P. (2007). Assessing professionalism in the context of an objective structured clinical examination: An in-depth study of the rating process. *Medical Education, 41*(4), 331-340. doi:10.1111/j.1365-2929.2006.02692.x

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdayle, NJ, USA: Lawrence Erlbaum.

McGill, D. A., van der Vleuten, C. P. M., & Clarke, M. J. (2011). Supervisor assessment of clinical and professional competence of medical trainees: A reliability study using workplace data and a focused analytical literature review. *Advances in Health Sciences Education, 16*(3), 405-425. doi:10.1007/s10459-011-9296-1

McGill, D. A., van der Vleuten, C. P. M., & Clarke, M. J. (2013). A critical evaluation of the validity and the reliability of global competency constructs for supervisor assessment of junior medical trainees. *Advances in Health Sciences Education, 18*(4), 701-725. doi:10.1007/s10459-012-9410-z

McLaughlin, K., Vitale, G., Coderre, S., Violato, C., & Wright, B. (2009). Clerkship evaluation–what are we measuring? *Medical Teacher, 31*(2), e36-e39. doi:10.1080/01421590802334309

McQueen, S. A., Petrisor, B., Bhandari, M., Fahim, C., McKinnon, V., & Sonnadara, R. R. (2016). Examining the barriers to meaningful assessment and feedback in medical training. *The American Journal of Surgery, 211*(2), 464-475. doi:10.1016/j.amjsurg.2015.10.002

Medical Practitioners Act, No. 95. (1995). Retrieved from http://www.nzlii.org/nz/legis/hist_act/mpa19951995n95242.pdf

Medigovich, K. (2012). *Satisfactory, good and outstanding nurses: Perceptions of nurses, their colleagues and patients.* (Doctoral thesis, Murdoch University, Australia). Retrieved from http://researchrepository.murdoch.edu.au/10894/

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research. A guide to design and implementation* (4th ed.). San Francisco, CA, USA: Jossey-Bass.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. doi:10.3102/0013189x018002005

Metheny, W. P. (1991). Limitations of physician ratings in the assessment of student clinical performance in an obstetrics and gynecology clerkship. *Obstetrics and Gynecology, 78*(1), 136-141. Retrieved from http://journals.lww.com/greenjournal/Fulltext/1991/07000/limitations_of_physician_ratings_in_the_assessment.27.aspx

Miller, G. E. (1990). The assessment of clinical skills / competence / performance. *Academic Medicine, 65*(9), S63-S67. doi:10.1097/00001888-199009000-00045

Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L. M., van der Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5), 1087-1102. doi:10.1007/s10459-013-9450-z

Moore, D. R., Cheng, M. I., & Dainty, A. R. J. (2002). Competence, competency and competencies: Performance assessment in organisations. *Work Study, 51*(6), 314-319. doi:10.1108/00438020210441876

Mossop, E. (2012). *Defining and teaching veterinary professionalism.* (Doctoral thesis, University of Nottingham, UK). Retrieved from http://eprints.nottingham.ac.uk/12694/1/Fully_corrected_thesis_Liz_Mossop.pdf

Mulder, M., Weigel, T., & Collins, K. (2007). The concept of competence in the development of vocational education and training in selected EU member states: A critical analysis. *Journal of Vocational Education and Training, 59*(1), 67-88. doi:10.1080/13636820601145630

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology, 67*(2), 161-164. doi:10.1037/0021-9010.67.2.161

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based behavior ratings and performance evaluations: Consequences for rating accuracy. *Journal of Applied Psychology, 71*(1), 39-44. doi:10.1037/0021-9010.71.1.39

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218-225. doi:10.1037/0021-9010.78.2.218

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189. doi:10.1111/j.2044-8317.1985.tb00832.x

Nasca, T. J., Gonnella, J. S., Hojat, M., Veloski, J., Erdmann, J. B., Robeson, M., . . . Callahan, C. (2002). Conceptualization and measurement of clinical competence of residents: A brief rating form and its psychometric properties. *Medical Teacher, 24*(3), 299-303. doi:10.1080/01421590220134141

National Board of Veterinary Medical Examiners. (2015). 2015-2016 North American Veterinary Licensing Examination (NAVLE). Bulletin of information for candidates. Bismarck, ND, USA: Author. Retrieved from https://www.nbvme.org/navle-general-information/candidate-bulletin/

Nestel, D., Walker, K., Simon, R., Aggarwal, R., & Andreatta, P. (2011). Nontechnical skills an inaccurate and unhelpful descriptor? *Simulation in Healthcare, 6*(1), 2-3. doi:10.1097/Sih.0b013e3182069587

Nich, C., & Carroll, K. (1997). Now you see it, now you don't: A comparison of traditional versus random-effects regression models in the analysis of longitudinal follow-up data from a clinical trial. *Journal of Consulting and Clinical Psychology, 65*(2), 252-261. doi:10.1037/0022-006x.65.2.252

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291-310. doi:10.1037/0033-295x.108.2.291

Non-technical. (2012). In *OED Online*. Retrieved from http://www.oed.com/view/Entry/256724?redirectedFrom=nontechnical#eid

Norcini, J. J. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*(10), 795-799. doi:10.7326/0003-4819-123-10-199511150-00008

Norman, G. R. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*(1), 37-49. doi:10.1007/s10459-009-9179-x

Norris, N. (1991). The trouble with competence. *Cambridge Journal of Education, 21*(3), 331-341. doi:10.1080/0305764910210307

North American Veterinary Medical Education Consortium. (2011). *Roadmap for veterinary medical education in the 21st century: Responsive, collaborative, flexible*. Washington, DC, USA: Author. Retrieved from http://www.aavmc.org/Veterinary-Educators/NAVMEC.aspx

O'Cathain, A. (2010). Assessing the quality of mixed methods research: Towards a comprehensive framework. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (2nd ed., pp. 531-555). Thousand Oaks, CA, USA: Sage.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, and Computers, 32*(3), 396-402. doi:10.3758/BF03200807

Osborne, J. W. (2015). *Best practices in logistic regression*. Thousand Oaks, CA, USA: Sage.

Outhwaite, W. (1987). *New philosophies of social science : Realism, hermeneutics and critical theory*. Basingstoke, UK: Macmillan Education.

Paget, M., Wu, C., McIlwrick, J., Woloschuk, W., Wright, B., & McLaughlin, K. (2013). Rater variables associated with ITER ratings. *Advances in Health Sciences Education, 18*(4), 551-557. doi:10.1007/s10459-012-9391-y

Paolo, A. M., & Bonaminio, G. A. (2003). Measuring outcomes of undergraduate medical education: Residency directors' ratings of first-year residents. *Academic Medicine, 78*(1), 90-95. doi:10.1097/00001888-200301000-00017

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. doi:10.1037/0021-9010.88.5.879

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*(1), 13-43. doi:10.1207/S15328031US0201_02

Prescott, L. E., Norcini, J. J., McKinlay, P., & Rennie, J. S. (2002). Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal evaluation of performance (LEP). *Medical Education, 36*(1), 92-97. doi:10.1046/j.1365-2923.2002.01099.x

Pritchard, W. (1989). Overview of the Pew report. *Journal of the American Veterinary Medical Association, 194*(7), 865-868.

Pulito, A. R., Donnelly, M. B., & Plymale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical Education, 41*(7), 667-675. doi:10.1111/j.1365-2923.2007.02787.x

Pulito, A. R., Donnelly, M. B., Plymale, M., & Mentzer, J. R. M. (2006). What do faculty observe of medical students' clinical performance? *Teaching and Learning in Medicine, 18*(2), 99-104. doi:10.1207/s15328015tlm1802_2

Punch, K. F. (2006). *Developing effective research proposals* (2nd ed.). London, UK: Sage.

Quarrick, E. A., & Sloop, E. W. (1972). A method for identifying the criteria of good performance in a medical clerkship program. *Journal of Medical Education, 47*(3), 188-197. doi:10.1097/00001888-197203000-00005

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167-190. doi:10.1007/BF02295939

Reddy, S., & Vijayakumar, S. (2000). Evaluating clinical skills of radiation oncology residents: Parts I and II. *International Journal of Cancer, 90*(1), 1-12. doi:10.1002/(sici)1097-0215(20000220)90:1<1::aid-ijc1>3.0.co;2-w

Rees, C. E., & Knight, L. V. (2007). Viewpoint: The trouble with assessing students' professionalism: Theoretical insights from sociocognitive psychology. *Academic Medicine, 82*(1), 46-50. doi:10.1097/01.acm.0000249931.85609.05

Regehr, C., Bogo, M., & Regehr, G. (2011). The development of an online practice-based evaluation tool for social work. *Research on Social Work Practice, 21*(4), 469-475. doi:10.1177/1049731510395948

Regehr, G., Bogo, M., Regehr, C., & Power, R. (2007). Can we build a better mousetrap? Improving the measures of practice performance in the field practicum. *Journal of Social Work Education, 43*(2), 327-343. doi:10.5175/JSWE.2007.200600607

Regehr, G., Eva, K., Ginsburg, S., Halwani, Y., & Sidhu, R. (2011). Assessment in postgraduate medical education: Trends and issues in assessment in the workplace. *A Paper Commissioned as part of the Environmental Scan for the Future of Medical Education in Canada Postgraduate Project*. Canada: Members of the Future of Medical Education in Canada Postgraduate Consortium. Retrieved from http://www.afmc.ca/pdf/fmec/13_Regehr_Assessment.pdf

Regehr, G., Ginsburg, S., Herold, J., Hatala, R., Eva, K., & Oulanova, O. (2012). Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. *Academic Medicine, 87*(4), 419-427. doi:10.1097/ACM.0b013e31824858a9

Rhind, S. M., Baillie, S., Brown, F., Hammick, M., & Dozier, M. (2008). Assessing competence in veterinary medical education: Where's the evidence? *Journal of Veterinary Medical Education, 35*(3), 407-411. doi:10.3138/jvme.35.3.407

Rhind, S. M., Baillie, S., Kinnison, T., Shaw, D. J., Bell, C. E., Mellanby, R. J., . . . Donnelly, R. (2011). The transition into veterinary practice: Opinions of recent graduates and final year students. *BMC Medical Education, 11*, 64. doi:10.1186/1472-6920-11-64

Rigdon, E. E., & Ferguson, C. E., Jr. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28*(4), 491-497. doi:10.2307/3172790

Root Kustritz, M. V., Molgaard, L. K., & Rendahl, A. (2011). Comparison of student self-assessment with faculty assessment of clinical competence. *Journal of Veterinary Medical Education, 38*(2), 163-170. doi:10.3138/jvme.38.2.163

Rosenbluth, G., O'Brien, B., Asher, E. M., & Cho, C. S. (2014). The "zing factor"–how do faculty describe the best pediatrics residents? *Journal of Graduate Medical Education, 6*(1), 106-111. doi:10.4300/JGME-D-13-00146.1

Roush, J. K., Rush, B. R., White, B. J., & Wilkerson, M. J. (2014). Correlation of pre-veterinary admissions criteria, intra-professional curriculum measures, AVMA-COE professional competency scores, and the NAVLE. *Journal of Veterinary Medical Education, 41*(1), 19-26. doi:10.3138/jvme.0613-087R1

Royal College of Physicians and Surgeons of Canada. (2015). Competence by design: Understanding milestones and EPAs. *Dialogue, 15*(2). Retrieved from http://www.royalcollege.ca/portal/page/portal/rc/resources/publications/dialogue/vol15_2/epa_milestones

Royal College of Veterinary Surgeons. (2014). *RCVS day one competences*. London, UK: Author. Retrieved from http://www.rcvs.org.uk/document-library/day-one-competences/

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment, 24*(2), 282-292. doi:10.1037/a0025697

Ryan, J. G., Mandel, F. S., Sama, A., & Ward, M. E. (1996). Reliability of faculty clinical evaluations of non-emergency medicine residents during emergency department rotations. *Academic Emergency Medicine, 3*(12), 1124-1130. doi:10.1111/j.1553-2712.1996.tb03372.x

Sadler, D. R. (2009a). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807-826. doi:10.1080/03075070802706553

Sadler, D. R. (2009b). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment and Evaluation in Higher Education, 34*(2), 159-179. doi:10.1080/02602930801956059

Saks, A. M., & Gruman, J. A. (2014). What do we really know about employee engagement? *Human Resource Development Quarterly, 25*(2), 155-182. doi:10.1002/hrdq.21187

Sanchez, J. I., & De La Torre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology, 81*(1), 3-10. doi:10.1037/0021-9010.81.1.3

Sandilands, D., & Zumbo, B. D. (2014). (Mis)alignment of medical education validation research with contemporary validity theory: The mini-CEX as an example. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (Vol. 54, pp. 289-310). Cham, Switzerland: Springer.

SAS Institute Inc. (2014). SAS/STAT® 13.2 user's guide. Cary, NC, USA: SAS Institute. Retrieved from http://support.sas.com/documentation/cdl/en/statug/67523/PDF/default/statug.pdf

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*(1), 73-103. doi:10.1080/00273170903504810

Scarlett, G. (2013). Report on partner practice visits. Sydney, Australia: Faculty of Veterinary Science, The University of Sydney

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi:10.1037/1082-989x.7.2.147

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321. doi:10.1177/0734282911406653

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement, 71*(1), 95-113. doi:10.1177/0013164410387348

Scholz, E., Hyams, J., Hayes, L., Raidal, S., Pollard-Williams, S., & Strong, M. (2015, May). *For love or money? Employers and undergraduate clinical education*. Paper presented at the Pan Pacific Veterinary Conference, Brisbane, Australia.

Schwind, C. J., Williams, R. G., Boehler, M. L., & Dunnington, G. L. (2004). Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Academic Medicine, 79*(5), 453-457. doi:10.1097/00001888-200405000-00016

Scott, C. S., Irby, D. M., Gilliland, B. C., & Hunt, D. D. (1993). Evaluating clinical skills in an undergraduate medical education curriculum. *Teaching and Learning in Medicine, 5*(1), 49-53. doi:10.1080/10401339309539588

Sebok, S. S., & Syer, M. D. (2015). Seeing things differently or seeing different things? Exploring raters' associations of noncognitive attributes. *Academic Medicine, 90*(11 Suppl.), S50-S55. doi:10.1097/ACM.0000000000000902

Shay, S. B. (2003). *The assessment of undergraduate final year projects: A study of academic professional judgment.* (Doctoral thesis, University of Cape Town, South Africa). Retrieved from https://open.uct.ac.za/handle/11427/13999

Shay, S. B. (2004). The assessment of complex performance: A socially situated interpretive act. *Harvard Educational Review, 74*(3), 307-329. doi:10.17763/haer.74.3.wq16l67103324520

Silber, C. G., Nasca, T. J., Paskin, D. L., Eiger, G., Robeson, M., & Veloski, J. J. (2004). Do global rating forms enable program directors to assess the ACGME competencies? *Academic Medicine, 79*(6), 549-556. doi:10.1097/00001888-200406000-00010

Silverman, D. (2014). *Interpreting qualitative data* (5th ed.). London, UK: Sage.

Skakun, E. N., Wilson, D. R., Taylor, W. C., & Langley, G. R. (1975). A preliminary examination of the in-training evaluation report. *Journal of Medical Education, 50*(8), 817-819. doi:10.1097/00001888-197508000-00012

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106*(1), 148-154. doi:10.1037/0033-2909.106.1.148

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, CA, USA: Sage.

Spratt, M., Carpenter, J., Sterne, J. A. C., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *American Journal of Epidemiology, 172*(4), 478-487. doi:10.1093/aje/kwq137

St-Onge, C., Chamberland, M., Levesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: Expert assessment of examinee performance. *Advances in Health Sciences Education, 21*(3), 627-642. doi:10.1007/s10459-015-9656-3

Stapleton, L. M. (2006). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling. A Second Course* (pp. 345-383). Greenwich, CT, USA: Information Age Publishing.

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika, 44*(2), 157-167. doi:10.1007/bf02293967

Sterkenburg, A., Barach, P., Kalkman, C., Gielen, M., & ten Cate, O. (2010). When do supervising physicians decide to entrust residents with unsupervised tasks? *Academic Medicine, 85*(9), 1408-1417. doi:10.1097/ACM.0b013e3181eab0ec

Stillman, P. L. (1986). Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine, 105*(5), 762-771. doi:10.7326/0003-4819-105-5-762

Stillman, P. L. (1991). Assessment of clinical skills of residents utilizing standardized patients. *Annals of Internal Medicine, 114*(5), 393-401. doi:10.7326/0003-4819-114-5-393

Stillman, P. L., Regan, M. B., Swanson, D. B., Case, S., Mccahan, J., Feinblatt, J., . . . Nelson, D. V. (1990). An assessment of the clinical skills of fourth-year students at four New England medical schools. *Academic Medicine, 65*(5), 320-326. doi:10.1097/00001888-199005000-00013

Stroud, L., Bryden, P., Kurabi, B., & Ginsburg, S. (2015). Putting performance in context: The perceived influence of environmental factors on work-based performance. *Perspectives on Medical Education, 4*(5), 233-243. doi:10.1007/s40037-015-0209-5

Stroup, W. W. (2013). *Generalized linear mixed models. Modern concepts, methods and applications*. Boca Raton, FL, USA: CRC Press.

Sultana, R. G. (2009). Competence and competence frameworks in career guidance: Complex and contested concepts. *International Journal for Educational and Vocational Guidance, 9*(1), 15-30. doi:10.1007/s10775-008-9148-6

Swanwick, T., & Chana, N. (2005). Workplace assessment for licensing in general practice. *British Journal of General Practice, 55*(515), 461-467. Retrieved from http://bjgp.org/content/55/515/461.short

Swing, S. R. (2002). Assessing the ACGME general competencies: General considerations and assessment methods. *Academic Emergency Medicine, 9*(11), 1278-1288. doi:10.1197/aemj.9.11.1278

Swing, S. R., Beeson, M. S., Carraccio, C., Coburn, M., Iobst, W., Selden, N. R., . . . Vydareny, K. (2013). Educational milestone development in the first 7 specialties to enter the next accreditation system. *Journal of Graduate Medical Education, 5*(1), 98-106. doi:10.4300/JGME-05-01-33

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA, USA: Pearson.

Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education, 18*(2), 291-303. doi:10.1007/s10459-012-9370-3

Tavares, W., & Eva, K. W. (2014). Impact of rating demands on rater-based assessments of clinical competence. *Education for Primary Care, 25*(6), 308-318. doi:10.1080/14739879.2014.11730760

Teherani, A., Hodgson, C. S., Banach, M., & Papadakis, M. A. (2005). Domains of unprofessional behavior during medical school associated with future disciplinary action by a state medical board. *Academic Medicine, 80*(10 Suppl.), S17-S20. doi:10.1097/00001888-200510001-00008

Teman, N. R., Minter, R. M., & Kasten, S. J. (2016). Utility of factor analysis in optimization of resident assessment and faculty evaluation. *The American Journal of Surgery, 211*(6), 1158-1163. doi:10.1016/j.amjsurg.2015.04.011

ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education, 39*(12), 1176-1177. doi:10.1111/j.1365-2929.2005.02341.x

ten Cate, O., Chen, H. C., Hoff, R. G., Peters, H., Bok, H., & van der Schaaf, M. (2015). Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide no. 99. *Medical Teacher, 37*(11), 983-1002. doi:10.3109/0142159x.2015.1060308

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. doi:10.1097/ACM.0b013e31805559c7

ten Cate, O., Snell, L., & Carraccio, C. (2010). Medical competence: The interplay between individual ability and the health care environment. *Medical Teacher, 32*(8), 669-675. doi:10.3109/0142159x.2010.500897

Theis, J. H. (2003). Veterinary medicine: A profession or a business? *Journal of Veterinary Medical Education, 30*(3), 207-210. doi:10.3138/jvme.30.3.207

Thomas, M. R., Beckman, T. J., Mauck, K. F., Cha, S. S., & Thomas, K. G. (2011). Group assessments of resident physicians improve reliability and decrease halo error. *Journal of General Internal Medicine, 26*(7), 759-764. doi:10.1007/s11606-011-1670-4

Thompson, B. (2004). *Exploratory and confirmatory factor analysis. Understanding concepts and applications*. Washington, DC, USA: American Psychological Association.

Thompson, W. G., Lipkin, M., Jr., Gilbert, D. A., Guzzo, R. A., & Roberson, L. (1990). Evaluating evaluation. *Journal of General Internal Medicine, 5*(3), 214-217. doi:10.1007/BF02600537

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29. doi:10.1037/h0071663

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. doi:10.1037/a0023353

Tourangeau, R. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68*(3), 368-393. doi:10.1093/poq/nfh035

Turnbull, J., & van Barneveld, C. (2002). Assessment of clinical performance: In-training evaluation. In G. R. Norman, C. P. M. Vleuten, & D. I. Newble (Eds.), *International handbook of research in medical education* (Vol. 7, pp. 793-810). Dordrecht, Netherlands: Kluwer Academic.

Tweed, M., & Ingham, C. (2010). Observed consultation: Confidence and accuracy of assessors. *Advances in Health Sciences Education, 15*(1), 31-43. doi:10.1007/s10459-009-9163-5

van Barneveld, C. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine, 80*(3), 309-312. doi:10.1097/00001888-200503000-00023

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242. doi:10.1177/0962280206074463

van der Vleuten, C. P. M. (2014). When I say ... Context specificity. *Medical Education, 48*(3), 234-235. doi:10.1111/medu.12263

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317. doi:10.1111/j.1365-2929.2005.02094.x

van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice and Research: Clinical Obstetrics and Gynaecology, 24*(6), 703-719. doi:10.1016/j.bpobgyn.2010.04.001

van Gelderen, I. (2015). *Investigating supervisors' experiences of veterinary intern placements.* (Master's thesis, University of Sydney, Australia). Retrieved from http://hdl.handle.net/2123/12928

van Lohuizen, M. T., Kuks, J. B., van Hell, E. A., Raat, A. N., Stewart, R. E., & Cohen-Schotanus, J. (2010). The reliability of in-training assessment when performance improvement is taken into account. *Advances in Health Sciences Education, 15*(5), 659-669. doi:10.1007/s10459-010-9226-7

Vandeweerd, J.-M., Cambier, C., Romainville, M., Perrenoud, P., Desbrosse, F., Dugdale, A., & Gustin, P. (2014). Competency frameworks: Which format for which target? *Journal of Veterinary Medical Education, 41*(1), 27-36. doi:10.3138/jvme.0413-062R1

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327. doi:10.1007/BF02293557

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41-71). Boston, MA, USA: Kluwer Academic.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*(1), 1-28. doi:10.1207/s15327906mbr2501_1

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research, 23*(1), 42-59. doi:10.1177/0962280212445834

Verhulst, S. J., Colliver, J. A., Paiva, R. E., & Williams, R. G. (1986). A factor analysis study of performance of first-year residents. *Journal of Medical Education, 61*(2), 132-134. doi:10.1097/00001888-198602000-00010

Veterinarians Act, No. 126. (2005). Retrieved from http://www.legislation.govt.nz/act/public/2005/0126/latest/DLM363859.html

Veterinary Council of New Zealand. (2012). Policy on competence and competence assessment. Wellington, NZ: Author. Retrieved from http://www.vetcouncil.org.nz/documentation/Policies/VCNZ_Policy_CompetenceAndCompetenceReview.pdf

Veterinary Council of New Zealand. (n.d.). Competency standards and performance indicators for veterinarians. Wellington, NZ: Author. Retrieved from http://www.vetcouncil.org.nz/documentation/VCNZ_CompetencyStandardsAndPerformanceMeasuresForVeterinarians.pdf

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*(1), 108-131. doi:10.1037/0021-9010.90.1.108

Volk, J. O., Felsted, K. E., Cummings, R. E., Slocum, J. W., Cron, W. L., Ryan, K. G., & Moosbrugger, M. C. (2005). Executive summary of the AVMA-Pfizer business practices study. *Journal of the American Veterinary Medical Association, 226*(2), 212-218. doi:10.2460/javma.2005.226.212

Walker, D., Roberts, J., & Mehlhorn, J. (2015). The importance of soft skill development for veterinary technology graduates and veterinary businesses. *Business and Economic Research, 5*(2), 315-326. doi:10.5296/ber.v5i2.8328

Walsh, D. A., Osburn, B. I., & Christopher, M. M. (2001). Defining the attributes expected of graduating veterinary medical students. *Journal of the American Veterinary Medical Association, 219*(10), 1358-1365. doi:10.2460/javma.2001.219.1358

Walsh, D. A., Zeck, S., Wall, J. A., Smith, B. P., & Wilson, W. D. (2012). Criterion-referenced evaluation of day one clinical competencies of veterinary students: VOLES-the VMTH (Veterinary Medicine Teaching Hospital) online evaluation system. *Journal of Veterinary Medical Education, 39*(1), 46-61. doi:10.3138/jvme.0411.046R

Warne, R. T., & Larsen, R. (2014). Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis. *Psychological Test and Assessment Modeling, 56*(1), 104-123. Retrieved from http://works.bepress.com/rwarne/2/

Watson, R., & Thompson, D. R. (2006). Use of factor analysis in Journal of Advanced Nursing: Literature review. *Journal of Advanced Nursing, 55*(3), 330-341. doi:10.1111/j.1365-2648.2006.03915.x

Weijs, C. A., Coe, J. B., & Hecker, K. G. (2016). Final-year students' and clinical instructors' experience of workplace-based assessments used in a small-animal primary-veterinary-care clinical rotation. *Journal of Veterinary Medical Education, 43*(1), 382-392. doi:10.3138/jvme.1214-123R1

Weitz, G., Vinzentius, C., Twesten, C., Lehnert, H., Bonnemeier, H., & Konig, I. R. (2014). Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Journal for Medical Education, 31*(4), Doc41. doi:10.3205/zma000933

Weller, J. M., Misur, M., Nicolson, S., Morris, J., Ure, S., Crossley, J., & Jolly, B. (2014). Can I leave the theatre? A key to more reliable workplace-based assessment. *British Journal of Anaesthesia, 112*(6), 1083-1091. doi:10.1093/bja/aeu052

Wetzel, A. P. (2012). Factor analysis methods and validity evidence: A review of instrument development across the medical education continuum. *Academic Medicine, 87*(8), 1060-1069. doi:10.1097/ACM.0b013e31825d305d

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine, 29*(28), 2920-2931. doi:10.1002/sim.3944

Whitehead, C. R., Selleger, V., van de Kreeke, J., & Hodges, B. (2014). The 'missing person' in roles-based competency models: A historical, cross-national, contrastive case study. *Medical Education, 48*(8), 785-795. doi:10.1111/medu.12482

Wilkinson, J. R., Crossley, J. G. M., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education, 42*(4), 364-373. doi:10.1111/j.1365-2923.2008.03010.x

Wilkinson, T. J., & Frampton, C. M. (2004). Comprehensive undergraduate medical assessments improve prediction of clinical performance. *Medical Education, 38*(10), 1111-1116. doi:10.1111/j.1365-2929.2004.01962.x

Wilkinson, T. J., Moore, M., & Flynn, E. M. (2012). Professionalism in its time and place–some implications for medical education. *New Zealand Medical Journal, 125*(1358), 64-73. Retrieved from https://www.nzma.org.nz/journal/read-the-journal/all-issues/2010-2019/2012/vol-125-no-1358/view-wilkinson

Williams, B., Brown, T., & Onsman, A. (2012). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3). Retrieved from http://ro.ecu.edu.au/jephc/vol8/iss3/1/

Williams, R. G., Chen, X. P., Sanfey, H., Markwell, S. J., Mellinger, J. D., & Dunnington, G. L. (2014). The measured effect of delay in completing operative performance ratings on clarity and detail of ratings assigned. *Journal of Surgical Education, 71*(6), e132-e138. doi:10.1016/j.jsurg.2014.06.015

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine, 15*(4), 270-292. doi:10.1207/S15328015TLM1504_11

Williams, R. G., Verhulst, S., Colliver, J. A., & Dunnington, G. L. (2005). Assuring the reliability of resident performance appraisals: More items or more observations? *Surgery, 137*(2), 141-147. doi:10.1016/j.surg.2004.06.011

Willis, N. G., Monroe, F. A., Potworowski, J. A., Halbert, G., Evans, B. R., Smith, J. E., . . . Bradbrook, A. (2007). Envisioning the future of veterinary medical education: The Association of American Veterinary Medical Colleges foresight project, final report. *Journal of Veterinary Medical Education, 34*(1), 1-41. doi:10.3138/jvme.34.1.1

Winterton, J. (2009). Competence across europe: Highest common factor or lowest common denominator? *Journal of European Industrial Training, 33*(8/9), 681-700. doi:10.1108/03090590910993571

Woloschuk, W., McLaughlin, K., & Wright, B. (2010). Is undergraduate performance predictive of postgraduate performance? *Teaching and Learning in Medicine, 22*(3), 202-204. doi:10.1080/10401334.2010.488205

Woloschuk, W., McLaughlin, K., & Wright, B. (2013). Predicting performance on the Medical Council of Canada Qualifying Exam Part II. *Teaching and Learning in Medicine, 25*(3), 237-241. doi:10.1080/10401334.2013.797351

Woloschuk, W., Myhre, D., Jackson, W., McLaughlin, K., & Wright, B. (2014). Comparing the performance in family medicine residencies of graduates from longitudinal integrated clerkships and rotation-based clerkships. *Academic Medicine, 89*(2), 296-300. doi:10.1097/ACM.0000000000000113

Wong, G., Greenhalgh, T., Westhorp, G., & Pawson, R. (2012). Realist methods in medical education research: What are they and what can they contribute? *Medical Education, 46*(1), 89-96. doi:10.1111/j.1365-2923.2011.04045.x

Wood, T. J. (2013). Mental workload as a tool for understanding dual processes in rater-based assessments. *Advances in Health Sciences Education, 18*(3), 523-525. doi:10.1007/s10459-012-9396-6

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education, 19*(3), 409-427. doi:10.1007/s10459-013-9453-9

Wothke, W. (1993). Nonpositive definite matrices in structural modeling In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256-293). Newbury Park, CA, USA: Sage.

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods, 47*(3), 756-772. doi:10.3758/s13428-014-0499-2

Yeates, P., O'Neill, P., Mann, K., & Eva, K. W. (2013). Seeing the same thing differently. *Advances in Health Sciences Education, 18*(3), 325-341. doi:10.1007/s10459-012-9372-1

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442. doi:10.1037/0033-2909.99.3.432

Appendix A:
Interview protocol

Section 1 – excellent student

Think back over your years supervising final year veterinary students while they worked in the practice/clinic/rotation with you. Think first about a really excellent student that you have had contact with. Without mentioning names, please talk about what made this person an excellent student.

Examples of prompts used if necessary:

- Prompt for a single actual student rather than generalized opinions.
- Ensure covers both personal characteristics as well as attributes related to clinical performance.
- Were there any weaknesses?
- Explore what led to any inferences or judgements. For example:
    - "what makes you say that?" or
    - "can you give me an example of what you mean?" or
    - "what did you observe that led to that opinion"?
- Prompt for knowledge and application of knowledge, work with animals, technical skills, work with others, and communication, if not given spontaneously.

Section 2 – weak student

Now I would like you to think about a really weak student that you have supervised. Again, without mentioning names, please talk about what made this person a weak student.

Examples of prompts used if necessary:

- Prompt for a single actual student rather than generalized opinions.
- Ensure covers both personal characteristics as well as attributes related to clinical performance.
- Were there any strengths?

- Explore what led to any inferences or judgements. For example:
    - "what makes you say that?" or
    - "can you give me an example of what you mean?" or
    - "what did you observe that led to that opinion"?
- Prompt for knowledge and application of knowledge, work with animals, technical skills, work with others, and communication, if not given spontaneously.

Section 3 – marginal student

Finally please reflect on a marginal student that you have had contact with. By marginal I mean one who you found it difficult to decide whether they were performing to an adequate standard or not. Thinking about the aspects of performance that you have just talked about, can you describe this person and how he or she differed from the excellent and weak students?

Were there any other aspects of this student's performance that come to mind when you think of this person?

Examples of prompts used if necessary:

- Prompt for a single actual student rather than generalized opinions.
- Ensure covers both personal characteristics as well as attributes related to clinical performance.
- Explore what led to any inferences or judgements. For example:
    - "what makes you say that?" or
    - "can you give me an example of what you mean?" or
    - "what did you observe that led to that opinion"?
- Prompt for knowledge and application of knowledge, work with animals, technical skills, work with others, and communication, if not given spontaneously.

Section 4 – anything else

Do you have anything else to add? Do you have any questions pertaining to what we have been talking about?

Appendix B:
In-training evaluation instrument

BVSc5 STUDENT EVALUATION

AFFIX PHOTO HERE

Student Name: _____

Name of Practice: _____

Dates at Practice: (From) _____ (To) _____

ASSESSING VETERINARIAN TO COMPLETE:

Name of Veterinarian: _____

OVERALL GRADE: ☐ Excellent ☐ Good ☐ Satisfactory ☐ Marginal ☐ Fail

Out of Hours: ☐ Satisfactory ☐ Marginal ☐ Unsatisfactory ☐ N/A

Has the intern been absent this rotation? ☐ Yes ☐ No If yes, how many days? _____

Signed: _____ Date: _____
(Veterinarian)

Please select the appropriate grade for each assessment area.
Feel free to make any comments regarding this students' performance while with your practice.
Your comments are appreciated and will be made available to the students.

PROFESSIONAL ATTITUDE

Professional Judgement and Development

Very mature clinical demeanour; responsible and reliable.	Interest is clearly shown; knows limits; responsible and reliable.	Punctual; well motivated; dependable	Sometimes late; Poor personal presentation; Poor case management; compromises animal welfare.	Late; lacks interest; unacceptable personal presentation. Case management jeopardises animal welfare.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Assigned Tasks

Displays initiative but knows limits.	Always follows through on assigned tasks, does more than expected.	Completes assigned tasks but does no more	Inconsistent effort does not always complete assigned tasks.	Unreliable.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal I	Fail	N/A

Participation in Rosters

Very good team member, above average interest, consistently dependable.	Good team member, dependable, interested and communicative.	Participates in all activities but minimal participation in discussions	Interest is inconsistent, minimal participation in discussion, inconsistently available.	Disinterested, does not participate in discussion, unavailable when needed.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal I	Fail	N/A

Comments

1

CLINICAL SKILLS

History

Thorough, complete and well-organised.	Thorough and complete.	Generally complete; symptoms generally all elicited	Important information often not included; symptoms not elicited.	Generally incomplete; disorganized; important information missing.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Physical Examination

Accurate (including subtle changes).	Accurate.	Generally thorough; only minor omissions	Tendency to be superficial; lapses in sequence, major findings missed.	Incomplete; superficial; cursory or inaccurate; major findings missed.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Technical Ability

Outstanding.	Very good; pays attention to patient comfort.	Proficient; exhibits appropriate care	Minor deficiencies in technical skills.	Generally careless or inept.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Animal Handling

Outstanding.	Empathetic and skilful.	Demonstrates a reasonable level of skill in animal handling and patient care	Animal handling could be improved but not dangerous	Unacceptable.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Comments

KNOWLEDGE

Knowledge

Outstanding knowledge, exceptional understanding; very current.	Displays very good grasp of the discipline.	Solid fund of knowledge	Minimal level of knowledge; inaccuracies and/or serious gaps in knowledge are evident.	Knowledge is very limited.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Use of Knowledge

Outstanding integration of clinical presentation with theoretical knowledge.	Integration abilities are very good.	History and physical examination are integrated to arrive at a satisfactory differential diagnosis and plan	Difficulty in developing differential diagnosis and plan.	Consistently unable to develop a differential diagnosis or plan.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Self-directed Learning

Excellent use of resources for SDL.	SDL skills are very well utilised.	Adequate SDL	Inconsistent effort in SDL.	Consistently poor effort in SDL.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Comments

COMMUNICATION

Client Skills

Excellent rapport and communication of medical information to clients.	Establishes trust; respectful; good communication of information.	Appropriate interaction and effective relationships with clients	Has difficulty in developing effective relationships with clients.	Poor communication with clients; lacking respect, empathy and integrity.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Interaction with Clinicians

Consistently responsible and reliable, respects others' opinions; successful in working with others.	Responsible and reliable; sensitive to concerns of other professionals.	Gets along well with most people; generally reliable; works as part of the team	Does not always work well with team; demonstrates lack of sensitivity or maturity.	Often creates friction; disrespectful; usually untidy and/or unprofessional.	Not Applicable
○	○	○	○	○	○
Excellent	Good	Satisfactory	Marginal	Fail	N/A

Comments

GENERAL COMMENTS

Strengths

Weaknesses

IMPORTANT: PLEASE ENSURE AN <u>OVERALL GRADE</u> IS GIVEN
(ON FRONT PAGE)

Thank you for taking the time to complete this Student Evaluation.

Please see over for <u>Programme Feedback</u>

Programme Feedback

We are interested to know how you rate our Veterinary Programme based on your experiences with our students. For example:

- What are our strengths and weaknesses?
- What are we doing right?
- Are there any areas where we can improve?
- Is there anything we should stop doing?

Your comments and feedback will be valuable to us to help us enhance our programme for future students. **Please Note:** Any comments made here will **NOT** be made available to students.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Thank you for taking the time to share with us your thoughts.

Please return to this evaluation to the **Veterinary Undergraduate Office**:

POST
Freepost 114094
Institute of Veterinary, Animal & Biomedical Sciences
Massey University,
Private Bag 11 222
Palmerston North 4442
New Zealand

FAX
+64 6 350 5654

EMAIL
vetug@massey.ac.nz

Appendix C:
Missingness analysis

The data for factor analysis and regression was incomplete. The incompleteness included missing data and an unbalanced number of observations, with students having different numbers of evaluations. Missingness is important because it not only reduces the sample size, but it can introduce bias, if the missingness is a result of an underlying effect that influences the effects of interest (Tabachnick & Fidell, 2013). This can invalidate inferences and limit the generalisations that are possible because the missingness may confine the study to a sample that is not representative. Therefore, it is important that the missingness is characterised in order to determine its possible effects on conclusions and is addressed if possible. The degree of bias introduced by missingness is related to the proportion of missingness, the strength of association between missingness and data (White & Carlin, 2010), and how the missingness is managed during the analysis. This section of the appendix presents the results of the analysis of missingness for both the factor analysis and regression analysis. It also discusses the how this was managed and the potential effects on conclusions drawn. It will begin with a brief overview of the problems caused by missing data and by unbalanced data.

The problem of missing data

Missing data causes problems because it reduces the sample size, but also because it can introduce bias. Whether it introduces bias depends on the reasons for the missingness and whether there is a relationship between the occurrence of missing data and what is being investigated. For example, in a study investigating the body weight of a population, data may be missing because the heaviest respondents do not answer a question regarding their weight. This will mean that the remaining data does not represent the entire population and the findings will be biased towards lighter members of the population.

One common method of managing missing data is complete case analysis, which involves only analysing cases with complete data and is sometimes referred to as listwise deletion. Many statistical software procedures can only analyse complete cases and this may not be apparent

to researchers. While simple and straightforward to manage, complete case analysis is not recommended because it reduces the sample size and can introduce bias (Allison, 2009). Even a small proportion of missing data can result in severe reductions in sample size if the missingness is scattered across many cases. The resulting loss of power can reduce the ability to detect significant differences and compromise inferences. Depending on which variables have missing data, bias is introduced unless the missingness is completely at random (Schafer & Graham, 2002), as explained below.

Another method of managing missing data is called available case analysis, which can be performed for some statistical procedures such as factor analysis that analyse covariance structures. Covariance and correlation matrices can be constructed using all available pairs of data, and because only some pairs of data are removed from analysis, this process is also known as pairwise deletion. Each covariance or correlation is then based on a different number of cases which makes it difficult to determine appropriate standard errors (Schafer & Graham, 2002). Like complete case analysis, bias is also introduced unless the missingness is completely at random (Allison, 2009).

Two preferred methods for managing missing data are maximum likelihood analysis and multiple imputation (Schafer & Graham, 2002). Both of these are statistical modelling procedures which make use of all the information available in the modelled data. They either implicitly or explicitly impute the missing data based on the distribution of the observed data, taking into account covariates (Bell & Fairclough, 2014). Maximum likelihood methods are incorporated into mixed modelling procedures often used for repeated measures or longitudinal data and properly account for missing data and unequal numbers of observations, as long as the missingness is missing at random (Bell & Fairclough, 2014), as explained below. Multiple imputation involves using a regression model to estimate and impute the missing data based on the relationships between variables (Allison, 2003). It introduces random variation so there is not perfect correlation between the imputed items and the known items, which would lead to inflated correlation estimates. Imputation is repeated to give multiple imputed datasets with slightly different estimates and standard errors. The estimates and standard errors are then combined by an averaging process which reduces the sampling variation and gives more accurate estimates and standard errors than any one single imputation would do. Multiple imputation involves more steps, but is able to be applied in a broader range of situations than maximum likelihood methods (Allison, 2003). The separation

of steps enables a different statistical model to be used for imputation than for analysis of the data. This is potentially advantageous because it allows the inclusion of covariates specifically of relevance to missingness but not to the substantive questions of the analysis (Bell & Fairclough, 2014). As is discussed below, the inclusion of covariates is an important mechanism by which bias is reduced by these methods because it changes the type of missingness and assists estimation.

Missingness can be categorised into three types that have differing implications (Table C.1) (Allison, 2009). If missingness is completely at random (MCAR) there is a reduction in sample size, but no bias is introduced, and therefore the missingness can be ignored if it is small in amount (Allison, 2009). The term MCAR can be confusing because the missingness may not be random in a general sense (Collins et al., 2001). However, it is random in respect to all the variables in the analysis model. This means that there should be no detectable relationship between the frequency of missingness and the values of any variables in the model. Also, although not directly testable, there should also be no reason to suspect that the missingness is related to the value that the missing variable would have had if it were not missing.

Table C.1: Definitions of types of missingness (Allison, 2009).

Type of missingness	Abbreviation	Distribution of missing data
Missing completely at random	MCAR	Unrelated to the value of other variables in the model and also unrelated to the value of the variable that is missing.
Missing at random	MAR	Related to the value of other variables in the model but unrelated to the value of the variable that is missing.
Not missing at random	NMAR	Related to the value of the variable that is missing

Commonly missingness does not meet the criteria for MCAR and is one of the other types. Missing at random (MAR) is predictable in that it is related to the values of other variables within the model, but it does not depend on the value of the missing variable itself. Therefore

for any given level of the other variable, the missingness of the missing variable is random. Whether MAR causes an analysis to be biased depends on which variable is missing and which variable the missingness is related to. If the missingness is in the dependent variable[3] and it is related to a covariate then the analysis is not biased by the missingness as long as the covariate is included in the model. This applies to complete case analysis, multiple imputation and maximum likelihood analysis methods of managing missingness (White & Carlin, 2010). For multiple imputation, it is the imputation model that must include the covariate, but not necessarily the analysis model. If missingness is in a covariate, and it is related to another covariate, then the analysis is not biased by the missingness as long as the other covariate is included in the model, regardless of whether complete case analysis, multiple imputation or maximum likelihood methods are used to manage the missingness (Allison, 2000, 2009). However, if the missingness is in a covariate and it is related to the dependent variable, then analysis with complete case analysis will result in bias, but analysis with multiple imputation or maximum likelihood will not be biased, as long as the covariate is included in the model (Allison, 2000).

Missingness that is related to the value of the missing variable is not missing at random (NMAR) and can also result in bias, again depending on which variable is missing and on the method of managing missingness. When the dependent variable is NMAR, bias is introduced with all methods but can be reduced if one or more covariates that are strongly related to the dependent variable can be included in the model (Collins et al., 2001). The covariate then provides the information from which to impute or model the missing variable (van Buuren, 2007). For example in a study of body weight, missingness may be associated with body weight if obese people are reluctant to be measured (NMAR). However there may be less missingness in a covariate such as waist circumference that is highly correlated to body weight. In this case, missing body weight data can be imputed or modelled from waist circumference (van Buuren, 2007). An alternative to including a covariate that is strongly related to the value of the missing variable is to include one that is strongly related to the missingness of the missing variable. Addition of such a covariate can convert missingness that is NMAR to MAR and thereby also reduce bias. For example, in a study of the scores of students in an examination, students who are likely to perform poorly may not attend the examination, thus missingness is related to the

[3] The dependent variable is the outcome variable or the variable of interest in the model.

dependent variable of examination scores and is therefore NMAR. However the missingness may also be highly correlated with pretest scores. Students with low pretest scores may be more likely to decide not to attend the examination and will have missing examination scores. Inclusion of pretest scores as a covariate in the model allows the missingness to be MAR, because for any given pretest score, the missingness in examination scores is then random (Dong & Peng, 2013).

When a covariate is NMAR complete case analysis is not biased (Allison, 2000), but multiple imputation and maximum likelihood are both biased (Bartlett, Carpenter, Tilling, & Vansteelandt, 2014; Spratt et al., 2010) because they model the data based on the covariates. Biased estimates of the dependent variable will result if the estimates of the covariates are themselves biased through imputation or maximum likelihood modelling. However, just as discussed before, inclusion of suitable additional covariates that are strongly correlated with either the missing covariate or the missingness can alleviate this bias.

NMAR is difficult to identify and manage because it is related to information that has not been collected. In practice most analyses proceed on the assumption of MAR, although it is likely that this is often violated (White & Carlin, 2010). If additional variables that are correlates of missingness (Dong & Peng, 2013) or of the missing variables can be included in the model this can reduce bias associated with NMAR (Collins et al., 2001).

The problem of unbalanced data

Another type of missingness occurs when there is a different number of observations for each subject. This type of missingness is also known as unbalanced data and is a frequent occurrence in longitudinal studies (Verbeke, Fieuws, Molenberghs, & Davidian, 2014). It can lead to problems because some forms of analysis, such as repeated measures ANOVA, require data to be balanced, which means they must have equal numbers of observations on each subject and at each time point (Nich & Carroll, 1997). However linear mixed modelling allows unequal numbers of observations on subjects and unequal spacing of observation times, assuming the data are missing at random (Demirtas, 2004). This is because it produces a series of regression models for the repeated measurements on a subject over time and then relates these to the independent variables that may lead to differences between subjects (Demirtas,

2004). Different timing and numbers of observations are accommodated because the regression model estimates an intercept and slope for each subject and can do this with more or less data points. Maximum likelihood methods fully account for the uncertainty produced by missing data, in the calculation of standard errors (Allison, 2012).

Conclusion

Because of the implications for inferences drawn from a data analysis it is important to plan for the management of missing data during research design. The missingness patterns in data should be examined and covariates that may correlate with missing variables or with the mechanism of missingness should be included in the data collection and analysis. Wherever possible, principled methods of managing the missingness such as maximum likelihood or multiple imputation should be used in preference to complete case analysis or available case analysis. Principled methods are available in standard software used for statistical analysis and are able to be used with an increasingly broad range of types of data and analyses.

Missingness analysis method

The variables and covariates for analysis were examined for missingness. The dependent variable for factor analysis was item score and the dependent variable for regression was overall grade. Other variables were placement, placement date, GPA, academic status of the supervisor (academic), the compulsory nature of the placement (core), and factor scores which were derived from item scores.

The frequency of missing data for each of these variables was tabulated. Items scored as not applicable were treated as missing because this category is not part of the ordinal scale used for scoring and cannot be considered a different level of the same continuum of excellence.

The relationship between missingness and the dependent variable, other variables of interest, and possible covariates, was examined in order to determine the degree of bias that might arise. Whether missingness in a variable was related to its own value was examined indirectly using other “proxy” variables whose value was likely to be closely related to the value of the

missing variable. For missing item scores, overall grade was used as a proxy, on the basis that it was likely to be closely related to item score. For missing overall grades, the final total mark awarded for all placement evaluations (final placement mark) for that student was used as a proxy. This was a weighted average of overall grades achieved over the year for a student and likely to be related to the overall grade achieved on any one placement.

A number of methods were used to explore the missingness, depending on the type of variable. Observed and expected frequency of missingness was tabulated and compared. Differences of greater than 5% were considered of importance. Pearson correlation coefficients were calculated for numerical variables. Correlations of greater than 0.3 were considered of importance. Boxplots and scatterplots were also used as appropriate for the variable to look for relationships. Any substantial trends were verified with ordinal logistic regression using the SAS procedure GLIMMIX. The ordinal logistic regression model accounted for the repeated measures on each student and included interactions with covariates. Reasons for missingness could be inferred in some cases and this was also considered.

Missingness analysis results

Missingness was present in evaluations, overall grades, and item scores (Table C.2). There was no missingness in the covariates used for analysis.

Table C.2: Proportion of missingness in evaluations, overall grades, and item scores.

	number	proportion
Total number of evaluations	**3466**	**100%**
Evaluation-level missingness		**4.3%**
evaluations with overall grades missing and all items missing	149	
All-item missingness: overall grade present but all items missing		**2.9%**
evaluations with all 12 items missing (but had an overall grade)	101	
evaluations with all 12 items not applicable (but had an overall grade)	1	
Some-item missingness: overall grade present but some items missing		**46.5%**
evaluations with 1-11 items scored as not applicable	1344	
evaluations with 1-11 items blank	167	
evaluations with 1-11 items mixed not applicable and blank	99	
Complete evaluations (no missingness)	**1605**	**46.3%**

Evaluation-level missingness

Missingness of entire evaluations was at a low level of 4.3% of evaluations. This was highly related to placement and greater than expected missingness was confined to 1 placement (PA6). There was no apparent relationship between evaluation-level missingness and final placement mark.

All-item missingness

Evaluations in which an overall grade was given but no items were scored were infrequent and affected 2.9% of the data. This all-item missingness was highly related to placement and greater than expected missingness was confined to 3 particular placements (PA7, EQ2 and SA5). All-item missingness in other placements was very infrequent (42 of 3466 (1.2%) of evaluations).

There was a tendency for all-item missingness to be associated with an overall grade of good, suggesting there could be an association with the value of item scores themselves. However this tendency did not remain after removal of data from the three particular placements in

which greater than expected all-item missingness was present. Therefore, the tendency seemed to be confined to those three placements. There was no apparent relationship between all-item missingness and final placement mark. There was a pattern of all-item missingness in relation to time but this was related to the time of placements and therefore considered an indirect effect.

Some-item missingness

In the 3215 evaluations for which some items were evaluated, there was only a low frequency (7.9%) of items missing, and most were because they were scored not applicable (Table C.3). However almost half of all evaluations (46.5%) had some items missing. The majority of these (91.9%) had only 1-3 items missing. Missingness was concentrated in certain items: participation in placements (Partic), history taking (History), physical examination (Exam), self-directed learning (SDL), and client communication (Client).

Table C.3: Item-level missingness in 3215 evaluations that had some items scored[1].

	number	proportion
Number of items in all 3215 evaluations	**38580**	**100%**
Items not applicable	2623	6.8%
Items blank	406	1.1%
Items scored	35551	92.1%

Note: [1] Evaluations that remained after deletion of those with no overall grade or no items evaluated.

There was no correlation (r=-0.05) between the number of items missing on an evaluation and the overall grade awarded at that evaluation. However, a potential relationship between the occurrence of some-item missingness and the value of the item was suggested by higher than expected frequencies of missing items in evaluations with an overall grade of satisfactory or

fail. Ordinal logistic regression analysis suggested a significant relationship between some-item missingness and overall grade (p=0.0018).

Greater than expected observed frequency of some-item missingness was found for some levels of the independent variables placement, academic, and core. This suggested that the frequency of some-item missingness depended on these variables. A significant relationship with some-item missingness was confirmed for academic (p=0.0037) and core (p<0.0001). The relationship between some-item missingness and placement was not able to be statistically tested because of separation of data leading to lack of maxima[4].

There was no correlation (r=-0.04) between GPA and the number of missing items on an evaluation and no apparent relationship between some-item missingness and GPA (p=0.097).

Relationship of item value with other variables

Boxplots of item score according to placement, academic, and core suggested relationships between item score and these variables. A significant relationship with item score was confirmed for each (p<0.0001). Boxplots of GPA according to item score suggested a weak relationship. A significant relationship between GPA and item score was confirmed statistically (p=0.0021).

[4] Separation occurs when the relationship between two variables is such that one always predicts the value of the other (Allison, 2008). It is frequently seen in datasets containing uncommon events, such that for some levels of one variable the value of the other variable is always the same. This can be visualised in a frequency table as there being a zero in one of the table cells because there have been no occurrences of a rare event in association with one level of the other variable. For example if the table involved the frequency of passing and failing scores for girls and boys, and there were no occurrences of failure amongst girls, then the table cell for that combination would be zero. In the case of this dataset, zero cells occurred because supervisors in some placements never or rarely awarded certain overall grades. The statistical calculations involving logarithms cannot be performed when there are zeros and parameter estimates cannot be generated.

Missingness arising from the different numbers of evaluations performed on each student

Students were evaluated different numbers of times depending on their choices of electives and whether they chose to include extra placements over and above the requirement. Evaluations were also unequally spaced in time because different placements were of different duration and each student did placements in a different order. The majority of students were evaluated between 14 and 20 times (Figure C.1:). One student who repeated the year was evaluated 35 times.

There was no relationship between the number of evaluations for a student and the final weighted mark given for placements (r=-0.06), which suggested that there was no relationship between the performance of the student and the number of placements.

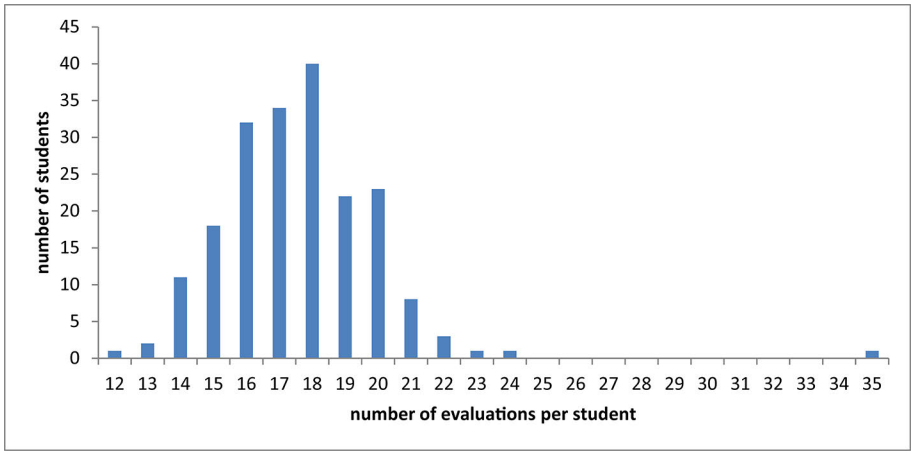


Figure C.1: Frequency distribution of the number of evaluations for each student.

Missingness in factor scores

The proportion of evaluations with missing factor scores ranged from 11.2% for the professional attitude factor to 30.2% for the knowledge factor (Table C.4). There was marked variation in the occurrence of missingness between placements, suggesting a relationship of factor score missingness with placement.

Only approximately half (49.9%) of all evaluations had a complete set of factor scores (Table C.5). Some placements had no or very few complete cases (for example PA3) and others had a high proportion of complete cases (for example SA1).

Table C.4: Percentage of evaluations with missingness in factor scores.

Placement	F1 clinical skills	F2 knowledge	F3 professional attitude	F4 communication	Total number of evaluations
EQ1	3.4%	1.7%	4.5%	2.3%	354
EXT	22.5%	22.1%	19.5%	9.4%	1040
OT4	34.3%	8.6%	0%	51.4%	35
PA1	24.9%	44.2%	2.3%	1.7%	346
PA2	19.3%	19.3%	17.3%	6.2%	243
PA3	26.2%	97.9%	0.9%	0.9%	233
PA4	53.5%	46.5%	31.3%	0%	198
PA5	25.2%	53.0%	21.7%	0%	115
PA6	1.0%	91.8%	1.0%	89.7%	97
SA1	0.7%	2.2%	0%	1.4%	279
SA2	80.3%	26.3%	0%	80.8%	198
SA3	3.4%	1.7%	0%	0%	59
SA4	0%	11.1%	0%	0%	18
Total	**23.4%**	**30.2%**	**11.2%**	**12.4%**	**3215**

Table C.5: Complete cases remaining when evaluations with any missing factor scores were deleted.

Placement	Complete cases (%)		Evaluations with missing factor scores (%)		Total evaluations
EQ1	315	(89%)	39	(11%)	354
EXT	534	(51.3%)	506	(48.7%)	1040
OT4	12	(34.3%)	23	(65.7%)	35
PA1	158	(45.7%)	188	(54.3%)	346
PA2	130	(53.5%)	113	(46.5%)	243
PA3	0	(0%)	233	(100%)	233
PA4	54	(27.3%)	144	(72.7%)	198
PA5	43	(37.4%)	72	(62.6%)	115
PA6	8	(8.2%)	89	(91.8%)	97
SA1	269	(96.4%)	10	(3.6%)	279
SA2	10	(5.1%)	188	(94.9%)	198
SA3	56	(94.9%)	3	(5.1%)	59
SA4	16	(88.9%)	2	(11.1%)	18
Total	**1605**	(**49.9%**)	**1610**	(**50.1%**)	**3215**

Implications of missingness for analysis

Factor analysis

The dependent variable for factor analysis was item score and the independent variables of interest were the items.

Evaluation-level and all-item missingness

The low level of evaluation-level and all-item missingness was managed by complete case analysis. Since this missingness was largely confined to four particular placement it was considered that for these placements (PA6, PA7, EQ2, and SA5) the results of the factor analysis may not be representative. For three of these the in-training evaluation was not

usually used and in the other, one particular supervisor disagreed with the use of the in-training evaluation for assessment of students on that placement. Therefore the lack of representativeness from these placements was not considered important for the overall conclusions. For the remaining placements there was no apparent relationship with the dependent variable as indicated by proxies. This suggests that for most placements, any bias from complete case analysis would be minimal.

Some-item missingness

Some-item missingness affected almost half of the evaluations and therefore complete case analysis would result in substantial loss of power and potential bias. Missingness in items appeared to be related to the dependent variable (item score) as indicated by the proxy overall grade and was also related to the potential covariates placement, academic, and core.

Some-item missingness was therefore managed by maximum likelihood estimation during construction of the correlation matrix for factor analysis using the SAS procedure MIXED. Covariates of placement as well as academic, core, and their interaction were added to the model in order to assist in reducing bias in the estimates.

Unbalanced data

The unbalanced data was modelled using linear mixed modelling with the SAS procedure MIXED. This enabled accurate construction of the covariance matrix for factor analysis for unbalanced data, assuming that the imbalance was not related to the value of the dependent variable (item score) (Demirtas, 2004). The mechanism of differences in numbers of evaluation for each student, and the lack of relationship between number of evaluations and final placement mark suggested this assumption was plausible.

Regression analysis

The dependent variable for regression analysis was overall grade and the independent variables were the factor scores, derived from item scores. Covariates of interest were placement, academic, and GPA.

Evaluation-level and all-item missingness

As for the factor analysis, evaluations with no overall grade and no items scored were not included in the regression analysis. Therefore, the results are not necessarily representative of the four placements with significant evaluation-level and all-item missingness. However for other placements minimal bias was expected because of the low level of missingness and lack of relationship with the dependent variable overall grade.

Some-item missingness

Some-item missingness lead to moderate missingness in factor scores which ranged from 11.2% to 30.2%. However because factor scores were independent variables in the regression model, the default analysis with PROC GLIMMIX used listwise deletion and complete case analysis (Allison, 2012; SAS Institute Inc., 2014). This effectively reduced the data by half for analysis. Bias due to the relationship of some-item missingness with the value of items was able to be effectively moderated by the covariates placement and academic, both of which had significant relationships with the missingness, and GPA, which had a significant relationship with the value of item scores. However this would not moderate bias due to a relationship of missingness with the value of the dependent variable overall grade. Therefore, complete case analysis might be expected to be biased as well as having reduced power.

Unlike in the factor analysis, where maximum likelihood estimation could model the data with missing dependent variables, modelling data with missing fixed effect covariates is not available through PROC MIXED and PROC GLIMMIX (SAS Institute Inc., 2014). It is possible through SAS PROC CALIS (Allison, 2012), however, this procedure is not suitable for ordinal data. Other specialised software such as MPlus would be needed to model missing covariates in a generalised linear mixed model (Allison, 2012).

Therefore, an alternative analysis with multiple imputation was explored. Missing items were imputed using SAS PROC MI with a fully conditional specification regression method, and including both the dependent variable overall grade and covariates placement, academic, core, and GPA in the imputation model. Twenty imputed datasets were constructed as recommended by Graham, Olchowski, and Gilreath (2007). The factor scores for all twenty datasets were derived using the factor coefficients from the factor analysis phase using the same method used for the regression analysis as detailed in Chapter 6. Regression was then

carried out using PROC GLIMMIX on each dataset (as detailed in Chapter 6) and the results combined using PROC MIANALYZE. This gives beta coefficients and *p* values for each level of each variable but does not give an overall type three test of fixed effects. The results of this analysis were compared to the complete case analysis. In both cases the *p* values indicated the same decision regarding the significance or insignificance of relationships at $p<0.05$ for each factor, academic, and GPA. The beta estimates for each factor were very similar and, importantly, remained in the same order from lowest to highest indicating the same relative strength of relationship with overall grade. Because no test of fixed effects was available with the imputed data, the overall significance of placement and its interactions could not be compared. There were some changes in which levels of placement and which placement by factor interactions were significantly related to overall grade, but the majority were the same. The comparison therefore suggested that the substantive conclusions of the analysis would be the same whether complete case analysis or multiple imputation was used for analysis. The complete case analysis results were therefore reported in Chapter 6 since the type three tests of fixed effects were available.

The finding that the same conclusions would be drawn with multiple imputation and complete case analysis does not indicate that neither is biased, as both could be equally biased. In this case, while the multiple imputation model did enable the ordinal nature of the data to be modelled using the fully conditional specification, it did not allow the dependency of repeated measures to be accounted for when imputing missing items. Therefore it did not model all aspects of the distribution of the data as is recommended (van Buuren, 2007). Methods for accounting for dependency during multiple imputation are, however, less well developed and available only in specialist software, and may not always be suitable when mixtures of categorical and continuous variables need to be imputed (Kenward & Carpenter, 2007). Formatting data so that there is only one entry per subject that incorporates all observations over time as separate variables (as opposed to one entry per observation) would allow imputation while preserving longitudinal relationships (Allison, 2009), but is not practical when there is significantly unbalanced data, as in my data.

Unbalanced data

Both complete case analysis and multiple imputation accounted for the different number of evaluations on each student through the use of PROC GLIMMIX for analysis.

Conclusion regarding the effect of missingness on this research

The findings of this missingness analysis suggest that the conclusions of the factor analysis would be little affected by missingness except in relation to the placements of PA6, PA7, EQ2 and SA5, for which they are likely not to be representative. However, because these placements do not usually use the in-training evaluation for assessment, this is of little consequence.

In relation to the regression analysis, there is a possibility that the results presented in Chapter 6 are biased by a relationship of missing factor scores to the value of overall grade. This may mean that the results only apply to certain overall grades or certain combinations of covariates with overall grade. Even though the majority of missing items were rated not applicable by supervisors rather than missing (as seen in Table C.3), Mazor, Clauser, et al. (2007) found that not applicable ratings were related to lower ratings in their study of medical students. This suggests that not applicable items might not be the result of random events unrelated to student performance. Therefore, despite the finding that principled missing data management with multiple imputation did not alter substantive conclusions, which suggests that significant biases were not present, there may still have been unaddressed NMAR mechanisms resulting in bias.

Other evidence supporting the validity of conclusions of the regression analysis comes from triangulation of the results from other parts of the research. The results of the regression analysis were found to be well aligned with the findings from the qualitative interview phase and the analysis of the content of the mark scheme for the in-training evaluation. This suggests that the conclusions reached were reasonable, despite the limitations produced by the factor score missingness.

Appendix D:
Details of factor analysis method

Suitability of the data for common factor analysis

Common factor analysis is a multivariate linear modelling technique. It assumes that that each variable (item scores in this case) is normally distributed and residuals are also normally distributed and independent (Tabachnick & Fidell, 2013). It assumes that variables are linearly related to factors although they are not assumed to be linearly related to each other (Gorsuch, 1997). However the data for analysis in the factor analysis (item scores) was ordinal in nature. To the extent that the interval between ordered categories is unknown and may not be equivalent, or even regular, the relationship between two ordinal variables is not linear, just as the relationship between two categorical variables is not linear. In addition, the distribution of ordinal variables is not always normal, and there can be significant skewness (with a tendency for most responses to be at one end of the response options) and kurtosis (with long thin tails or short thick tails in the distribution of responses). The nature of data for this research was therefore not consistent with the assumptions of factor analysis.

The significance of this for inferences depends on the extent of non-normality and non-linearity and the type of factor analysis performed. Severe violations of normality may cause biased maximum likelihood parameters and standard error estimates (Fabrigar et al., 1999; Schmitt, 2011) but the method remains fairly robust to lower level violations (Gorsuch, 1997). Bias is least when sample size is high, correlation between variables and factors is high and there are at least five ordinal response categories (Finney & DiStefano, 2006). Mild to moderate levels of skewness and kurtosis (up to -2.0 to 2.9) have been shown to cause little or no bias in parameter estimates (factor pattern and structure coefficients) with maximum likelihood estimation (Muthén & Kaplan, 1985). Standard errors, however, tended to be overly small unless skewness and kurtosis were less than -1.0 to 1.0 (Muthén & Kaplan, 1985). This means that while factor pattern and structure coefficients are likely not to be biased with low level departures from normality, confidence intervals around them will be narrower than they should be and this may suggest a significant contribution of a variable to a factor that is not

the case. Statistical measures of model fit will tend to suggest that more factors need to be retained, when in fact a model may fit well.

Using polychoric correlation matrices for factor analysis instead of Pearson correlations is an alternative for modelling ordinal data which may better represent the data. They are generally larger than Pearson correlations calculated from the same data (Bartholomew, Steele, Moustaki, & Galbraith, 2008). The use of polychoric correlations as input data for factor analysis still leads to biased standard errors when used with maximum likelihood estimation (Rigdon & Ferguson, 1991), but the bias is less than seen with Pearson correlations (Babakus, Ferguson, & Jöreskog, 1987). Polychoric correlations can be constructed using the SAS procedure CORR. Since the procedure uses maximum likelihood estimation it produces good estimates in the presence of missing items, but does not account for the dependency of repeated measures over time or the unbalanced nature of the data. These are significant shortcomings which can introduce substantial bias (Bliese & Hanges, 2004; Glass et al., 1972). Therefore, despite their advantages, polychoric correlation matrices were not used for this analysis.

From this discussion, it is apparent that as well as sample size and the number of categories in the ordinal responses, the extent of skewness and kurtosis of the data influences the severity of the consequences of violating assumptions of normality for exploratory factor analysis. Therefore, the degree of skewness and kurtosis in item scores was calculated using the SAS procedure MEANS. Estimates of parameters such as eigenvalues and pattern and structure coefficients were considered accurate with maximum likelihood estimation if skewness and kurtosis were between -2.0 and 2.9. Statistics derived from estimates of standard errors, including statistical methods of model fit, number of factors to retain and confidence intervals around structure and pattern coefficients were considered accurate if skewness and kurtosis were between -1.0 and 1.0. If skewness and kurtosis lay outside this range, the more conservative significance level of α=0.01 rather than the usual α=0.05 was used in order to limit type I errors.

Preparation of the adjusted correlation matrix for factor analysis

The dataset contained multiple observations on each student, sometimes by the same supervisor or supervisory group. To account for the intercorrelations between these repeated measurements, the method of Cook et al. (2010) was used. This involved using a linear mixed model to estimate a correlation matrix that appropriately accounts for the repeated measures, then performing the common factor analysis on this adjusted correlation matrix.

Item scores were converted to long format for analysis using a data step in SAS. The SAS procedure MIXED was then used to perform the linear mixed modelling using item score as the dependent variable and item type as the independent variable. Placement type (placement), academic status of the supervisor (academic), and compulsory nature of the placement (core) were included in the model as covariates in order to minimise bias due to missing item scores (Box D.1). The between-student correlation matrix was outputted from PROC MIXED. This was prepared for use as the input data for factor analysis by assembling it together with the mean and standard deviation of the item scores which are required by the factor analysis procedure. The sample number is also required and the method of determining the sample number used by Cook et al. (2010) was applied. Therefore the sample number used was the number of discrete placement-student pairs. This number was smaller than the actual number of observations, since students went to some placements more than once, but it was greater than the number of students. As Cook et al. (2010) explain, the appropriate method of calculating sample size in this type of analysis is debatable. Specifying an incorrect sample size does not alter estimates of eigenvalues or pattern and structure coefficients but does alter the standard errors calculated and thus statistical tests of model fit, number of factors to retain as well as the confidence intervals around pattern and structure coefficients.

For the purposes of comparison a Pearson correlation matrix of the raw data which was calculated using the SAS procedure CORR. Missing pairs of correlations were not included in the calculation but all available data was utilised. The dependency between repeated measures on the same student over time was ignored.

Box D.1: SAS code for preparation of the adjusted correlation matrix.

```
proc sort data=datasetname;
     by item studentID plmtdate;
run;
proc mixed data=datasetname;
     class studentID plmtdate placement academic core item ;
     model itemscore=item placement academic core
           academic*core /solution;
     repeated  / subject=studentID*plmtdate type=UN rcorr=3;
     ods output rcorr=rcorrdatasetname;
run;
```

Note. The PROC SORT command sorted the data by subject so that the MIXED procedure could determine the proper location of the repeated evaluations on students.

The effect of item type (item), placement type (placement), the academic status of the supervisor (academic), and the elective nature of the placement (core) on the item score (itemscore) was modelled in the model statement. The interaction of academic status of the supervisor and elective nature of the placement was also included (academic*core).

The repeated statement modelled the repeated measurements on each student over time, specified by placement date (plmtdate).

An unstructured covariance matrix for the repeated measurements on each student was specified with the type=UN command and this does not impose any predetermined structure on the within-student covariance matrix, allowing each variance and covariance to be different.

The R correlation matrix was outputted. This is the Pearson correlation matrix of the residuals calculated using maximum likelihood and thereby accounting for missingness. It is therefore not identical to a Pearson correlation matrix of the residuals. It represents the between-student correlations after accounting for the within-student correlations that are present because of the repeated evaluations for each student. The third R correlation matrix was chosen as the output with the rcorr=3 command. Although the R correlation matrices for each evaluation are the same, the first and second ones contained missing items and hence the correlation matrix output was incomplete for those and the third one was used instead.

Suitability of sample size for factoring

The minimum sample size required to have reasonable power and precision for exploratory factor analysis depends on characteristics of the data. Smaller sample sizes are more likely to be sufficient with increasing strength of communalities, number of variables loading on each factor, and strength of structure and pattern coefficients and as the presence of crossloading (which is when a variable has a substantial loading on more than one factor) decreases (Costello & Osborne, 2005). Rules of thumb are not appropriate (Schmitt, 2011) but guidelines based on the characteristics of the data suggest that sample sizes should be no smaller than 200 when communalities are in the range of 0.4-0.7 and there are at least 3 measured variables per factor, and at least 400 when communalities or the number of variables per factor is lower (Fabrigar & Wegener, 2012). The effective sample size based on the number of discrete placement-student pairs, as discussed above, was 3034 and is thus a good sample size.

Suitability of the matrix for factoring

The suitability of the data for common factor analysis was assessed using four different methods. Firstly the adjusted correlation matrix was examined to determine if sufficient correlations (over 0.3) were present to indicate the presence of common factors (B. Williams, Brown, & Onsman, 2012). Secondly Bartlett's test of sphericity was performed. This involves a chi squared test evaluating the significance of correlations in the sample matrix (Gorsuch, 1983). A significant finding (p<0.05) indicates the presence of one or more common factors. Thirdly, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was determined with values over 0.5 considered suitable for factor analysis and values of 0.8 ideal (Spicer, 2005; B. Williams et al., 2012). Lastly, partial correlations were also examined, with small values suggesting the presence of common factors and the suitability of the data for factor analysis (Spicer, 2005). All evaluations were performed using SAS PROC FACTOR.

Extraction and rotation of factors

Extraction of factors was conducted using maximum likelihood estimation in the SAS procedure FACTOR because it provides statistical tests of model fit which assist in determining the number of factors to retain, and because it provides confidence intervals around pattern and structure coefficients which assist in determining the significance of a variable's loading on a factor. Robust versions are not currently available for use in SAS unless raw item scores are used as the input data (as opposed to a correlation matrix) which would have precluded accounting for dependency and missingness.

Squared multiple correlations were used as the initial communality estimates for this study. These are a widely used because they are a lower bound to the population communality (Fabrigar & Wegener, 2012). The initial estimates are, however, of little importance when iterative methods such as maximum likelihood analysis are used to derive the final communality estimates because the final communalities converge on the same value irrespective of the starting value.

An oblique rotation method, direct oblimin (also known as quartimin), was chosen because correlation of factors was likely. The separate dimensions of the in-training evaluation were

related by the fact that all are aspects of overall competency in the veterinary workplace. Oblique rotations allow factors to be correlated but do not force them to be correlated if they are not (Fabrigar et al., 1999), and hence are ideal in this type of situation. Since different oblique rotation methods use different strategies for factor simplification, the analysis was also repeated using two other oblique rotation methods representing contrasting strategies. This has been recommended by several authors (Asparouhov & Muthén, 2009; Finch, 2013; Sass & Schmitt, 2010) because each strategy can lead to substantively different conclusions (Schmitt & Sass, 2011). However, for this data the conclusions were not different and the analysis is therefore not shown.

Determining the number of factors to retain

Determining the number of factors to retain for rotation in exploratory factor analysis has been described as the most challenging part of factor analysis (Fabrigar & Wegener, 2012). The goal is to find the number of factors that best represent the underlying constructs, acknowledging that models are not true representations, but, rather, are useful simplifications (MacCallum, 2003). Therefore we need to balance fit with simplicity. A guideline for achieving this balance is provided by Fabrigar and Wegener (2012). The number of factors retained should be that which gives a substantially better fit than one fewer factor, but not appreciably worse than one more factor, and provides interpretable factors that relate to constructs of interest.

Rotated factor solutions were therefore produced for different numbers of factors and examined to determine their interpretability and model fit. A number of procedures that have been used to help determine the number of factors to retain were also examined. Applying multiple procedures is generally recommended (Bandalos & Finney, 2010; Beavers et al., 2013; MacCallum, 2009; Ruscio & Roche, 2012). Agreement between procedures gives greater confidence in the suitability of the number of factors to retain (Ruscio & Roche, 2012), however the accuracy of these procedures varies according to characteristics of the data such as the sample size, the degree of non-normality, the number of factors, the number of variables per factor, the strength of factor loadings, and the strength of interfactor correlations (Garrido, 2012). Procedures were therefore chosen which were most likely to be accurate given the characteristics of this data determined by a preliminary analysis (Table D.1). These

characteristics were a large sample size; items assessed using a 5-option Likert-type response format; mild-moderately skewed and kurtotic distribution; few factors; few variables per factor; strong factor loadings; and strong interfactor correlations. Where procedures indicated different numbers of factors to retain, the greatest consideration was given to procedures reported to be more accurate for this type of data, and to the usefulness and interpretability of the solution. The procedures used were limited to those that could be computed with Excel or SAS or using freely available software. Some procedures were also included for comparison purposes, even though they are known to be less accurate, because they are commonly reported in the literature on factor analysis of in-training evaluations. Thus the procedures undertaken were the eigenvalues greater than one rule, the scree test, parallel analysis, minimum average partial method, chi squared test of model fit and chi squared difference test, Tucker-Lewis index (TLI), root mean squared error of approximation (RMSEA), evaluation of residuals and partial correlations and interpretation of the importance of factors. The use of the variance explained was considered but was not performed.

Table D.1: Expected performance of various procedures used to determine the number of factors to retain in data with large sample size, items assessed using a 5-option Likert-type response format, mild-moderately skewed and kurtotic distribution, few factors, few variables per factor, strong factor loadings, and strong interfactor correlations.

Method	Likely to underfactor with this data	Likely to overfactor with this data
Eigenvalues > 1		X
Scree test		X
Parallel analysis	X	
Minimum average partial	X	
Residual size		X
Chi squared test of model fit		X
Chi squared difference test		X
TLI		X
RMSEA	X	X

Abbreviations: TLI: Tucker-Lewis index; RMSEA: root mean squared error of approximation

Eigenvalues greater than one rule

The eigenvalues greater than one rule is simple to apply: the number factors is given by the number of eigenvalues greater than 1.0. The appealing logic of this rule is that only factors which account for more variance than any one variable itself are retained (Velicer et al., 2000). However this logic only makes sense if the rule is applied to the eigenvalues obtained during principal components analysis and not those produced during common factor analysis. This is because in principal components analysis a value of 1 represents the variance shared equally by each variable, however this is not the case in common factor analysis and the value of 1 is meaningless. Performing a principal components analysis to obtain eigenvalues with which to determine the number of common factors to retain in common factor analysis could be done, but does not make theoretical sense (Fabrigar & Wegener, 2012). Other theoretical issues with the use of this rule include that it assumes a precise estimation of the eigenvalues such that a strict cut-off of 1.0 can be applied (B. Thompson, 2004). The use of confidence intervals around eigenvalues has been suggested to alleviate this difficulty (Larsen & Warne, 2010) but in application did not improve the accuracy of this rule (Warne & Larsen, 2014).

The accuracy of the eigenvalues greater than one rule has been widely reported with studies showing that the rule has a tendency to either underestimate or overestimate the number of factors depending on the characteristics of the data. Fabrigar et al. (1999) sum up the research by saying that “in fact, we know of no study of this rule that shows it to work well” (p. 278). In simulated ordinal data the accuracy of the eigenvalues greater than one rule depended to a large extent on the number of variables per factor, the size of factor loadings, the number of factors, the sample size, as well as the skewness (Garrido, 2012). Although large sample size, small number of variables and high factor loadings, such as are present in my data, tended to improve the accuracy of the rule, in these conditions it was still found inaccurate at least 40% of the time and tended to overestimate the number of factors by 0.8-2.02 factors. The conclusion of numerous reviewers is that this rule should not be used (for example Baglin, 2014; Costello & Osborne, 2005; Fabrigar et al., 1999; Gaskin & Happell, 2014; Gorsuch, 1983; Preacher & MacCallum, 2003; Velicer et al., 2000; Zwick & Velicer, 1986) and it was only included in this analysis for comparison with the numerous studies of in-training evaluation that have used it.

MacCallum (2009) observed that the eigenvalues greater than one rule is appropriately applied to the observed (unreduced) correlation matrix rather than the reduced matrix (which is the

matrix with the diagonals reduced to communalities), however eigenvalues from the observed matrix are not available in the output from the SAS procedure FACTOR. Therefore the reduced matrix had to be used. The preliminary eigenvalues (those calculated using the initial estimate of communality) were used and the number of factors to retain was determined by the number of eigenvalues greater than 1.0.

Scree test

The scree test involves examining a plot of the eigenvalues to separate those which are contributing significantly to the variance and those that make only a minor contribution (Gorsuch, 1983). Because the largest eigenvalues are extracted first, the data values begin high and then become lower. The rate of change is initially rapid but then slows. The shape of the data when plotted was compared to a mountain with scree at the bottom by Cattell (1966). To perform the test, a straight line is drawn through the smaller eigenvalues that appear in the later data points. The number of factors is determined to be the number of data points for the initial eigenvalues that lie above this line. This method can be difficult to apply when there is not one single clear break (Velicer et al., 2000) or when the smaller eigenvalues can be summarised by more than one straight line that encompasses greater or fewer data points. Velicer et al. (2000) reviewed studies of the accuracy of the scree test and reported that many found it to be quite accurate but that it was less accurate when there were low communalities, smaller sample sizes, low factor loading or crossloading. Under these conditions there is a tendency for it to overfactor.

The scree test can be performed on either the observed correlation matrix or the reduced correlation matrix (MacCallum, 2009). Fabrigar and Wegener (2012) recommends using the reduced correlation matrix in common factor analysis because it represents the common factor variance. Like the eigenvalues greater than one rule, the scree test was therefore performed on the eigenvalues from the preliminary reduced matrix based on the initial communality estimates. This is the default in SAS PROC FACTOR.

Parallel analysis

Parallel analysis is widely recommended as being one of the more accurate methods for determining the number of factors to retain (Fabrigar & Wegener, 2012; Garrido, Abad, & Ponsoda, 2013; Hayton, Allen, & Scarpello, 2004; Schmitt, 2011; Velicer et al., 2000). The aim

of parallel analysis is to determine the number of factors that account for more variance than could occur by chance. It is performed by comparison of random datasets to the sample dataset (Fabrigar & Wegener, 2012). Possible variations on this method involve generating the random data eigenvalues using common factor analysis rather than principal components analysis; using polychoric matrices rather than Pearson correlation matrices; and combining the random datasets using the 95th percentile values rather than the means.

Some authors have recommended that parallel analysis should be performed using a common factor method, rather than a principal components method, when common factor analysis is being undertaken (Fabrigar & Wegener, 2012). However, while it makes theoretical sense to use the same method for estimating the number of factors as for the actual analysis, there are several reasons why a principal components method is preferred for parallel analysis. One is that in common factor analysis the communalities used to derive the eigenvalues are only estimates and therefore it is not appropriate to use them as strict cut-points for comparison with random values (Garrido et al., 2013). Another, is that the variables in the random datasets are uncorrelated and therefore have no common variance. This is therefore inconsistent with the common factor model (Garrido et al., 2013). In addition, a common factor approach results in an increasing tendency for more of the sample eigenvalues to be larger than the randomly generated eigenvalues, indicating more factors, as sample size increases (Timmerman & Lorenzo-Seva, 2011). In contrast, Garrido et al. (2013) has shown that parallel analysis using a principal components method for generating both the sample and random eigenvalues has better accuracy in data with high interfactor correlations, large factor loadings, few variables per factor, large sample size, and mild to moderate skewness, such as my data.

It is important that during parallel analysis the same methods are used for generating both the sample and the random eigenvalues, even if a different estimation is then used to examine the factor solution once the number of factors has been determined (Garrido et al., 2013). Parallel analysis was therefore undertaken by performing principal components analysis on both the sample and random datasets, using the adjusted Pearson correlation matrix that accounted for the repeated measures and the missing data. The SAS macro by O'Connor (2000) was used, to generate 1000 random eigenvalues for comparison and combined using the mean criterion, which has been shown to be more accurate than the 95th percentile criterion when factors are highly correlated, as in my data (Cho, Li, & Bandalos, 2009; Crawford et al., 2010; Garrido et al., 2013).

Velicer's minimum average partial (MAP)

The minimum average partial (MAP) method of determining the number of factors was proposed by Velicer (1976). It involves performing a principal components analysis and examining the correlations that remain after sequentially removing the variance attributable to each factor (called partial correlations). The correlations will decrease as common variance is removed, but once only unique variance remains, further extraction of variance will cause the correlations to rise (Velicer et al., 2000). The point at which all common variance is removed can therefore be determined by looking for the point at which this rise begins. This determines the number of common factors. Thus although the method utilises principal components analysis it is based on a common factor model (Velicer, 1976). Several studies have indicated that MAP is not as accurate as parallel analysis in many circumstances but more accurate than other methods such as the eigenvalues greater than one rule or scree test (Garrido, Abad, & Ponsoda, 2011; Ruscio & Roche, 2012; Zwick & Velicer, 1986). MAP tends to underestimate the number of factors, especially when there are low factor loadings and small numbers of variables per factor (Garrido et al., 2011).

MAP was performed using the SAS macro produced by O'Connor (2000). The version utilising the squared partial correlations was used as it has been shown be more accurate than the version that uses a power of four, in a simulation study of ordinal data (Garrido et al., 2011).

Measures of model fit

Model fit is used to determine the number of factors by comparing the fit after different numbers of factors have been extracted. A good fit means that the model is a good summary of the data, and the observed correlations are close to the correlations predicted by the model. However, as previously discussed, the model fit needs to be balanced with model simplicity. Looking for good model fit involves examining the residuals, which are the differences between the model and the data. There are several different methods of examining residuals.

Residual covariances and partial correlations

The output from exploratory factor analysis includes a matrix of residual covariances, and a standardised version of the same, which is called the partial correlation matrix. The lower the off-diagonal residuals, the more closely the model approximates the data. Generally a model is

considered a good fit if most of the off-diagonal residuals are less than 0.05 (Basto & Pereira, 2012; R. B. Kline, 2013; B. Thompson, 2004). Values above 0.10 are considered evidence of poor fit (R. B. Kline, 2013). For partial correlations, as the values are bounded by ±1, values close to zero indicate a good model fit. It is quite helpful to consider the pattern of off-diagonal residuals and partial correlations to see whether they are similar in value. If a subset is higher than the others, it can indicate the need for another factor explaining those variables (Gorsuch, 1983; R. B. Kline, 2013). The residual covariance matrix and partial correlation matrix are included in the standard output of the SAS procedure FACTOR. The absolute values of the off-diagonal residuals as well as their mean, standard deviation, and sum were considered in determining the best model fit. The absolute values of the off-diagonal partial correlations were also considered. To aid this, patterns of partial correlations were examined by constructing a three-dimensional graph of the partial correlation matrix using the same method as for the adjusted matrix.

Chi squared test of model fit and chi squared difference test

Under maximum likelihood estimation the model fit can be evaluated statistically using chi squared tests. Two different hypotheses can be examined (Fabrigar & Wegener, 2012). One is that the number of factors extracted is sufficient to give a good model fit. The other, which utilises the chi squared difference test, is that a model is a better fit than one with one less factor. Both of these tests have a good theoretical basis, but their usefulness is limited by the fact that they are overly sensitive when sample size is large. Insubstantial differences between models with different numbers of factors can be statistically significant, leading to rejection of well-fitting models. Thus, both these tests tend to overestimate the number of factors. In addition the chi-squared tests have a tendency to overfactor when data are not normally distributed (T. A. Brown, 2006) and with higher factor loadings (Barendse, Oort, & Timmerman, 2014). In simulated ordinal data, Ruscio and Roche (2012) found that the chi squared test of model fit specified too many factors in 41% of cases. The chi squared difference test tends to select even more factors than the chi squared test of model fit (Barendse et al., 2014).

The chi squared test and chi squared difference test were estimated using SAS. The SAS procedure FACTOR gives the chi squared test statistics and significance levels in the output. The chi squared difference test was then performed using the formula provided by Fabrigar and Wegener (2012) in a data step in SAS. Each test was performed on models with extraction

of 1, 2, 3, and 4 factors. The number of factors was determined to be the fewest that produced a non-significant *p* value ($p>0.05$).

Descriptive fit indices

A variety of model fit indices have been developed which aim either to balance the model fit with model complexity or to measure the model fit relative to other models rather than absolutely, similarly to the chi squared difference test (T. A. Brown, 2006). They also aim to reduce the inaccuracies seen with large sample size that occur with the chi squared statistic (Hu & Bentler, 1999). T. A. Brown (2006) suggests that researchers should report at least one model fit index of each type. Only two descriptive fit indices will be considered here because of their availability and known behaviour in data similar to mine. These are the Tucker-Lewis index (TLI) and the root mean square error of approximation (RMSEA).

The TLI is a comparative fit index which also compensates for the complexity of the model and therefore combines both types of model fit index (T. A. Brown, 2006). It penalises more complex models (those with more factors) by measuring model fit per degree of freedom rather than absolutely. It is derived from the chi squared statistic and therefore has a tendency to overfactor rather than underfactor. However, in a simulation study of ordinal data it performed well, especially for data with large sample size, where it correctly determined the factor structure in 91% of occasions (Garrido, 2012). This level of accuracy was second only to parallel analysis in these conditions. Garrido (2012) concluded from his study that when both the TLI and parallel analysis agree about the number of factors to retain, researchers can have great confidence in the result. The TLI tends not to be substantially influenced by the factor loadings, number of factors, or the interfactor correlation (Garrido, 2012). TLI was calculated for models with extraction of 1, 2, 3, and 4 factors using SAS PROC FACTOR. The number of factors was determined to be the fewest that produced a TLI of greater than 0.95. This cut-point was recommended by Hu and Bentler (1999) and was used in the evaluations by Garrido (2012).

Like the TLI, the RMSEA takes into account model complexity by using degrees of freedom as a divisor. Therefore, it will tend to favour models with fewer factors unless there is a sufficiently better fit from a model with more factors. It allows for some discrepancy between the model and the data and therefore provides a measure of whether a model is a reasonable fit rather

than a perfect fit, in contrast to the chi squared statistic (T. A. Brown, 2006). An acceptable model fit is often defined as models with values less than 0.05-0.06 (Fabrigar & Wegener, 2012; Hu & Bentler, 1999; R. B. Kline, 2013; MacCallum, Browne, & Sugawara, 1996), with zero indicating a perfect fit (Fabrigar & Wegener, 2012), however Garrido (2012) found that a cut-point of 0.02 gave much greater accuracy in simulated ordinal data. Confidence intervals can be calculated for RMSEA and should be reported (T. A. Brown, 2006). These assist researchers in determining whether a model with more factors is substantially better than another. Overlapping confidence intervals would suggest that there is no advantage in the more complex model (Fabrigar & Wegener, 2012). In simulated ordinal datasets RMSEA is quite accurate, correctly determining the number of factors in 84% of cases when sample size is high and when a cut-point of 0.02 is used (Garrido, 2012). It was less accurate than parallel analysis and TLI, but more accurate than eigenvalues greater than one or MAP. Sample size and the degree of non-normality had the greatest influence on its accuracy in ordinal data (Garrido, 2012; Hutchinson & Olmos, 1998). High factor loadings and interfactor correlations also influence the accuracy of RMSEA to a lesser degree (Garrido, 2012). RMSEA was calculated for models with extraction of 1, 2, 3, and 4 factors using the formula given by Fabrigar and Wegener (2012) from the chi squared values output from SAS PROC FACTOR. Confidence intervals for each RMSEA were calculated using FITMOD software (Browne, 1992). The number of factors was determined to be the fewest that produced an RMSEA of less than 0.02.

Variance explained

The percentage of variance explained by the factors is of interest because it indicates how much of the variance of the data the model is accounting for. In terms of this study, where the goal is to determine the factors accounting for the scores awarded on the in-training evaluation, accounting for as much variance as possible is appropriate. In common factor analysis it is the percent of common variance explained that should be considered, rather than the total variance explained, since the total variance is not being modelled (Gorsuch, 1983). However, it is not possible to calculate the common variance explained as it is set at 100% through iteration in order to estimate the factors (SAS Institute Inc., 2014). Therefore, variance explained was not used as a criterion to judge the number of factors to retain.

Interpretation of the importance of the factors

An important part of determining the number of factors to retain is to examine the structure and pattern coefficients to determine the importance of each factor in the solution. Aspects

considered include the presence of specific factors[5], the strength and significance of factor loadings, the number of variables loading on a factor, and the presence of cross-loading[6]. A solution was sought that contained the number of factors that reflected the data and were most interpretable and useful for the purposes of the analysis.

A factor might not be useful if it has fewer than three variables loading strongly on it (Gorsuch, 1983). A common factor must, by definition, represent at least two variables that are correlated, and the presence of a third variable strengthens the solution. However Gorsuch points out that for data with few variables, even factors with only one substantial loading and two or three minor loadings may be of importance (Gorsuch, 1983). When as few as one variable loads on a factor, that factor does not share its variance with any other variable and is therefore not a common factor but a specific (Gorsuch, 1983). It offers no more value to the solution than the original variable did and is therefore not any more useful than the variable itself. Another situation occurs when a factor contains only variables that also load strongly on other variables (Gorsuch, 1983). Such a factor may be difficult to define as it has no variables that uniquely characterise it (Gorsuch, 1983) and, since it does not have a major influence on any variable, it could not be considered useful (Fabrigar & Wegener, 2012).

To assess the strength of a factor loading, arbitrary cut-points have been proposed, commonly ranging from 0.3-0.5 (Henson & Roberts, 2006) with 0.3 often used (P. Kline, 1994; Tabachnick & Fidell, 2013). However, a loading of just over 0.3 may or may not be significantly different from a loading of zero, depending on the confidence intervals of the estimate. Combined with the magnitude of the loadings, confidence intervals help to establish whether a loading is both substantial and significant, and therefore provide a better way to determine the importance of a variable to a factor (Schmitt & Sass, 2011). The provision of confidence intervals is therefore an important advantage of maximum likelihood factor estimation.

It is helpful to extract more factors than are likely to be necessary and then examine each solution (Gorsuch, 1983). As more factors are extracted from a dataset, specific factors start to appear in the solutions. Common factors tend to initially be stable when more factors than

[5] Specific factors are those which load on only one variable (item) (Gorsuch, 1983).
[6] Cross-loading occurs when a variable has significant loadings on more than one factor (Costello & Osborne, 2005).

necessary are extracted, but eventually start to break up into specifics themselves. When as many factors as variables are extracted, each factor is usually a specific. Examining this pattern of the appearance of specifics as more factors are extracted can help determine the appropriate number of factors.

Therefore, in this study, the factor structure with extraction of 1, 2, 3, 4, and 5 factors was examined. The 99% confidence intervals around both the structure and pattern coefficients were used to determine whether loadings were significantly different to zero, significantly different to 0.4 (an arbitrary moderate loading) and significantly different to 0.6 (an arbitrary strong loading). Confidence intervals are generated in SAS PROC FACTOR when maximum likelihood estimation is used.

Appendix E:
Additional results for factor analysis

Table E.1: Adjusted correlation matrix used for the main factor analysis.

	Animal	Technic	Exam	History	K_use	Kknow	SDL	Assignd	Judgemt	Partic	Comclin	Client
Animal	1	0.64	0.56	0.55	0.50	0.44	0.40	0.47	0.48	0.45	0.45	0.49
Technic	0.64	1	0.59	0.59	0.55	0.50	0.42	0.47	0.47	0.45	0.44	0.49
Exam	0.56	0.59	1	0.71	0.58	0.55	0.44	0.46	0.44	0.43	0.40	0.43
History	0.55	0.59	0.71	1	0.59	0.56	0.47	0.47	0.49	0.48	0.45	0.48
K_use	0.50	0.55	0.58	0.59	1	0.73	0.54	0.45	0.49	0.48	0.46	0.49
Kknow	0.44	0.50	0.55	0.56	0.73	1	0.52	0.44	0.46	0.44	0.40	0.42
SDL	0.40	0.42	0.44	0.47	0.54	0.52	1	0.48	0.51	0.53	0.47	0.45
Assignd	0.47	0.47	0.46	0.47	0.45	0.44	0.48	1	0.67	0.64	0.58	0.53
Judgemt	0.48	0.47	0.44	0.49	0.49	0.46	0.51	0.67	1	0.67	0.65	0.59
Partic	0.45	0.45	0.43	0.48	0.48	0.44	0.53	0.64	0.67	1	0.67	0.59
Comclin	0.45	0.44	0.40	0.45	0.46	0.40	0.47	0.58	0.65	0.67	1	0.71
Client	0.49	0.49	0.43	0.48	0.49	0.42	0.45	0.53	0.59	0.59	0.71	1

Table E.2: Initial and final communality estimates for each item in the in-training evaluation. Initial estimates are squared multiple correlations and ranged between 0.43 and 0.62. Final communality estimates ranged between 0.47 and 0.76.

Item	Communality estimates	
	Initial	Final
Animal	0.50	0.55
Technic	0.54	0.60
Exam	0.58	0.67
History	0.60	0.66
K_use	0.62	0.75
Kknow	0.58	0.71
SDL	0.43	0.47
Paassig	0.55	0.65
Judgemt	0.60	0.68
Paparti	0.60	0.67
Comclin	0.62	0.76
Client	0.57	0.68

Additional factor matrices

Three separate factor matrices are produced from oblique rotation and Gorsuch (1983) recommends that all three are reported to aid interpretation by readers. The factor pattern matrix was presented in Table 5.4 (page 130) with the other results. The factor pattern coefficients represent the unique contribution of the factor to predict the variable (item) (Gorsuch, 1983). The factor structure matrix is presented in Table E.3. Factor structure coefficients represent the total contribution of a variable (an item) to a factor, including the unique contribution of the item as well as indirect contributions through the correlations between factors (Gorsuch, 1983). They are therefore greater than the corresponding pattern coefficients. The third matrix is the reference structure matrix Table E.4. The reference structure correlations are similar to the pattern coefficients in representing the unique contribution of a factor to a variable, but are correlations rather than standardised regression coefficients (Gorsuch, 1983).

Table E.3: Factor structure coefficients produced by extraction of 1-, 2-, 3-, and 4-factor solutions on the adjusted Pearson correlation matrix, using maximum likelihood estimation and direct oblimin rotation.

Item domains	Items	1 factor	2-factor solution		3-factor solution			4-factor solution			
Clinical skills	Animal	0.69	0.69	0.58	0.73	0.49	0.57	0.73	0.51	0.53	0.48
	Technic	0.71	0.74	0.58	0.77	0.55	0.57	0.77	0.57	0.53	0.47
	Exam	0.70	0.80	0.53	0.81	0.63	0.52	0.81	0.64	0.51	0.37
	History	0.74	0.81	0.58	0.81	0.64	0.57	0.80	0.65	0.56	0.43
Knowledge	K_use	0.73	0.77	0.59	0.69	0.83	0.58	0.69	0.86	0.54	0.45
	Kknow	0.69	0.74	0.54	0.64	0.86	0.52	0.64	0.84	0.51	0.36
	SDL	0.66	0.61	0.61	0.54	0.61	0.61	0.53	0.62	0.60	0.47
Professional attitude	Assignd	0.72	0.60	0.75	0.58	0.49	0.75	0.58	0.50	0.80	0.57
	Judgemt	0.76	0.61	0.81	0.59	0.52	0.81	0.58	0.53	0.82	0.66
	Partic	0.75	0.59	0.81	0.57	0.50	0.81	0.55	0.52	0.81	0.68
Communica-tion	Comclin	0.73	0.55	0.83	0.54	0.45	0.83	0.53	0.47	0.74	0.85
	Client	0.72	0.59	0.76	0.59	0.48	0.76	0.59	0.50	0.66	0.80

Note. Shaded cells indicate coefficients that are significantly greater or equal to 0.6 (α=0.01). Those with values between 0.6-0.7 are lightly shaded and those over 0.7 are more darkly shaded.

Table E.4: Reference structure correlations produced by extraction of 1-, 2-, 3-, and 4-factor solutions on the adjusted Pearson correlation matrix, using maximum likelihood estimation and direct oblimin rotation.

Item domains	Items	1 factor	2-factor solution		3-factor solution			4-factor solution			
Clinical skills	Animal	0.69	0.39	0.11	0.43	-0.06	0.10	0.44	-0.05	0.02	0.07
	Technic	0.71	0.46	0.06	0.44	0.00	0.05	0.45	0.01	-0.01	0.06
	Exam	0.70	0.61	-0.08	0.50	0.07	-0.08	0.49	0.06	0.01	-0.08
	History	0.74	0.56	-0.01	0.44	0.08	0.00	0.43	0.08	0.04	-0.03
Knowledge	K_use	0.73	0.50	0.04	0.09	0.47	0.05	0.06	0.51	-0.04	0.08
	Kknow	0.69	0.50	0.01	0.02	0.57	0.00	0.02	0.53	0.02	-0.03
	SDL	0.66	0.24	0.25	0.00	0.27	0.27	-0.01	0.27	0.20	0.03
Professional attitude	Assignd	0.72	0.07	0.46	0.06	0.02	0.47	0.07	-0.02	0.46	-0.03
	Judgemt	0.76	0.03	0.53	0.01	0.04	0.54	0.02	0.03	0.40	0.08
	Partic	0.75	0.00	0.56	-0.02	0.04	0.57	0.00	0.04	0.37	0.14
Communica-tion	Comclin	0.73	-0.07	0.62	-0.02	-0.02	0.62	-0.01	0.04	0.13	0.45
	Client	0.72	0.05	0.49	0.08	-0.01	0.49	0.10	0.05	0.01	0.45

Note. Shaded cells indicate coefficients that are greater or equal to 0.2. Those with values between 0.2-0.3 are lightly shaded, those between 0.3-0.4 are more darkly shaded and those over 0.4 are the most darkly shaded.

Appendix F:
Additional details for regression analysis

Additional details for regression method

Box F.1 shows the SAS code used in the final model for the regression analysis in Phase 3.

Box F.1: SAS code for generalised linear mixed modelling of overall grade with various independent variables.

```
proc sort data= datasetname;
    by studentID week ;
run;
proc glimmix data= datasetname method=laplace empirical=mbn
              plots=oddsratio (logbase=e);
    class studentID academic (ref='1') placement (ref='EXT');
    model overall (descending)= Factor1 Factor2 Factor3 Factor4
          academic placement GPA Factor2*placement Factor3*placement
          /link=cumlogit dist=multinomial ddfm=bw  solution cl
          oddsratio (label) ;
    random intercept week/subject=studentID type=sp(pow)(week) ;
    covtest 'week*studentID significant? ' 0. /cl est;
run;
```

Note. The PROC SORT command sorted the data by subject so that the GLIMMIX procedure could determine the proper location of the repeated evaluations on students.

In the GLIMMIX procedure the Laplace approximation was specified and the Morel Bokossa Neerchal (MBN) empirical estimator was used for bias correction. The class statement specified the categorical variables and the reference values for academic and placement.

The model statement models the effects of the specified variables and any quadratic terms and interactions on the overall grade awarded. The ordinal nature of overall grade was modelled by including the cumulative logit link function and specifying the multinomial distribution. The between within method of calculating the denominator degrees of freedom was used. The descending option specified that the probability of higher grades would be modelled.

The random statement modelled the repeated measurements on each student over time. The type= command specified the structure of the imposed covariance matrix for the repeated measurements on each student. The spatial power (sp(pow)(week)) covariance structure was used after determining that it gave the best model fit.

The covtest statement provides a test of the significance of the random effects (of the repeated measurements on each student over time).

Additional results from the regression analysis

These tables present descriptive statistics for data used in the regression analysis of Phase 3 and the parameter estimates for both the main model and the full model used for analysis.

Table F.1: Distribution of overall grade according to covariate grouping.

Placement type	N	N Miss	Mean	SD	Min	Median	Max
EQ1	370	0	7.7	1.37	0	8	10
EQ2	28	0	8	0	8	8	8
EXT	1057	0	9.01	1.25	0	10	10
OT4	35	0	8.34	1.49	4	8	10
PA1	346	0	7.74	1.12	0	8	10
PA2	249	0	8.63	1.42	4	8	10
PA3	233	0	9.12	1.14	4	10	10
PA4	198	0	8.3	1.27	4	8	10
PA5	115	2	8.4	1.43	0	8	10
PA6	98	147	5.94	0.61	0	6	6
PA7	2	0	8	0	8	8	8
SA1	281	0	7.57	1.85	0	8	10
SA2	198	0	8.03	0.74	4	8	10
SA3	59	0	7.76	1.18	6	8	10
SA4	18	0	8.33	1.85	4	8	10
SA5	30	0	8	0	8	8	8

Academic status	N	N Miss	Mean	SD	Min	Median	Max
academic	1315	147	7.83	1.56	0	8	10
nonacademic	2002	2	8.65	1.31	0	8	10

Year	N	N Miss	Mean	SD	Min	Median	Max
2012	1649	49	8.37	1.51	0	8	10
2013	1668	100	8.29	1.42	0	8	10

Note. Overall grade was converted to a numerical score as follows: excellent = 10, good = 8, satisfactory = 6, marginal = 4, fail = 0. Abbreviations: N: number of items scored; N Miss: number of items with missing scores; SD: standard deviation; Min: minimum; Max: maximum.

Table F.2: Parameter estimates and their standard errors and 95% confidence intervals for the main model.

Effect	Level	Estimate	SE	DF	P value	95%CI	
						Lower	Upper
Intercept	10	-1.29	0.34	192	0.0002	-1.97	-0.62
Intercept	8	6.55	0.80	192	<.0001	4.97	8.12
Intercept	6	12.48	1.23	192	<.0001	10.05	14.91
Intercept	4	16.75	1.68	192	<.0001	13.44	20.06
Factor1 (clinical skills)		1.30	0.21	1392	<.0001	0.88	1.72
Factor2 (knowledge)		0.86	0.22	1392	<.0001	0.44	1.29
Factor3 (professional attitude)		2.14	0.25	1392	<.0001	1.64	2.63
Factor4 (communication)		0.56	0.21	1392	0.0062	0.16	0.97
academic	0	0.03	0.33	196	0.929	-0.62	0.67
academic	1	0.00	.	.	.	.	.
placement	EQ1	-1.96	0.41	818	<.0001	-2.77	-1.15
placement	OT4	-1.07	0.82	818	0.194	-2.68	0.54
placement	PA1	-2.54	0.54	818	<.0001	-3.59	-1.49
placement	PA2	-0.51	0.36	818	0.149	-1.21	0.18
placement	PA4	0.95	0.5	818	0.061	-0.04	1.94
placement	PA5	1.41	0.88	818	0.111	-0.33	3.15
placement	PA6	-1.47	0.62	818	0.018	-2.69	-0.25
placement	SA1	-1.66	0.49	818	0.0007	-2.62	-0.7
placement	SA2	-0.62	0.97	818	0.525	-2.51	1.28
placement	SA3	-0.32	0.33	818	0.340	-0.97	0.34
placement	SA4	-1.29	0.9	818	0.152	-3.06	0.48
placement	EXT	0	.	.	.	.	.
GPA		0.10	0.08	192	0.247	-0.07	0.27

Note. Abbreviations: SE: standard error; DF: degrees of freedom; 95%CI: 95% confidence interval. Statistically significant estimates are shaded. This model does not include interaction terms.

Table F.3: Parameter estimates and their standard errors and 95% confidence intervals for the regression model.

Effect	Level	Estimate	SE	DF	P value	95%CI	
						Lower	Upper
Intercept	10	-1.37	0.46	192	0.003	-2.28	-0.47
Intercept	8	6.90	2.59	192	0.008	1.79	12.02
Intercept	6	13.66	4.42	192	0.002	4.94	22.38
Intercept	4	18.57	6.22	192	0.003	6.30	30.83
Factor1 (clinical skills)		1.43	0.32	1372	<.0001	0.80	2.05
Factor2 (knowledge)		0.21	0.78	1372	0.790	-1.33	1.74
Factor3 (professional attitude)		2.71	0.76	1372	0.000	1.21	4.21
Factor4 (communication)		0.57	0.23	1372	0.016	0.11	1.02

Effect	Level	Estimate	SE	DF	P value	95%CI	
						Lower	Upper
academic	0	0.05	0.70	196	0.945	-1.33	1.43
academic	1	0	.	.	.	.	.
placement	EQ1	-2.05	1.16	819	0.078	-4.34	0.23
placement	OT4	-1.23	1.76	819	0.485	-4.69	2.23
placement	PA1	-2.6	1.45	819	0.072	-5.44	0.24
placement	PA2	-0.57	0.48	819	0.232	-1.51	0.37
placement	PA4	1.18	0.8	819	0.138	-0.38	2.74
placement	PA5	1.51	0.91	819	0.098	-0.28	3.3
placement	PA6	-1.91	1.09	819	0.079	-4.04	0.22
placement	SA1	-1.47	0.84	819	0.080	-3.12	0.18
placement	SA2	-0.95	1.01	819	0.348	-2.93	1.03
placement	SA3	0	0.7	819	0.999	-1.37	1.37
placement	SA4	-2.09	1.59	819	0.190	-5.21	1.03
placement	EXT	0	.	.	.	.	.
GPA		0.13	0.18	192	0.489	-0.23	0.49
Factor2-placement interaction	EQ1	0.86	0.88	1372	0.327	-0.86	2.58
Factor2-placement interaction	OT4	1.93	1.44	1372	0.182	-0.9	4.76
Factor2-placement interaction	PA1	1.61	0.51	1372	0.002	0.6	2.61
Factor2-placement interaction	PA2	0.31	0.84	1372	0.708	-1.33	1.95
Factor2-placement interaction	PA4	-0.38	0.98	1372	0.697	-2.31	1.54
Factor2-placement interaction	PA5	-1.41	2.2	1372	0.521	-5.73	2.9
Factor2-placement interaction	PA6	0	.	.	.	.	.
Factor2-placement interaction	SA1	1.31	0.54	1372	0.015	0.25	2.38
Factor2-placement interaction	SA2	-3.99	2.05	1372	0.051	-8.01	0.02
Factor2-placement interaction	SA3	0.51	1.04	1372	0.627	-1.54	2.55
Factor2-placement interaction	SA4	-1.38	1.56	1372	0.379	-4.44	1.69
Factor2-placement interaction	EXT	0	.	.	.	.	.
Factor3-placement interaction	EQ1	-0.52	0.44	1372	0.243	-1.39	0.35
Factor3-placement interaction	OT4	-1.93	1.81	1372	0.285	-5.47	1.61
Factor3-placement interaction	PA1	-1.95	0.76	1372	0.011	-3.45	-0.46
Factor3-placement interaction	PA2	-0.41	0.83	1372	0.624	-2.03	1.22
Factor3-placement interaction	PA4	-1.83	1.28	1372	0.153	-4.34	0.68
Factor3-placement interaction	PA5	-0.76	2.86	1372	0.789	-6.36	4.84
Factor3-placement interaction	PA6	0	.	.	.	.	.
Factor3-placement interaction	SA1	-0.61	0.51	1372	0.228	-1.61	0.38
Factor3-placement interaction	SA2	2.96	1.95	1372	0.129	-0.87	6.79
Factor3-placement interaction	SA3	0.2	0.63	1372	0.756	-1.05	1.44
Factor3-placement interaction	SA4	3.84	1.6	1372	0.016	0.71	6.97
Factor3-placement interaction	EXT	0	.	.	.	.	.

Note. Abbreviations: SE: standard error; DF: degrees of freedom; 95%CI: 95% confidence interval. Statistically significant estimates are shaded.

Appendix G:
Ethics Committee letter of approval

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA

10 January 2014

Elizabeth Norman
IVABS
PN412

Dear Elizabeth

Re: HEC: Southern B Application – 13/94
Placement supervisor conceptions of veterinary student performance

Thank you for your letter dated 9 January 2014.

On behalf of the Massey University Human Ethics Committee: Southern B I am pleased to advise you that the ethics of your application are now approved. Approval is for three years. If this project has not been completed within three years from the date of this letter, reapproval must be requested.

If the nature, content, location, procedures or personnel of your approved application change, please advise the Secretary of the Committee.

Yours sincerely

Dr Jill Wilkinson, Acting Chair
Massey University Human Ethics Committee: Southern B

cc Dr Peter Rawlins
Institute of Education
PN500

Dr Linda Leach
Institute of Education
PN500

A/Prof Sally Hansen, Director
Institute of Education
PN500

Mrs Roseanne MacGillivray
Institute of Education
PN500

Massey University Human Ethics Committee
Accredited by the Health Research Council
Research Ethics Office
Massey University, Private Bag 11222, Palmerston North 4442, New Zealand **T** +64 6 350 5573 +64 6 350 5575 **F** +64 6 350 5622
E humanethics@massey.ac.nz animalethics@massey.ac.nz gtc@massey.ac.nz **www.massey.ac.nz**

Appendix H: Request for access – Director, Student Management, Massey University

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

13th January, 2014

Dr
Director - Student Management
Massey University

Regarding: permission to access student records for research

Dear Dr

I am writing to you in your capacity of guardian of the databases that contain student information to request access to anonymised student's records for research purposes, where explicit consent for access has not been sought from the students.

The students involved are those who studied in the fifth (final) year of the Bachelor of Veterinary Science programme in 2011, 2012 and 2013. The records involved are the electronic records of assessment results including the final scores and component scores for the end of fourth year and the end of fifth year.

The reason for not seeking consent from students to use their records for research purposes is that the research would be harmed if not all students consented. A retrospective sample is an efficient way to collect a suitable sample size of data (3 years' worth) to allow valid quantitative analysis. All (or almost all) of the students involved will have already graduated and left the University. Contacting all 300 for consent would be difficult or impossible. Removal of the results of ex-students who are not contactable, do not respond to requests for consent or who do not give consent may significantly skew the dataset and affect the conclusions reached because their results may differ from those of students who are contactable and do give consent.

The purpose of this research is to investigate the process of judgement of the complex performance of veterinary students during their work in professional contexts in the Veterinary Teaching Hospital and external sites. The assessment of students in-training is high stakes for both the student and the community and is currently under-researched. The form of assessment currently used in the veterinary programme is widely practiced in other universities and in other disciplines like medicine, yet its validity and reliability is also widely criticised. A mixed methods study combining a retrospective quantitative analysis of results of previous assessment with a qualitative interview study of lecturers and other

Te Kunenga ki Pūrehuroa
Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand T +64 6 356 9099 www.massey.ac.nz

supervisors will form the work for my Doctorate of Education. Understanding of what supervisors consider when judging veterinary student performance may inform new perspectives on what it is to be a competent veterinarian and how best to align assessment with the constructs.

The privacy of students, lecturers and supervisors will be respected and the individuals involved will not be identified in any outputs from the research. All data will be anonymised before I receive it, however , in reporting the research, Massey University will be identified as the institution involved. The data will be held securely on Massey University password-protected computers and servers and will be destroyed after 5 years. The findings will be shared with others in my Doctoral thesis and in conference presentations and publications.

The risk of this research to the reputation of Massey University and the Institute of Veterinary, Animal and Biomedical Sciences has been considered and discussed with Professor , Head of Institute and Professor , Programme Director for the Bachelor of Veterinary Science programme, both of whom approve of the research. There is the potential for the study to indicate deficiencies in our current assessment practices. The risks of harm to the reputation of the institution will be minimised by the fact that any problems that may be found are likely to be similar to those reported for many other institutions across a range of disciplines including medicine, and are also likely to be typical of those found in other veterinary schools. In addition the research findings may suggest areas in which assessment practice can be improved, to the benefit of the institution, staff and students. Such active evaluation of assessment practice and utilisation of results to improve practice is likely to be looked on favourably by accrediting bodies.

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 13/94. If you have any concerns about the conduct of the research, please contact Dr Nathan Matthews, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 350 5799 x 80877, email humanethicsouthb@massey.ac.nz .

Yours sincerely,

Liz Norman
Candidate for the Doctor of Education
Director of the Master of Veterinary Medicine programme

Appendix I: Request to conduct research – Pro-Vice Chancellor, College of Sciences

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

13th January, 2014

Professor
Pro Vice-Chancellor
College of Sciences
Massey University

Regarding: permission to conduct educational research in IVABS

Dear Professor ,

I am writing to you to seek your approval to conduct educational research in IVABS.

The purpose of this research is to investigate the process of judgement of the complex performance of veterinary students during their work in professional contexts in the Veterinary Teaching Hospital and external sites. The assessment of students in-training is high stakes for both the student and the community and is currently under-researched. The form of assessment currently used in the veterinary programme is widely practiced in other universities and in other disciplines like medicine, yet its validity and reliability is widely criticised. A research study combining a retrospective quantitative analysis of results of previous assessment with a qualitative interview study of lecturers and other supervisors will form the work for my Doctorate of Education. Understanding of what supervisors consider when judging veterinary student performance may inform new perspectives on what it is to be a competent veterinarian and how best to align assessment with the constructs.

The study involves me accessing anonymised final year BVSc student records for 2011, 2012 and 2013 as well as conducting interviews with staff who supervise students while on their clinical rotations. Consent will not be sought from students to access their records for the purposes of research. The reason for not seeking consent from students to use their records for research purposes is that the research would be harmed if not all students consented. As well as your own consent, I will be seeking consent from Professor to conduct this research in IVABS, to access those student records held internally within IVABS and to conduct the interviews with staff during working hours. I will be seeking consent to access centralised student records from Dr for the study. All data will be held securely on Massey University password-protected computers and servers and will be destroyed after 5 years.

Te Kunenga ki Pūrehuroa
Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand T +64 6 356 9099 www.massey.ac.nz

The findings of this research will be shared with others, though my Doctoral thesis and in conference presentations and publications. The privacy of students, lecturers and supervisors will be respected and the individuals involved will not be identified in any outputs from the research. All student record data will be anonymised before I receive it, however, in reporting the research, Massey University will be identified as the institution involved.

The risk of this research to the reputation of Massey University and the Institute of Veterinary, Animal and Biomedical Sciences has been considered and discussed with Professor , Programme Director for the Bachelor of Veterinary Science programme, who approves of the research. There is the potential for the study to indicate deficiencies in our current assessment practices. The risks of harm to the reputation of the institution will be minimised by the fact that any problems that may be found are likely to be similar to those reported for many other institutions across a range of disciplines including medicine, and are also likely to be typical of those found in other veterinary schools. In addition the research findings may suggest areas in which assessment practice can be improved, to the benefit of the institution, staff and students. Such active evaluation of assessment practice and utilisation of results to improve practice is likely to be looked on favourably by accrediting bodies. The study plan has also been discussed with Emeritus Professor , previous Programme Director for the Bachelor of Veterinary Science and longstanding member of one of the international accrediting committees, who indicated that he thought such evaluation of our assessment practice would be looked on favourably during accreditation.

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 13/94. If you have any concerns about the conduct of the research, please contact Dr Nathan Matthews, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 350 5799 x 80877, email humanethicsouthb@massey.ac.nz .

Yours sincerely,

Liz Norman
Candidate for the Doctor of Education
Director of the Master of Veterinary Medicine programme

Appendix J: Request to conduct research – Head of Institute, IVABS

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

13th January, 2014

Professor
Head of Institute
Institute of Veterinary, Animal and Biomedical Sciences
Massey University

Regarding: permission to conduct educational research in IVABS, to access student records held internally within IVABS and to conduct the interviews with academic staff during working hours

Dear Professor ,

I am writing to you to seek your approval to conduct educational research in IVABS, to access student records held internally within IVABS and to conduct the interviews with academic staff during working hours.

The purpose of this research is to investigate the process of judgement of the complex performance of veterinary students during their work in professional contexts in the Veterinary Teaching Hospital and external sites. The assessment of students in-training is high stakes for both the student and the community and is currently under-researched. The form of assessment currently used in the veterinary programme is widely practiced in other universities and in other disciplines like medicine, yet its validity and reliability is widely criticised. A research study combining a retrospective quantitative analysis of results of previous assessment with a qualitative interview study of lecturers and other supervisors will form the work for my Doctorate of Education. Understanding of what supervisors consider when judging veterinary student performance may inform new perspectives on what it is to be a competent veterinarian and how best to align assessment with the constructs.

The study involves me accessing anonymised final year BVSc student records for 2011, 2012 and 2013 as well as conducting interviews with staff who supervise students while on their clinical rotations. Consent will not be sought from students to access their records for the purposes of research. The reason for not seeking consent from students to use their records for research purposes is that the research would be harmed if not all students consented. I anticipate interviewing approximately 15 members of the academic staff who agree to participate in the study. Potential participants are those who act in a supervisory role for

Te Kunenga ki Pūrehuroa
Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand T +64 6 356 9099 www.massey.ac.nz

students on clinical rotations over the range of disciplines. Each interview is expected to take 30-60 minutes. These would be conducted at a time convenient for the staff member involved, which may, with your consent, include working hours.

As well as your own consent, I will be seeking consent from Professor to conduct this research in IVABS. I will be seeking consent to access centralised student records from Dr for the study. All data will be held securely on Massey University password-protected computers and servers and will be destroyed after 5 years.

The findings of this research will be shared with others, though my Doctoral thesis and in conference presentations and publications. The privacy of students, lecturers and supervisors will be respected and the individuals involved will not be identified in any outputs from the research. All student record data will be anonymised before I receive it, however, in reporting the research, Massey University will be identified as the institution involved.

The risk of this research to the reputation of Massey University and the Institute of Veterinary, Animal and Biomedical Sciences has been considered and discussed with Professor Programme Director for the Bachelor of Veterinary Science programme, who approves of the research. There is the potential for the study to indicate deficiencies in our current assessment practices. The risks of harm to the reputation of the institution will be minimised by the fact that any problems that may be found are likely to be similar to those reported for many other institutions across a range of disciplines including medicine, and are also likely to be typical of those found in other veterinary schools. In addition the research findings may suggest areas in which assessment practice can be improved, to the benefit of the institution, staff and students. Such active evaluation of assessment practice and utilisation of results to improve practice is likely to be looked on favourably by accrediting bodies. The study plan has also been discussed with Emeritus Professor , previous Programme Director for the Bachelor of Veterinary Science and longstanding member of one of the international accrediting committees, who indicated that he thought such evaluation of our assessment practice would be looked on favourably during accreditation.

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 13/94. If you have any concerns about the conduct of the research, please contact Dr Nathan Matthews, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 350 5799 x 80877, email humanethicsouthb@massey.ac.nz .

Yours sincerely,

Liz Norman
Candidate for the Doctor of Education
Director of the Master of Veterinary Medicine programme

Appendix K:
Information sheet for potential participants

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

Placement supervisor conceptions of veterinary student performance
INFORMATION SHEET

Researcher Introduction

My name is Liz Norman. I am a senior lecturer in veterinary science at Massey University and also a Doctor of Education student at the Institute of Education, Massey University. I am conducting this research in my role as a Doctoral student. My supervisors are Dr Peter Rawlins and Dr Linda Leach from the Institute of Education at Massey University. This study is taking place with the consent of Professor , Head of Institute, Institute of Veterinary Animal and Biomedical Sciences and Professor , Programme Director for the Bachelor of Veterinary Science degree.

Project Description and Invitation

This project is an investigation of the aspects of a veterinary student's performance in the workplace that their supervisors and teachers consider important. It is an interview study which will compare the thoughts of supervisors from different types of veterinary practice. The purpose is to identify broad themes across and between discipline areas. It is part of a larger study that it is hoped will inform improvements to our final year training and assessment so that our students are best prepared for veterinary work.

As someone who has helped us train final year veterinary students in the past, I invite you to participate in an interview which will help me to understand your expectations of students.

Participant Identification and Recruitment

You have been invited to participate because I understand from our records that you have worked in a supervisory capacity with veterinary students from Massey University as they work on their practice-based rotations. You will be one of at least 15 supervisors interviewed. I will be selecting potential participants to represent a range of types of veterinary practice including large animal and small animal, from within the university as well as from private practices.

Project Procedures

Your participation would involve you indicating by email that you agree to participate and returning the attached consent form after signing it. I can collect this at the beginning of the interview if you prefer.

We will arrange for me to conduct the interview with you at a time that is convenient for you, at your place of work or another location if you prefer. The interview will involve you responding to questions I pose for you during a private meeting. The interview is expected to take between 30 and 60 minutes.

Te Kunenga ki Pūrehuroa
Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand T +64 6 356 9099 www.massey.ac.nz

If you wish the interview to take place during your working hours, I will need to obtain permission for this from your employer. Please let me know this so I can approach your employer.

Data Management

I will record the interview so that I can analyse it later. The recording will be de-identified before being transcribed. You will have an opportunity to view the transcript and make any corrections if you so wish.

The audio recording, research analysis and findings will be kept on password protected Massey University computers and servers used only by me. The signed form indicating your consent to participate will be stored in a locked office by my supervisor. Information that indicates your identity will not be shared with anyone, even staff of the Institute of Veterinary Animal and Biomedical Sciences, unless with your specific consent. The data will be stored for five years and will then be destroyed.

All reporting on findings will protect your identity and that of your practice.

Participant's Rights

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

- decline to answer any particular question;
- ask for the recorder to be turned off at any time during the interview;
- withdraw from the study before or during your interview or up until 30th May 2014;
- ask any questions about the study at any time during participation;
- provide information on the understanding that your name will not be used unless you give permission to the researcher;
- be given access to a summary of the project findings when it is concluded.

Project Contacts

Researcher:
Liz Norman
Institute of Veterinary Animal & Biomedical Sciences
Massey University
Palmerston North, NZ
E.J.Norman@massey.ac.nz
+64 6 356 9099 ext 7898

Supervisors
Dr Peter Rawlins
Institute of Education
Massey University
Palmerston North, NZ
p.rawlins@massey.ac.nz
+64 6 356 9099 ext 84403

Dr Linda Leach
Institute of Education
Massey University
Palmerston North, NZ
L.J.Leach@massey.ac.nz
+64 6 356 9099 ext 84457

Please feel free to contact either myself or my supervisors with any questions regarding this project.

Compulsory Statements

This project has been reviewed and approved by the Massey University Human Ethics Committee: Southern B, Application 13/94. If you have any concerns about the conduct of the research, please contact Dr Nathan Matthews, Chair, Massey University Human Ethics Committee: Southern B, telephone 06 350 5799 x 80877, email humanethicsouthb@massey.ac.nz.

Appendix L:
Consent form for interview participants

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

Placement supervisor conceptions of veterinary student performance
PARTICIPANT CONSENT FORM

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I agree to the interview being sound recorded.

I wish/do not wish to have my recordings returned to me.

I agree to participate in this study under the conditions set out in the Information Sheet.

Signature: _____ **Date:** _____

Full Name - printed _____

Te Kunenga ki Pūrehuroa

Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand **T** +64 6 356 9099 **www.massey.ac.nz**

Appendix M:
Transcriber confidentiality agreement

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

Placement supervisor conceptions of veterinary student performance
TRANSCRIBER'S CONFIDENTIALITY AGREEMENT

I …… (Full Name - printed) agree to transcribe the recordings provided to me.

I agree to keep confidential all the information provided to me.

I will not make any copies of the transcripts or keep any record of them, other than those required for the project.

Signature: _____ **Date:** _____

Te Kunenga ki Pūrehuroa
Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand **T** +64 6 356 9099 **www.massey.ac.nz**

Appendix N:
Authority for release of transcripts

MASSEY UNIVERSITY
COLLEGE OF SCIENCES
TE WĀHANGA PŪTAIAO

Placement supervisor conceptions of veterinary student performance
AUTHORITY FOR THE RELEASE OF TRANSCRIPTS

I confirm that I have had the opportunity to read and amend the transcript of the interview(s) conducted with me.

I agree that the edited transcript and extracts from this may be used in reports and publications arising from the research.

Signature: _____ **Date:** _____

Full Name - printed _____

Te Kunenga ki Pūrehuroa
Institute of Veterinary, Animal and Biomedical Sciences
Private Bag 11222, Palmerston North 4442 New Zealand **T** +64 6 356 9099 **www.massey.ac.nz**