The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Feedback after summative OSCEs:

Optimising Delivery and Engagement

Dr Myo Min Soe

Thesis submitted for PhD in medical education

School of Medicine

The University of Nottingham

Acknowledgements

My supervisors, Professor Gillian Doody and Dr Mohsen Tavakol, went above and beyond from the very beginning to the end. I consider myself very lucky to have had these supervisors.

The University of Nottingham awarded me a scholarship and research funding. Dr Thomas Hewson and Dr Robert Cullum helped recruit participants and acted as peer interviewers. The medical school's Assessment Unit and administrators helped implement my feedback intervention.

Dr Kimberley Edwards and Dr Christopher Harrison assessed my thesis fairly and gave me detailed feedback for improvement. Dr Steven Agius acted as a qualitative supervisor and helped me make amendments for resubmission. My research benefited greatly from their input.

My family supported me throughout all these years in any way they could.

All of you kindly helped me complete this work. It could not have been completed otherwise. I sincerely appreciate you all very much.

Contents

Table of figures

Table of tables

Abstract

Background

Few opportunities for learning exist after high stakes summative assessments. Providing students feedback after summative Objective Structured Clinical Examinations (OSCEs) could be a learning opportunity. However, as engagement is often poor, it is necessary to have an understanding of the learners' perspective to design a summative feedback programme. Once summative feedback has been given, it is necessary to evaluate its quality and quantity and whether the learners engage with it as the examiners and learners may have little motivation and opportunity to engage with summative feedback.

Evaluating whether the learners read the feedback they receive belongs to the behavioural aspect of feedback engagement. Looking beyond this, cognitive and emotional aspects of feedback engagement can be better understood by defining and measuring specific learner attitudes that influence feedback engagement. Currently, there is no evidence of such constructs in medical education. Feedback orientation is a construct that represents a learner's receptivity to feedback. It is composed of Utility, Accountability, Feedback self-efficacy and Social Awareness dimensions. Feedback Orientation Scale is a twenty item five-point questionnaire that measures these constructs in business management. This study aims to generate the evidence of feedback orientation in medical education.

Methods

This is a mixed methods study conducted at the University of Nottingham. Firstly, a qualitative phenomenological study was conducted using semi-structured peer interviews to explore the learners' perception of feedback in

the context of summative assessments in order to inform the design of the intervention.

Secondly, a feedback intervention was implemented in which examiners used digital tablets to record written feedback to students during actual high stakes summative OSCEs. This feedback was delivered online at the same time as the results. The quality of the comments was assessed using a set criteria. Feedback website log-in data were collected. Learners were categorised into Honours, Pass and Fail groups. Quantity distribution of feedback across these groups and their log-in rates were analysed.

Thirdly, confirmatory factor analysis, item analysis and ANOVA were performed on FOS data collected from medical students. FOS scores were analysed in relation to the learners' course, training year and gender.

Results

Interviews

The students (N = 16) wanted specific personalised examiner feedback to be offered to everyone as they were receiving only generic feedback after summative OSCEs at the time. Some felt they will likely not use the feedback because they have already passed the exam. One student felt that one-on-one feedback they received after failing an exam was unsatisfactory as the feedback giver neither appeared to know nor care about anything apart from what was written on the mark sheet.

Feedback intervention

Using digital tablets in summative OSCEs to produce feedback was feasible, acceptable and efficient. Exams were delivered smoothly without any disruptive incidents.

Quality

Seventy examiners assessed 287 students in eight stations resulting in 2296 assessments. Nearly two thirds of these (N = 1600) produced free text feedback, of which four were illegible. No unprofessional or derogatory comments were found. More than half of the comments (N = 857) were descriptions of two or more actions performed by the student. A quarter of the free text feedback (N = 387) contained both a description what the student did and a suggestion for improvement. A minority contained only a general description or unspecific judgement (N = 240).

Quantity

Nearly all assessments (N = 2266) produced either free text or pre-written "tickbox" feedback. All students received either free text or tickbox feedback in all stations. All students received both types of feedback in at least two stations and nearly all students (N=271, 94.43%) did so in at least four stations out of eight. Quantity distribution of both free text and tickbox counts followed closely to normal distribution (mean = 125.73 and 32.45, SD = 50.88 and 5.37, skew = 0.67 and 0.12 respectively). Failed students received significantly more free text feedback than the rest (P = 0.00).

Engagement

A large majority (N=263, 91.64%) of students viewed their feedback after the main exam. Honours students' view rate (92.59%) was higher than that of the passed students (89.74%). A smaller proportion of students viewed their feedback after the resit (78.95%). A smaller proportion of passed students at resit (N=26, 83.87%) viewed their feedback while only three out of seven failed students did so.

FOS study

Nearly a third of all students (N = 422) responded to the FOS questionnaire. All items had high reliability (SE = 0.03, coefficient alpha = 0.89).

Confirmatory factor analysis produced good fit indices which supported the applicability of the model to the data.

Overall feedback orientation was high (mean = 3.99). Utility domain had the highest mean at 4.3 and the Feedback self-efficacy domain had the lowest mean at 3.7.

The overall scores were significantly higher in students from training years four and five ($p = 0.02$). Males had a significantly higher self-efficacy score than females ($p = 0.02$). GEM students had significantly higher scores in Utility and Social Awareness domains than non-GEMs ($p = 0.02$ and 0.03 respectively). The Utility scores increased in year four and five significantly ($p = 0.02$), while Social Awareness scores were significantly lower in the first year ($p = 0.00$). Females were significantly stronger in Utility than males when they are GEM students ($p = 0.02$). There was no other significant difference in any domain according to student variables.

Discussion

Feedback after summative OSCEs was successful: quality and quantity of feedback for all students was adequate and learner engagement was high. This success could be attributed to the information gained during the literature review and peer interviews. Examiner training should include specific information regarding what constitutes good feedback in order to further improve feedback quality.

Finding a direct relationship between academic performance and feedback engagement was congruent with existing evidence. It is not surprising because it is logical to assume that higher performing students are more motivated learners who would engage with feedback, or any learning opportunity, more than lower performing or unmotivated learners would.

Examiners showed a tendency to help poor performing learners correct their mistakes than help competent ones perfect their practice by giving significantly more feedback to poor performing students. This is likely because examiners felt it was important to help students pass and also to prevent incompetent practice. More evidence should be attained on the examiners' feedback giving behaviour.

This study generated strong psychometric evidence of feedback orientation in medical education using data from a large number of students. FOS was reliable to measure the same constructs in medical education as it did in business management. None of its items were redundant. Overall feedback orientation of the medical students was strong. It was not surprising to find more senior students and GEMs having higher feedback orientation and males having more Feedback self-efficacy than females. Students were aware of the utility of feedback more than they were confident about their ability to use the feedback they receive. Despite these variations, no demographic group appeared to be severely weak at any aspect of feedback orientation.

Medical educators can use FOS to assess how receptive their learners are to feedback, both as a group and individually, and give feedback accordingly. Future research should focus on the relationship between the learners' feedback orientation and their actual engagement with feedback, for example, in terms of whether they read the feedback they receive. Targeted interventions to improve each aspect of feedback orientation should be developed.

In conclusion, this study generated evidence on providing feedback to undergraduate medical students after summative OSCEs and on their feedback orientation.

Chapter 1. Introduction

This thesis is concerned with providing undergraduate medical students feedback after summative clinical skill assessments, evaluating the quality and quantity feedback, the learners' engagement with feedback and their feedback orientation. The context, the theoretical background and the rationale of the study are discussed in this chapter.

1.1 Feedback

A research-based, operational definition of feedback in clinical education is:

> ***"Specific information about the comparison between a trainee's observed performance and a standard, given with the intent to improve the trainee's performance." (Ridder, Stokking et al. 2008)***

The power and importance of feedback in education is well established (Hattie and Timperley 2007, Norcini 2010). While feedback is powerful, its impact is not always positive. While a systematic review in the context of clinical practice found that feedback had a positive impact on physician performance in 74% of the studies reviewed (N = 41) (Veloski, Boex et al. 2006), negative effects and problems with feedback are not uncommon in assessments in higher education (Gibbs 2006). Students have reported that assessment related feedback is often difficult to understand (Weaver 2006), non-specific, providing only vague and sweeping generalisations (Higgins, Hartley et al. 2001), or difficult to act on (Gibbs 2006, Poulos and Mahony 2008).

Consequently, providing feedback to students is a challenging area in undergraduate medical education. In the UK, the National Student Survey shows that medical students are less satisfied with feedback than other areas of learning experience, such as assessment and teaching, according

to the Office for Students 2018 NSS database. There have been calls to provide feedback after all forms of assessment (Norcini, Anderson et al. 2011).

However, a dichotomy in assessments exists in medical education in the form of formative and summative assessments. Often exclusively, formative assessments are used to produce feedback while summative assessments are used to make judgements about a learner's competency to progress to the next stage of learning or practice (Sinclair and Cleland 2007). In traditional curriculums, it is uncommon for a single assessment to serve both formative and summative purposes simultaneously as it has been assumed to be difficult (Wass, Van der Vleuten et al. 2001). However, there is recent evidence arguing against this assumption (Harrison, Könings et al. 2017).

1.2 Summative assessment culture

Most medical schools use traditional high stakes summative assessments at the end of a period of training to access the learners' clinical skills competency (Hauer, Teherani et al. 2008). As a result, summative assessment culture is a dominating influence on learning in undergraduate medical education (Harrison, Könings et al. 2017). This culture has shaped the learners and educators' approaches to assessment and feedback (Harrison, Könings et al. 2016).

Harrison et al identified several barriers to engagement with feedback and learning for medical students in a summative context (Harrison, Konings et al. 2015). One of the most influential barriers was the fear of failure. Summative assessments tended to induce strong emotions in learners to avoid failure at all costs, which in turn influences their social interactions with peers, family and teachers. Consequently, motivation for learning stemmed mainly from the pressure to not fail exams which were viewed

as end-point hurdles to which all learning was directed (Harrison, Konings et al. 2015). This environment fostered a learner mentality to concentrate on overcoming a relatively low hurdle of minimally acceptable competence rather than aspiring to become the best practitioner one can be because the latter is not normally rewarded (Harrison, Konings et al. 2013, Burdick, Boulet et al. 2018). Another study has found that only high performing students and female students tend to actively seek feedback on a past assessment for reflection while the rest do not seem to be motivated to learn beyond the bare minimum to pass the next summative assessment (Sinclair and Cleland 2007).

As end-of-the-year high stakes summative assessments are seen as a culmination of a year-long learning journey, giving detailed personalised feedback after summative assessments is uncommon. It is common only to give marks and pass-fail information as feedback after summative assessments (Hauer, Teherani et al. 2008) which is viewed as a tool for closure rather than a learning resource (Boud and Molloy 2013, Harrison, Konings et al. 2015). Furthermore, in a typical traditional undergraduate medical education curriculum, students go through a period of training and take exams to move on to their next phase of training or practice. This segmental approach to teaching contributes to the learner perception of assessments as an end-point hurdle (Sender-Liberman, Liberman et al. 2005, Eva, Munoz et al. 2010).

Due to these factors, students tend to judge the usefulness of feedback based on its relevancy to passing summative assessments in the near future (Harrison, Konings et al. 2015). Learners engage poorly with feedback if they do not perceive it to be directly relevant for their next summative assessment even if it may be useful for learning (Winstone, Nash et al. 2016). Formative assessments are often viewed as “trial runs”

for summative ones: feedback after formative assessments is used mainly to help the learner prepare for upcoming summative assessments (Rushton 2005, Harrison, Könings et al. 2017). Consequently, learners demand more feedback while engaging poorly with the feedback they receive (Boud and Molloy 2013). This is a missed opportunity for learning.

1.3 The shift away from the summative assessment culture

For nearly two decades, there have been calls within the general educational literature for a move away from high stakes assessments at the end of a period of learning, also known as assessment of learning, towards multiple low-stakes assessments throughout the period of learning combined with feedback, also known as assessment for learning (Shepard 2000). Within medical education, there has been a growing interest in programmatic assessment, which is aimed at incorporating both assessment for learning and assessment of learning (van der Vleuten, Schuwirth et al. 2012).

Any individual assessment method has its inherent limitations in terms of educational value and statistical strength. Therefore, Van Der Vleuten et al asserted that using various assessment methods in combination is likely to produce a more complete picture of a learner's competence than standalone high stakes summative assessments. In programmatic assessment, each assessment should provide meaningful feedback to learners but should not determine pass or fail decisions on their own (Schuwirth and Van der Vleuten 2011). Rather than a single high stakes summative assessment at the end, programmatic assessment collects data through multiple assessment methods throughout the course of learning until sufficient evidence is available to inform a robust overall pass-fail decision. The choice of assessments should depend on the educational justification and how they fit into the curriculum (van der

Vleuten, Schuwirth et al. 2012). Over the course of training, a feedback dialogue should take place between the learner and a mentor about the learner's performance so that appropriate actions can be taken in time (van der Vleuten, Schuwirth et al. 2012).

Strengths of programmatic assessment include making assessment more meaningful for learning as it generates feedback and being able to generate evidence from multiple different assessment methods (Wilkinson and Tweed 2018). Several medical schools including the Cleveland Clinic Lerner College of Medicine, and Maastricht University School of Medicine and Keele University School of Medicine have successfully implemented programmatic assessment in undergraduate medical education. Multiple studies have been conducted at these centres on challenges and benefits of programmatic assessment (Dannefer and Henson 2007, Bierer, Dannefer et al. 2008, Schuwirth and Van der Vleuten 2011, Harrison, Könings et al. 2016).

However, most medical schools have not adopted programmatic assessment as several challenges in implementation remain difficult. Firstly, adopting programmatic assessment represents a paradigm shift away from familiar and trusted traditional summative assessments (Obe and Hogard 2016). As most educators have only experienced a summative assessment culture in their training and practice, they often find it difficult to conceptualise such a different approach as programmatic assessment and consequently resist the change (Harrison, Könings et al. 2017). Many programmatic models require the use of a portfolio and large scale changes to assessments methods. Creating new assessment opportunities, re-mapping the learning objectives from one big summative assessment to multiple smaller assessments and setting up a decision-making process that includes committees of experts making holistic

judgements of a learner's competency based on data from multiple assessments could be daunting tasks for the curriculum planners. It is often associated with higher initial costs and requires redistribution of resources to make increased use of routinely collected data and observations, such as workplace based assessments (WBAs), giving feedback in both qualitative and quantitative forms and supporting students to use feedback (van der Vleuten and Heeneman 2016). However, these costs can be offset, at least partially, over long term as there would be less need for expensive high stakes examinations (van der Vleuten and Heeneman 2016).

Another challenge is the need to assure the students and other stakeholders who have a vested interest in how the school processes the assessment data to inform pass-fail decisions (Wilkinson and Tweed 2018). As the medical graduates are awarded licenses to practice medicine, the competence of the graduates needs to be supported by clear evidence, not only to the academic community but also to the general public. Although programmatic assessment can indeed generate strong and transparent evidence, it is not as familiar as the long established traditional summative assessments for the general public, students and many of the educators.

Furthermore, the present study took place at an institution which used only traditional summative assessments and showed no indication towards a shift to programmatic assessment at the time. It was not feasible for the present study to implement a truly programmatic assessment approach to replace traditional summative assessments in this context.

1.4 Feedback after summative assessments

A more feasible step towards assessments for learning is giving personalised feedback to all students after summative assessments. It

adds a formative element to summative assessments. This approach utilises data generated by summative assessments regarding a student's strengths and weaknesses. Medical schools invest significant resources in developing psychometrically defensible methods for high stakes assessments (Wass, Van der Vleuten et al. 2001), such as the objective structured clinical exams (OSCEs), which generate large amounts of data. Instead of using this data only to generate an overall mark to inform a pass-fail decision, this approach creates a learning opportunity to inform the learners of their strengths and weaknesses in more detail.

Secondly, according to the constructivist view point, assessment drives learning in three phases: pre-assessment, pure-assessment and post-assessment effects (Newble and Jaeger 1983, Dochy, Segers et al. 2007). As discussed earlier, summative assessments tend to drive pre-assessment learning to be aimed at the minimally required competence but not at excellence. The perception of summative assessments as end-point hurdles also limits post-assessment learning. Providing students personalised feedback after summative assessment would offer an opportunity, or even an obligation, for post-assessment learning (Harrison, Konings et al. 2013).

Furthermore, summative feedback could potentially improve student satisfaction, as this would address the students' calls for more feedback. However, while students consistently demand more feedback, they often do not use the feedback they receive (Handley, Price et al. 2011, Winstone, Nash et al. 2016). Withey (2013) highlighted a so-called feedback paradox, in which students clearly recognise the importance of feedback, and frequently complain about the quality of feedback they receive, yet also make limited use of it. The resources committed to the

production of feedback would be wasted if the students are either not able or willing to utilise feedback.

1.4.1 Challenges

The first practical challenge in generating and delivering personalised feedback to every medical student is time, especially for clinical skill assessments. Objective Structured Clinical Examinations (OSCEs) is a ubiquitous method for assessing clinical skills in undergraduate medicine (Setyonugroho, Kennedy et al. 2015). Hundreds of students take summative assessments at the same time in most medical schools (Hauer, Teherani et al. 2009). The OSCE method provides a framework for examiners to objectively assess students, but offers little opportunities for feedback. Giving immediate feedback during high stakes summative OSCEs could be distracting and or cause the candidates undue stress. Organising individual feedback meetings after OSCEs for hundreds of students may also not be feasible. Examiners need to be able to generate feedback based on their observations during OSCEs without disrupting the assessment process. As high stakes summative assessments typically occur at the end of the academic year, feedback needs to reach the students before they move on to their next phase of training or practice. Generating and delivering individualised feedback to all students in a timely manner is a challenge.

If immediate or delayed verbal feedback is not feasible, it is also not practical for the examiners marking OSCEs to write down feedback for students during the exam using pen and paper. Time and space are inevitably limited and consequently, the information is often poorly constructed and illegible (Harrison, Molyneux et al. 2015, Denison, Bate et al. 2016). In addition, it is not a small undertaking to relate the pen-written examiner comments to students, as the mark sheets need to be

scanned and sorted for each student (Denison, Bate et al. 2016). This labour-intensive process inevitably restricts individualised feedback to only those who fail.

Furthermore, medical schools are often reluctant to include any information about the questions in feedback because it is expensive to create high quality questions. Therefore the content of feedback needs to be informative for the student without giving away the questions too much. The content of feedback is arguably the most important factor as most students will engage with feedback only if it contains useful, relevant or encouraging information (Winstone, Nash et al. 2016). Feedback after summative assessments needs to be more than an account of what the student did and what the correct answers were because most students are not likely to find this information relevant to their training or practice by the time they receive it. To remain relevant and to avoid giving away the questions, summative feedback should address repeating patterns of behaviour, such as communication skills and professionalism.

Production of feedback should be complemented with promotion of feedback (Sargeant, Mann et al. 2009). Students should be informed how the feedback was produced and what purpose it was meant to serve. Otherwise, a perception gap could appear between the educator and the learner regarding feedback (Evans, Zeun et al. 2014).

One of the aims of this research project was to deliver personalised feedback to all students after summative objective structured clinical skill exams (OSCEs). The researcher, Myo Min Soe (M.M.S.), conducted this research project in collaboration with the supervisors Gillian Doody (G.D.) and Mohsen Tavakol (M.T.). As discussed above, there are potential learning benefits from doing so if the challenges can be addressed. A literature review was conducted on feasible and effective approaches to

feedback after summative and formative clinical skill assessments in order to address these challenges. Another reason was that medical students at the University of Nottingham were not receiving any personalised feedback after summative OSCEs although they received some feedback after summative knowledge assessments. A detailed description of their assessments and feedback is presented later in this chapter.

1.5 Engagement with feedback

Provision of feedback should enable a learning process to occur, in which both the educator and the learner actively engage (Boud and Molloy 2013, Winstone, Nash et al. 2016). As discussed earlier, the summative assessment culture is a limiting factor to learner engagement with feedback. Engagement with feedback is a complex nuanced phenomenon (Handley, Price et al. 2011), but there are several commonly accepted steps to the process (Hattie and Timperley 2007). Firstly, a learner needs to be receptive to feedback. Secondly, they need to have the capacity to interpret the message. Thirdly, the learner needs to set their own learning goals and take action to achieve them (Hattie and Timperley 2007, Shute 2008). The learner's receptivity to feedback is influenced by their emotions elicited by the content of feedback, their previous experiences in receiving feedback, the context in which feedback is given and their relationship with the feedback giver (Fredricks, Blumenfeld et al. 2004). In addition, the learner's cognition of the utility of feedback and their motivation to learn also play a role. The learner's emotions and cognition influence their behaviour, in terms of whether they will avoid feedback or passively accept it or proactively seek it, and how they will reflect on the message and whether or not they will make changes in their behaviour (Fredricks, Blumenfeld et al. 2004).

Some learners want feedback to build their confidence rather than to address their deficiencies (Eva, Armson et al. 2012). Some want feedback but fear receiving information that could challenge their perception of self (Mann, van der Vleuten et al. 2011). Receiving negative or constructive feedback has a profound effect on an individual's self-perception, even in a low stakes multisource feedback context (Sargeant, Mann et al. 2008). Emotional impact could be much higher in a high stakes summative assessment context. Consequently, students may be reluctant to engage with summative feedback.

The first step of feedback engagement can be objectively measured and analysed, in terms of whether the learners accept the feedback they receive by reading or listening to it. Sinclair and Cleland (2007) did a statistical analysis on how many students read individualised feedback reports for their essays after a summative exam. More than fifty percent of students did not pick up their feedback. In contrast, Harrison et al. reported a higher recipience rate (Harrison, Konings et al. 2013). Analysis of the log-ins to a website hosting OSCE station grades and structured feedback showed nearly all (97%) of the students visited it at least once but only half of them viewed all pages on the website (Harrison, Konings et al. 2013). However, the authors acknowledged that many of the students may have been logging in only to view their marks in each station out of curiosity rather than to access detailed feedback. Furthermore, the website did not contain any examiner comments at the time. These two were the only papers which measured learner engagement with feedback after summative assessments. This limited body of evidence shows that the learner's behavioural propensity to receive summative feedback is often poor. Therefore, it is necessary to

investigate whether the learners read the feedback they receive if one desires to give them feedback after assessments.

These studies also investigated the relationship between feedback engagement and academic performance. Sinclair and Cleland (2007) found poor performing male students commonly did not use feedback as a learning tool. Higher performing students and female students tended to actively seek feedback. Researchers mentioned that making individual feedback reports was labour intensive. Since students demand more "useful" feedback, it is not enough to focus only on the production of feedback (Sinclair and Cleland 2007). A reason why students did not collect written feedback for their essay writing may have been that there were no future summative exams that would assess their essay writing. Similarly, Harrison, Molyneux et al. (2015) found high performing students visited all pages and rated the feedback website positively. Borderline pass students viewed the least number of pages on the site. Failed students had to go to remediation in person and use the website to receive feedback. However, a study in a formative context at Cardiff University reported that factors such as previous academic performance, personal demographics or computer literacy did not predict whether the students will read their feedback online (Stone 2014). Stone found that engagement with feedback depended more on previous emotional reactions with some students seeing little value in feedback because of negative experiences. Another study set up voluntary remediation sessions for students who failed exams (Winston, Van der Vleuten et al. 2010). They found that most of the poor performing students did not attend these sessions. Poor performer's attendance rates increased after the school made the remediation sessions mandatory for every failed student. Beyond medical education, organisational psychology literature

by Moss et al. has proposed that active feedback avoiding behaviour exists especially among poor performing workers (Moss, Sanchez et al. 2009). They tended to actively avoid feedback when they are aware that their performance is poor. Other studies have also demonstrated that poorer performers are less likely to seek feedback (Northcraft and Ashford 1990, Morrison and Bies 1991). Overall, the evidence supports a direct relationship between a learner's performance and their feedback engagement. The present study aims to further contribute to the evidence on this topic in a summative context.

Furthermore, as feedback engagement could be often poor, the first step in designing a feedback intervention should be to investigate the learners' perspective to their assessment and feedback programme. Learners do not universally perceive that clinical performance assessments provide them with valuable developmental input on their performance but their perception of the value of feedback is influenced by the environment, the assessor and themselves (Scarff, Bearman et al. 2019). Understanding their perception of feedback and the context in which it is given could help design a feedback programme that attracts engagement.

1.6 The construct of feedback orientation

It is clear in the evidence discussed above that the learners' awareness on the utility of feedback and their sense of accountability to engage with the feedback process is often weak. While there is some evidence on the behavioural aspect of learner engagement with feedback in medical education, no study to date has explicitly identified or measured the attitudes that influence a learner's overall receptivity to feedback in medical education.

Feedback literacy has been proposed as a construct that influences engagement with feedback. Feedback literacy involves knowing the

purpose of feedback, utility of feedback, dealing with emotions regarding feedback and acting on the feedback (Sutton 2012). High performing learners often have high feedback literacy which enables them to reflect more deeply on the feedback they receive (Sutton 2012). However there has not yet been any quantitative evidence of feedback literacy.

Considering a rich body of evidence outside medical education, Sanders encouraged medical educators to consider insights from non-medical educational research on feedback (Sandars 2015). Occupational psychologists have done extensive research on feedback seeking behaviour in the workplace.

The construct of feedback orientation was developed originally in the human resources management context. London and Smither (2002) proposed feedback orientation as a construct that represents a learner's tendency to be open to and use the feedback they receive. This construct covers the emotional and cognitive aspects of the learners' attitudes to feedback which influences their behaviour. Feedback orientation is defined as an individual's overall receptivity to feedback (London and Smither 2002). Learners with strong feedback orientation are more likely to know the potential learning and social benefits of feedback, feel obligated to engage with the feedback process and feel confident about their ability to make the best use of feedback. As with any other variable, a learner's feedback orientation could be influenced over time by their own effort or the learning environment created by social norms, peer attitudes and curriculums. Feedback orientation is composed of a number of dimensions including liking feedback, behavioural propensity to seek feedback, cognitive propensity to process feedback mindfully and deeply, sensitivity to others' views of oneself, belief in the value of feedback, and feeling accountable to act on feedback. Based on these dimensions, Linderbaum

and Levy (2010) defined four constructs: Utility, Accountability, Social Awareness, and Feedback self-efficacy.

1.6.1 Utility

Utility is defined as an individual's tendency to believe that feedback is useful in achieving desired outcomes. Utility represents the belief that feedback can lead to valued outcomes as opposed to Defensiveness which represents the belief that feedback can cause unpleasant emotions. It is based on London and Smither's dimension "belief in the value of feedback". Perceived usefulness of feedback has been found to influence the motivation to accept, seek, and use feedback (Brett and Atwater 2001, Kanfer, Frese et al. 2017).

1.6.2 Accountability

Accountability refers to an individual's tendency to feel a sense of obligation to react to and follow up on feedback. It is based on London and Smither's dimension titled "feeling accountable to act on feedback". The theory of accountability suggests that those who feel accountable will behave in ways to increase the likelihood that others will view them in a positive way (Hall, Frink et al. 2017). The learners are accountable for utilising the feedback they receive, as much as the educators are accountable for making accurate ratings and providing useful feedback.

1.6.3 Social Awareness

Social Awareness refers to an individual's tendency to use feedback to be aware of others' views of oneself and to be sensitive to these views. This dimension was based on the idea that subjective norms drive behaviour (Ajzen and Fishbein 1977). The Social Awareness factor is based on the construct of public self-consciousness i.e. the extent to which individuals see themselves as social objects and are aware of the observation of others in public contexts (Fenigstein, Scheier et al. 1975). Individuals with

higher public self-consciousness tend to have a greater desire for feedback (Levy, Albright et al. 1995). While the sense of Accountability stems from internal pressure to respond to feedback, Social Awareness stems from external pressure.

1.6.4 Feedback self-efficacy

Feedback self-efficacy refers to an individual's perceived competence of self to interpret and respond to feedback appropriately. Self-efficacy is a context-specific evaluation of competence, whereas self-esteem is a global evaluation. People are more apt to exhibit particular behaviours if they believe they have control over their performance (Ajzen and Fishbein 1977, Linderbaum and Levy 2010). However, there has not yet been any direct evidence of the relationship between feedback self-efficacy and seeking or using feedback.

Linderbaum and Levy developed a twenty item questionnaire, named Feedback Orientation Scale (FOS), to measure these four aspects of feedback orientation. They performed exploratory factor analysis to identify the factors and validate the questionnaire in human resources management context (Linderbaum and Levy 2010). Other studies in similar contexts have used FOS (Braddy, Sturm et al. 2013, Lilford, Caruana et al. 2014). These studies found the data from FOS valid in their context, supporting its validity argument for the business workplace. They agreed with the four factor theoretical framework to feedback orientation.

Although there has not yet been any evidence of FOS in medical education, studies that explored student engagement with feedback touched on some of the constructs related to feedback orientation. As discussed above, due to the summative assessment culture, medical students tend to perceive the usefulness of feedback only in terms of how directly relevant it is to passing the next exam (Harrison, Konings et al.

2015). High performing students are more open to feedback, while borderline and failed students tend to not engage well with feedback (Winston, Van Der Vleuten et al. 2010, Winston, Van der Vleuten et al. 2010, Harrison, Konings et al. 2013, Winstone, Nash et al. 2016). These findings indicate that poor performing medical students may have a weak sense of utility and accountability regarding feedback.

In addition to learner performance, the learners' demographic backgrounds may also influence their receptivity to feedback. Reactions to receiving feedback and subsequent decision making are different between genders (Niederle and Vesterlund 2011, Berlin and Dargnies 2016). Learners of different genders and training years displayed significantly different reactions to patients' feedback and showed empathy differently (Greco, Brownlea et al. 2001, Hojat, Vergare et al. 2009, Kataoka, Koide et al. 2009). The course of study may also affect student attitudes towards feedback because graduate entry medicine (GEM) students are generally older and have had more diverse life experiences than undergraduate entry students (Non-GEMs). Therefore it has been hypothesised that such student variables may influence student attitudes and emotions related to feedback (Sargeant, Mann et al. 2008). If the educators know how well their learners are orientated to feedback engagement, they can take appropriate actions to improve the learner's weaker aspects of feedback orientation in conjunction with giving feedback.

1.7 Context

This PhD course and its research took place at the University of Nottingham, School of Medicine (UONSOM). It has approximately 1500 registered undergraduate students. The school uses a traditional undergraduate medical curriculum based on taught modules followed by

summative assessments. It is a large UK medical school with a traditional curriculum that could be considered typical for a UK medical school.

1.7.1 Course structure

There are two courses to the B.M.B.S. degree (Bachelor of Medicine, Bachelor of Surgery); A100 and A101. The A100 course is an undergraduate entry (Non-GEM) five-year long Bachelor of Medicine Bachelor of Surgery (B.M B.S.) course. A101 leads to the same degree but it is graduate entry to medicine (GEM) in which people who already have an undergraduate degree could join medical school.

In the first two years of training in the A100 course, basic medical science is taught in a series of modules. In the first half of the third year, students undertake a supervised research project, leading to the degree of BMedSci. Starting from the second half of the third year, students receive clinical training. The periods of training are termed as clinical phases; CP1 (clinical phase one, year three), CP2 (clinical phase two, year four) and CP3 (clinical phase three, year five), where they rotate through a series of placements at local teaching trusts. Summative exams in all training years are composed of knowledge and clinical skill exams. Clinical skills assessments use the objective structured clinical examination format (OSCE) and knowledge assessments use single-best-answer multiple choice questionnaires (MCQs).

1.7.2 Summative assessment and feedback in 2015

After summative knowledge exams, all students in all years received learning objective based feedback in the form of a list of learning objectives with red, yellow and green signs attached to each objective. This was named the Traffic Lights System and it was generated automatically using ROGO, a digital assessment management software developed by the University of Nottingham.

Every summative OSCE involved approximately 300 students and seventy examiners. In 2015, school administrators printed the mark sheets, transported them to the exam sites and collected them afterwards. Then they checked missing and illegible data on the mark sheets immediately to clarify any clerical errors while the examiners were still there. If any mistakes were found later, the protocol was to award the student full marks for the item. The marked papers were then scanned using a special scanner to digitalise the marks onto a Microsoft Excel™ spreadsheet. All these tasks needed to be performed securely and quickly, requiring experienced administrators. In 2015 CP3 OSCEs, it took two administrators three full working days to complete all the logistic tasks.

After summative OSCEs, only year one and year two students received immediate verbal examiner feedback. There was no personalised feedback to students after the OSCEs starting from CP1 (year three). In CP1, CP2 and CP3 until 2015, students received their own marks for each OSCE station, the mean mark of the whole cohort for each station and the three most common mistakes that students in the whole cohort did. In 2015, the number and length of OSCE stations varied according to the level of training: CP1 OSCEs had shorter and fewer number of stations than CP3 OSCEs. It was possible for a student to fail a few stations but achieve an overall pass. Only failed students were offered one-on-one meetings with a member of the Assessment Unit or a module leader. The Assessment Unit was composed of a director of assessments and clinical teaching fellows, all of whom were specialist clinicians with medical education backgrounds. A module leader is a senior educator, in charge of delivering a module, who could either be a clinician or a non-clinician. In 2015, there was no system to allocate which educator meets which student, or training for feedback givers, or a detailed guidance about how feedback

should be delivered. The feedback givers in these meetings did not have any additional information other than the student's OSCE mark sheets. It was common for students to dispute the examiners' decisions and what actually happened in the exams. The educators discussed with students regarding their performance, their current situation and advise whether to take the resit or repeat the whole year. The students were free to choose any option no matter the advice. Those who failed the same exam three times were terminated from the course.

This was the educational context in 2015. In 2016, the school moved from traditional OSCE checklists to domain marking. Digital marking was piloted in December 2016. From then on, all students in all years received written feedback online after summative OSCEs. These changes were described in more detail in Chapter 4.

1.7.3 Student voice

The faculty and student representatives hold monthly meetings to discuss what changes students want and what the school is doing. These are named as Learning Community Forums. These forums are meant to enable regular interaction between the students and the faculty, to raise issues and transparently monitor progress in addressing them. Students also voice their opinions via the National Student Survey (NSS).

1.8 Research question, aim and objectives

To summarise, summative assessments limit the learners' motivation and opportunities for learning. Feedback after summative assessments can potentially be an opportunity for learning. It utilises existing high stakes summative assessments to generate feedback. However, learner engagement with feedback is not guaranteed. Therefore, designing the feedback intervention requires a sound understanding of the learners' perspective. Furthermore, after the implementation, the quality and

quantity of feedback needs to be evaluated along with the learners' engagement. Learners' feedback orientation should also be measured to gain a better understanding of the learners' attitudes towards feedback engagement. Doing so would provide a learning opportunity for the medical students at Nottingham, contribute to the small body of evidence on summative feedback and generate original evidence on feedback orientation in medical education.

The overarching research question was:

"What are the elements of a feasible and acceptable approach to the feedback process to students following high stakes summative OSCEs?"

The overarching aim was to develop, implement and evaluate a feasible and acceptable programme for delivering detailed individualised feedback to all students after summative clinical skills assessments in undergraduate medicine, to construct a validity argument for the use of Feedback Orientation Scale (FOS) in medical education and to assess the learners' feedback orientation.

The objectives were:

1. To review the literature on various approaches to feedback for students after clinical skill assessments to identify optimal timing, mode of communication, feedback giver and content
2. To identify the perceived needs of the learners regarding feedback after high stakes summative assessments
3. To pilot, implement and evaluate summative examiner feedback to students

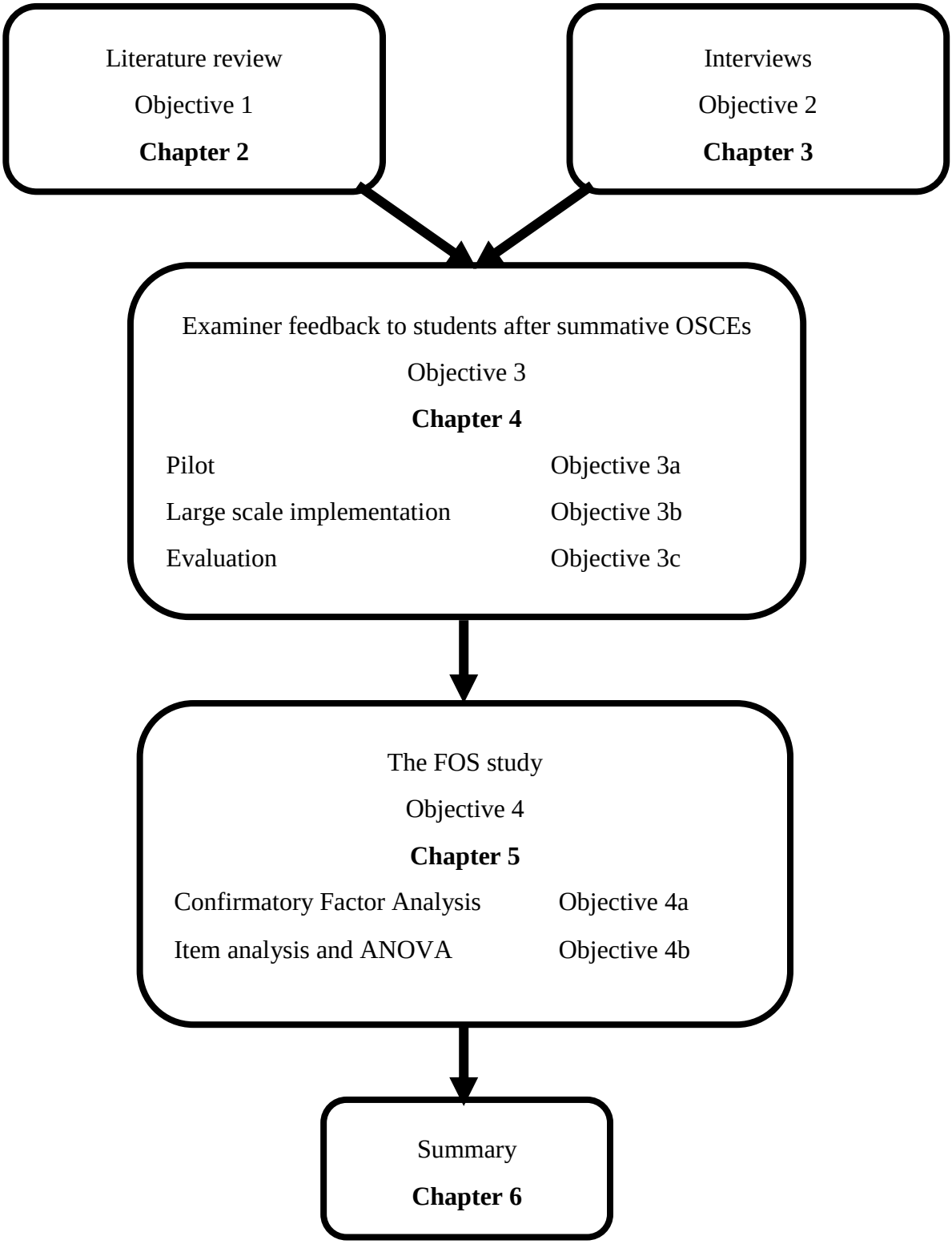
a. To pilot the summative feedback at a high stakes summative OSCE using digital marking
b. To implement summative feedback for the whole school
c. To evaluate the quality and quantity of the examiners' feedback comments and the learners' engagement with feedback
    i. To appraise the content of feedback using a set criteria and to analyse its distribution across the student cohort.
    ii. To analyse the feedback distribution across the stations and student performance.
    iii. To analyse the students' log-in rates to the feedback website in relation to their performance.
    iv. To consider the intervention's acceptability, feasibility and efficiency

4. To construct a validity argument for the use of Feedback Orientation Scale in medical education and to assess the learners' feedback orientation.
    a. To examine the dimensionality and internal consistency of the Feedback Orientation Scale (FOS) in the context of medical education,
    b. To measure the relationship between feedback orientation and student demographics in terms of gender, course and training year.

1.8.1 Programme of research

To meet the objectives, this thesis follows a programme of research as shown in Figure 1. The chronological order in which the research was done is shown in the timeline below.

Figure 1 Programme of research



1.8.2 The timeline

February 2015	Literature review.	Objective 1, Chapter 2.
August 2015	Interviews.	Objective 2. Chapter 3.
December 2015	Member checking for interviews.	Objective 2. Chapter 3.
December 2016	Digital marking and feedback pilot.	Objective 3a. Chapter 4.
February 2017	Digital marking and feedback at a large scale summative OSCE. - Emailed feedback to students via Microsoft Outlook™ mail-merge	Objective 3b. Chapter 4.
April 2017	Feedback Orientation Scale data collection and analysis.	Objective 4. Chapter 5.
January 2018	Software update - Enabled students to log in and view their own feedback. - Enabled administrators to see the students' website usage data - Outlook™ mail-merge no longer required.	
February 2018	Evaluation of feedback content and student engagement	Objective 3c. Chapter 4.

1.8.3 Structure of the thesis

Three studies were performed in the following order over the course of this PhD; the interview study, the summative feedback evaluation study and the feedback orientation study. Firstly, a literature review on various feedback approaches after clinical skill assessments identified the best evidence on the subject. The interview study identified the local learners' perceived feedback needs. The literature review and the interviews informed the design of the feedback programme. Implementation of the

summative feedback programme was followed by an evaluation of the examiners' feedback, the learners' engagement with feedback and a study of the learners' feedback orientation.

To illustrate this process, this thesis was presented as a main introduction and a literature review followed by a series of "paper style" chapters and a summary. The researcher assumed that this format was more appropriate than a traditional thesis format (introduction, literature review, methods, results and discussion) to demonstrate the logic and evolution of this thesis.

Chapter one is the introduction. It presents the background and rationale for the present research.

Chapter two presents a traditional literature review on the topic of interest.

Chapter three presents a qualitative inquiry on the students' perspective on feedback in the context of summative assessments.

Chapter four presents the development, pilot, large scale implementation and evaluation of a novel feedback programme to all students after summative OSCEs. Evaluation includes appraising the quality of the examiners' feedback comments to students, analysing the quantity distribution of feedback across the student cohort, and analysing of the students' engagement with feedback in terms of whether or not they logged in to the feedback website.

Chapter five presents a validity argument for the use of Feedback Orientation Scale (FOS) in medical education. This study tests the psychometric properties of FOS to determine its suitability for use in medical education, followed by item analysis and analysis of the FOS scores of learners from different demographic backgrounds.

Chapter six presents an overall discussion on the evidence generated by the whole project. It includes reflection, limitations and implications for future research and practice. This chapter is followed by the reference list and the appendix.

1.8.4 Philosophical approach

This thesis followed the philosophical approach of pragmatism and used mixed methods.

An educational environment is multifactorial. It is shaped by interactions between numerous factors (Conrad and Serlin 2005). Therefore, educational research should take a holistic approach using multiple methods (Berliner 2002). The realities social sciences try to understand are so complex that researchers should consider multiple views to truly understand them (Creswell, Clark et al. 2003). Pragmatism is suitable for such a purpose. Pragmatism is a paradigm underpinning research which is real-world practice oriented, problem-centred, pluralistic and "focused on consequences of actions" (Creswell, Clark et al. 2003). Pragmatism approach is flexible. It uses multiple methods to gather information and generate evidence. As such, it is often seen as an excuse for "philosophical anarchy", where the researchers do convenient research without a strong philosophical justification.

Pragmatism is between the positivist paradigm that defines knowledge as an objective reality and the interpretivist paradigm that defines knowledge as a subjective construct. The assumption in positivism is that only objective quantitative methods could generate generalizable findings across context and populations. Unlike the positivist approach, pragmatism does not consider reality to be a single unchangeable overarching truth which could be uncovered by research. On the other hand, pragmatism also differs from the interpretivist paradigm by not

assuming that all reality is subjectively constructed in the interaction between the researcher and the object of research. Pragmatism accepts both objectivity and subjectivity are important for knowledge creation and considers knowledge to be the researchers' interpretation of an “external reality”. Pragmatism assumes that this external reality is translated into knowledge through the lenses of the researchers. Pragmatic construct of knowledge has been defined as “truth being what works at the time” (Maudsley 2011). In other words, pragmatic assumption of knowledge driven from research is not a universal truth but it is only an adequate explanation for the outcome. Differing from other research paradigms, pragmatism does not dictate specific methods based on the ontological and epistemological assumptions. It assumes that both qualitative and quantitative methods are equally valid ways to research as long as they are consistent with the research questions and problems to be investigated. Therefore, mixed-methods research is often driven by this paradigm (Creswell, Clark et al. 2003).

It was assumed that this study’s outcomes were influenced by the researcher’s methodological choices and the context in which the research took place. Following the paradigm of pragmatism, both qualitative and quantitative methods were used. The researcher did not assume this study’s findings to be absolute truths but only as knowledge generated in its particular context which may be applicable in similar contexts.

Chapter 2. A literature review on feedback to learners after clinical skill assessments

2.1 Introduction

Chapter 1 presented a rationale for giving feedback to students after summative OSCEs. In order to inform the design of this intervention, the evidence on various approaches to feedback were explored. A traditional literature review on peer reviewed published literature was done. Evidence on giving individualised feedback to learners in both summative and formative contexts was considered. This review considered learners in all healthcare education, including undergraduate medical students, nurses, dentists, pharmacists and postgraduate clinical specialist trainees. The quality of all papers included in this review were assessed using quality appraisal checklists developed by the Joanna Briggs Institute. The aim was to identify the optimal timing, mode of communication, feedback giver and content to give feedback after summative OSCEs.

2.2 Tradition of literature reviews

For centuries, the academic community has appreciated the need for unbiased analysis of the collective evidence from all available research. James Lind, the 18th century doctor who conducted one of the first ever clinical trials to identify the cause of scurvy, wrote,

> ***"As it is no easy matter to root out prejudices … it became requisite to exhibit a full and impartial view of what had hitherto been published on the scurvy … by which the sources of these mistakes may be detected. Indeed, before the subject could be set in a clear and proper light, it was necessary to remove a great deal of rubbish"(Lind 1753)***

Gathering all available literature and summarising the best evidence remains the essence of literature reviews ever since. The twentieth century saw the establishment of explicit methods for different types of reviews for an exponentially growing body of evidence, driven by the rise of evidence based practice starting from the 1970s (Chalmers, Hedges et al. 2002, Grant and Booth 2009). Transparency of the methods,

reproducibility of the results and comprehensiveness of the literature search are distinctive characteristics that set systematic reviews apart from traditional reviews. Systematic reviews require considerably more time and resources than the traditional ones as at least two reviewers are usually required to strengthen the reproducibility and comprehensiveness of the literature search, appraisal and data extraction. On the other hand, traditional reviews allow more freedom as there is no need to set a strict criteria for paper selection, allowing them to address broader research questions. A traditional literature review was done for this study.

2.3 Previous reviews

There are two published reviews which considered feedback after assessments in medical education. Both are systematic reviews. Their areas of interest are not identical to that of the present study but are similar and relevant.

The first is a scoping review on feedback to all learners in medical education, published in Academic Medicine (Bing-You, Hayes et al. 2017). The second is a narrative review of medical specialist trainees' perspectives of the assessment messages they receive, published in Medical Education (Scarff, Bearman et al. 2019). In both papers, review questions were clearly and explicitly stated. The inclusion criteria and search strategies were appropriate for the review questions. Both used adequate databases. Multiple reviewers were used for paper screening and selection. Scarff et al used medical education research study quality instrument (MERSQI) tool to assess the quality of the selected papers while the scoping review by Bing-You et al did not use any appraisal criteria, stating that it is not the purpose of a scoping review to appraise the quality of the studies. Bing-You et al used two researchers for data extraction to minimize errors but Scarff et al used only one. Likelihood of

publication bias was recognised as a limitation by both studies as they considered peer reviewed published articles written in English only. Their recommendations for practice and further research were logical and drawn directly from the reported data. Therefore, both reviews were judged to be of good quality and with valid findings.

The Bing-You et al review's scope was broad as it covered feedback in all medical education contexts including teaching and assessment (Bing-You, Hayes et al. 2017). This review identified 650 relevant papers, showing that there is a large body of evidence on the topic. Most papers were published in recent years. Despite stating they did not objectively appraise the quality of the papers they reviewed, Bing-You et al reported finding few studies that used rigorous methods such as randomised educational trials while many were of lower quality levels of evidence such as single, observational feedback interventions with small numbers of students (Bing-You, Hayes et al. 2017). However, they reported that many papers contained only general descriptions of new or altered curricular approaches, and a gap for high quality evidence on the feedback process remained.

Furthermore, they highlighted the importance of emotions and the credibility of the feedback givers and that the repeated and immediate feedback was commonly seen to be beneficial for learning (Bing-You, Hayes et al. 2017). This was a relevant finding for the present study's context. The researcher assumed that strong emotions are inevitable in a high stakes summative context. The faculty or clinician examiners who are experts in the subjects they assess are more credible than standardised patients or peers. Immediate feedback is common in formative assessments but Bing-You et al did not report any evidence of immediate feedback in summative assessments.

The Scarff et al's review is more recent as it was published in 2019. It is a narrative systematic literature review of medical specialist trainees' perspectives of the assessment messages they receive (Scarff, Bearman et al. 2019). Postgraduate clinical specialist trainees are termed as "residents" in the US and "registrars" in the UK. All undergraduate clinical training placements are termed as "clerkships" and students in clinical placements as "clerks" in the US. Non-specialist junior doctors in training are termed as "foundation year doctors" in the UK. While Scarff et al considered both summative and formative contexts, their review's scope was limited to workplace based assessments (WBAs) and in-training assessments (ITAs) for specialist trainees. They excluded non-workplace based assessments such as the OSCEs. As this review focused on the learners' perspective, it showed that many trainees valued developmental feedback they received and they tended to disengage from the feedback process when they feel the messages are not constructive for their skill development. Their perceptions are influenced by the environment, the assessors and themselves (Scarff, Bearman et al. 2019). These findings support the notion that the learners do not universally value or engage with the feedback they receive. Therefore, the educators need to not only deliver feedback but also be aware of the learners' perception to and engagement with feedback.

No further relevant reviews were found. However, it was noted that literature reviews are not the only way to gather data on various approaches to feedback and assessment. In 2008, Hauer et al contacted the deans in every medical school in the United States for information on various remediation methods in medical schools and the consequences students faced after failing assessments (Hauer, Teherani et al. 2008, Hauer, Teherani et al. 2009). Their research did not identify any school

that delivered personalised feedback to all students after summative assessments. The common approach was to offer failed students one-on-one meetings with faculty to discuss why they failed and what to do next. Hauer et al's approach is useful for collecting data on current practices in medical schools but not necessarily the evidence to support the effectiveness of those practices. Furthermore, this type of data collection was deemed unsuitable for the present study because of the possibility of low reply rates, the lack of incentives to offer and the researcher's preference for peer reviewed evidence to descriptive accounts.

2.4 Feedback after summative clinical skill assessments

Several studies which explored feedback to students after summative clinical skill assessments were identified.

In 2013 at Keele, a website was set up to deliver feedback to medical students after summative OSCEs (Harrison, Konings et al. 2013). The feedback content on multiple web pages included a breakdown of scores, stations and skills, and guidance on reflecting and responding to feedback. But it did not include the examiners' comments. Students were able to access the feedback online a few hours after receiving pass-fail decisions. All students in year three (N = 138) were involved. Nearly all of them (95.7%) viewed their feedback. However, a quarter of students (27%) accessed only a few pages on the website. As the website contained pass-fail information and a breakdown of marks for each station, these students may have been visiting the site only out of curiosity to see which stations they passed. It would be interesting to find out how students engage with a website containing only the examiner comments without any pass-fail information or marks. High performing students used the website more thoroughly than the barely passing ones. This was the only paper the researcher was able to find on learner receptivity to OSCE feedback in a

summative context. The content of feedback appears to have become more elaborate by 2015 at the same school. The examiners were able to use digital tablets to tick off pre-written feedback items and record verbal comments (Harrison, Molyneux et al. 2015). Students were able to access them along with other feedback on the website. The learners and examiners were surveyed to evaluate the intervention's acceptability, which was found to be high. Both papers were written by the same author at the same location. All students appeared to be subjected to the intervention as it was a change in the school's assessment and feedback programme. Measurement of the learners' website view rates and user satisfaction were objective and valid. Using simple descriptive statistics was appropriate to the aim of the studies. Therefore the findings from these papers were taken at face value.

A more recent paper compared the quality and quantity of examiner feedback comments generated by the use digital marking and paper marking in large scale summative and formative OSCEs (Denison, Bate et al. 2016). The researchers rated the quality of the examiners' comments using a pre-determined criteria. They found a significant increase in quality ($p<0.001$) and the quantity of comments, in terms of word count, increased by approximately 10%. The increase in the number of comments for students who barely passed the exam was approximately 22%. In addition, the researchers reported that the improvement in efficiency was ***"well in excess of (Denison et al's) expectations"*** (Denison, Bate et al. 2016). Feedback and raw marks were available within minutes of finishing the OSCE. Tablet devices were individually expensive, but are reusable for multiple purposes. Overall cost effectiveness was improved especially for large exams which used to consume larger numbers of single-use computer-readable marking sheets (Denison, Bate et al. 2016).

Another study saw the use of standardised patients (SPs) as examiners in formative assessments and then piloting in a partially summative one (Bord, Retezar et al. 2015). The marks from the OSCE accounted for twenty percent of the whole year's marks used for making pass-fail decisions. Although the authors did not explicitly refer to programmatic assessments, this approach made their OSCEs relatively lower stakes summative assessments like those in programmatic assessment. The examiner and the SP gave the students immediate verbal feedback after completing the station in an emergency medicine clerkship. Eighty students were involved. Bord et al stated that this OSCE format served its purpose well in their context but they only described the process without any evaluation. While it may be feasible for standardised patients (SPs) to give immediate feedback in low stakes OSCEs in a clerkship with relatively small numbers of students, this approach is unlikely to be feasible in a large scale summative OSCE.

A similar innovation was done using near peers (Reiter, Rosenfeld et al. 2004). Final year students assessed and gave immediate verbal feedback to first years in formative OSCEs. The candidates found the feedback useful. However, there was no exploration to why they found it useful or an objective assessment of the quality of feedback. Reiter et al proceeded to compare the near peer examiners' marks to the expert examiners' marks to assess the suitability of final year students as examiners in summative OSCEs. The statistical methods for analysis were clearly stated and they were suitable for the study's purpose. The near peers' marks were significantly higher than those given by faculty examiners (Reiter, Rosenfeld et al. 2004). Therefore, Reiter et al concluded that while the final year clerks' feedback was perceived by the first years to be useful in a formative context, it was not applicable in a summative one.

Another study evaluated the feasibility of using team based assessment as a summative workplace based assessment (Sharma, Cui et al. 2012). Students were given assessment forms for various clinical skills. They were responsible for asking the members of faculty to assess them during their clinical placement. The faculty can mark them and give immediate verbal feedback using the assessment form. This approach replaced the end-of-the-placement summative clinical skill assessments. This study used a simple user acceptability survey to show that this approach to summative workplace based assessment was feasible and acceptable.

The earliest paper the researcher identified was published in 1986 by R.M. Harden, the educator credited for pioneering OSCEs. Harden et al's study found that immediate verbal feedback by examiners in a summative OSCE did not cause the students undue stress (Black and Harden 1986). In fact, a survey on the 37 participants in his study showed that they preferred to receive feedback immediately. Another study in 1989 supported this finding and added that students who received immediate feedback performed better in a repeat OSCE which immediately followed the feedback session (Hodder, Rivington et al. 1989). This study compared the performances of a control group and an intervention group, each consisting of 50 students. However, it was not clear whether the Hodder study was done in a summative OSCE. Both the 1986 Harden study and the 1989 Hodder study used four minute OSCE stations with two minute gaps in between for feedback. Both studies were done in the UK. They did not consider implications in larger exams with hundreds of students. They also did not clarify what assessment methods and criteria they used to assess the student's overall clinical competency. This is important because it was common in UK medical schools to use a combination of long case, short case and OSCE marks to determine pass-fail decisions in the late

1980s and early 1990s (Fowell and Bligh 1998). Therefore, the summative OSCEs in these studies may have been at lower stakes than those of today. The researcher was not able to find any more recent evidence to support the use of immediate verbal feedback in summative OSCEs although it found common use in formative OSCEs (Hill, Guinea et al. 1994, Wik, Myklebust et al. 2005). Although both studies were of good quality, their findings were not applicable to the present study as the context of summative clinical skill assessments has changed over the decades.

These are the only papers of acceptable quality the researcher was able to find which explicitly addressed feedback after summative clinical skill assessments. The evidence body is small and relatively basic as all studies were of cross sectional research design. Higher level of evidence such as quasi experimental or randomised controlled trials were not found. Further evidence was searched for in the formative context on the approaches to optimise the content of feedback, timing, mode of communication and feedback giver.

2.5 Feedback after formative clinical skill assessments

2.5.1 Timing and mode of communication

Timing and mode of communication are intertwined with each other. No study reported delivering written feedback immediately although verbal feedback was delivered either immediately or after a period of time. Delayed verbal feedback was often involved a faculty member and a student watching a video of the student's encounter with patients. In the wider educational literature, immediate feedback was recommended wherever possible so that a learner has more time and opportunities to refine their performance (Anders Ericsson 2008). However, in the context

of clinical skill assessments, only a few studies explored the effectiveness of different timings and modes of communication.

2.5.1.1 Immediate verbal feedback

A quasi-experimental study which used trained simulated patients to give feedback found that students who received immediate feedback performed better in a re-assessment after four months (Hoefer, Sterz et al. 2017). However, this evidence is questionable because the experimental comparison was between the students who received feedback to those who did not receive any. It was not a comparison between different feedback timings. Possible confounding factors were also not considered. Another study considered the educational impact of immediate verbal feedback. A small randomised controlled trial (N = 24) showed that the immediate verbal feedback by simulated patients at formative assessments was as effective as faculty teaching in improving history taking and consultation skills (Hughes, Wilson et al. 2016). They randomly assigned participants into two groups. However, being an educational intervention, it was not possible to blind the participants or the faculty to the allocations. It appears that the two groups were treated equally other than the intervention. Their outcome measurement and comparison were valid as they used a summative OSCE after a few months to compare the performance of the two groups. However, it was not clear if the outcome assessors were blinded to the group allocations. Statistical analysis was suitable and the conclusions were logical. However, as the sample size (N = 24) was small, it is difficult to determine if their results were generalizable. Although the researcher was unable to find more studies which assessed the educational impact of formative immediate verbal feedback in clinical skill assessments, a large

body of evidence supports its feasibility and acceptability (Turan, Uner et al. 2011, Schlegel, Woermann et al. 2012).

2.5.1.2 Delayed verbal feedback

Delayed verbal feedback was found to be effective at helping students show empathy in clinical communication (Kalish, Dawiskiba et al. 2011). Delayed verbal feedback with video received high user satisfaction rating in a study which involved a tutor and students watching their videoed clinical encounters together (Perron, Louis-Simonet et al. 2016, Perron, Louis-Simonet et al. 2016).

2.5.1.3 Delayed written feedback

Similar to the findings in summative assessment, delayed written feedback to students using digital devices in formative assessments showed high acceptability and feasibility (Treadwell 2006). From 2004 to 2006, a study in Pretoria, South Africa explored the usability of Personal Digital Assistants (PDAs) in OSCEs (Treadwell 2006). PDAs were handheld digital devices similar to the IPad^{TM}s with small display screens. Data could be typed into them using a physical keypad. Treadwell et al's study was conducted over three years. It compared paper and PDA assessment formats. The two cohorts' sizes and demographics were similar and were recruited from the same population. The outcome measurements for feasibility, reliability, efficiency and acceptability were valid and reliable. Appropriate statistical methods were used for analysis. This study's overall quality was good although confounding factors which may have affected the results were neither explicitly sought for nor strategies to deal with them mentioned. Using PDAs for marking in formative OSCEs was reliable because there was high correlation between examiner marks generated by paper marking and digital marking (Treadwell 2006). It was also more efficient than paper marking as the administrative processing time after

the exam was shorter. The acceptability was high according to a survey on the students and the examiners. These findings further support the use of digital technology.

Another study considered if differences in timing and feedback giver has an impact on digital feedback. A strong interrater reliability and concurrent validity was seen between immediate verbal feedback by a simulated patient and delayed written feedback by a faculty member (Ferenchick, Solomon et al. 2013).

Written feedback has been associated with portable digital devices in medical education since the technology became widely available (Snodgrass, Ashby et al. 2014). Digital devices used to be neither fast nor easy enough to record written comments during assessments (Kneebone, Bello et al. 2008), prompting the educators to collaborate with the technological industry to create mobile devices suitable for education (Kramer 2009). Due to the advancement of technology, more recent studies have shown that the use of digital marking and online feedback is feasible and acceptable, not only for written data but also for voice recordings (Coulby, Hennessey et al. 2011, Stone 2014, Harrison, Molyneux et al. 2015). Students found that technology could serve as an icebreaker in approaching patients and faculty for workplace based assessment (Coulby, Hennessey et al. 2011). It also found common use in workplace based clinical assessment and teaching, exhibiting high feasibility, reliability and user satisfaction in large scale studies (Torre, Simpson et al. 2007, Ferenchick, Solomon et al. 2013).

2.5.1.4 Delayed verbal feedback with video

Another widely used approach to formative assessment is video feedback. Its acceptability and feasibility is well established in the formative assessment context (Garry and Stirling 2012, Lee, Hwang et al. 2013,

Hulsman and van der Vloodt 2015), although the researcher was not able to find any evidence to support its use in a summative context. This approach allows various feedback givers, such as the faculty, simulated patients and peers, to view the learner's performance and give feedback at their convenience. It creates an opportunity for a two way feedback discussion, rather than a one way delivery from the giver to the recipient (Hulsman and van der Vloodt 2015).

However, to the best of the researcher's knowledge, there is no evidence that the content of feedback generated after watching a video is better than immediate feedback or delayed written feedback. In one study, two different patient instructors were found to generate largely the same feedback at two points in time; one immediately after the assessment and another a few weeks later after watching the video (Yudkowsky, Downing et al. 2004). Learner satisfaction was equally high for both. Video feedback was found to have no effect on the accuracy of the learners' self-assessment. One study used the Google Glass™ to record the patient's first-person view of a students' communication skills (Hoonpongsimanont, Feldman et al. 2018). Google Glass™ were spectacles with small cameras which could record videos by voice command. Watching a first-person video of themselves did not make any difference to the learners' self-ratings before and after the intervention. However, there were multiple confounding factors that could have influenced this finding, such as the content of feedback and the learners' receptivity to feedback.

2.5.2 Feedback giver

An assessment is valid only when the assessor possesses a first-hand knowledge of the subject and has observed the student's performance (Wilkinson and Tweed 2018). Educators are known to give feedback without actually observing the students' performance in formative

workplace based assessments (WBAs) (Daelmans, Hoogenboom et al. 2005) although such problems have not been reported in medical school assessments such as OSCEs. In addition, the feedback giver should be committed to the process with the intention to improve the learner's performance towards predetermined goals (Ende 1983).

2.5.2.1 Non-faculty examiners

The researcher was unable to find any evidence of using non-faculty examiners in summative assessments but a large body of evidence supports the feasibility non-faculty examiners, such as simulated patients (SPs) and medical students, to mark and give feedback formatively (Mavis, Turner et al. 2006, McLaughlin, Gregor et al. 2006, Moineau, Power et al. 2011, Vogel, Nel et al. 2012, Burgess, Clark et al. 2013, Young, Montgomery et al. 2014). Peers and SPs are able to observe and complete detailed checklists reliably (Yudkowsky, Downing et al. 2004). However, multiple studies have found that non-faculty examiners have a tendency to give higher marks and or global impression ratings than the faculty do, which weakens their applicability a summative context (Reiter, Rosenfeld et al. 2004, McLaughlin, Gregor et al. 2006, Burgess, Clark et al. 2013).

In workplace based assessments, learners were often given an assessment checklist and made responsible for approaching multiple assessors for assessment and feedback (McLeod, Mires et al. 2012). The checklists provide assessors a framework to objectively assess and give feedback to the student. Similar to a log book or a portfolio, this approach encourages interaction between learners, patients and educators and forces the learner to act more proactively in the feedback process. In formative mock OSCEs in medical schools, senior students were often used as assessors and simulated patients for junior students. One study

found that the students did not mind the extra work but they appreciated such arrangements (Young, Montgomery et al. 2014). However, students have reported concerns about bias due to pre-existing relationships between peer examiners and candidates and expressed doubts about the peer examiners' ability to notice everything the candidates have done (Moineau, Power et al. 2011). As such, the junior students did not view the senior ones as suitable examiners for summative assessments while they are highly acceptable in formative assessments (Burgess, Clark et al. 2013, Burgess, Roberts et al. 2013).

2.5.2.2 Faculty examiners

In terms of quantity, specialist trainees are known to give undergraduate students more immediate verbal feedback than other examiners in the workplace such as consultants and junior doctors do (Fernando, Cleland et al. 2008). Another large scale study for workplace based assessment showed that provision of feedback by all assessors was poor overall but academic trainees were the most consistent providers of feedback (Fernando, Cleland et al. 2008). This may be because medical students felt more comfortable to approach relatively junior specialist trainees than other assessors in the workplace while the latter were willing to assess and give feedback. However, there is evidence to the contrary. In Daelman et al's qualitative study in the context of implementing a novel formative workplace based assessment, neither the students nor the clinical staff actively tried to engage with each other despite the implementation of a formative WBA programme (in-training assessment or ITA) (Daelmans, Hoogenboom et al. 2004, Daelmans, Hoogenboom et al. 2005). Students felt the specialist trainees usually seemed to be busy with their own work although they did their best to address any perceived weakness of the students they have assessed (Daelmans, Overmeer et al.

2006). These studies showed that provision of feedback for workplace based assessments by the faculty varied in quality and quantity. This further highlights the need to evaluate the quality and quantity of feedback once the intervention is implemented.

2.5.3 Content

The content of feedback is a major factor which influences learner engagement with feedback (Winstone, Nash et al. 2016). However, only a few studies have objectively assessed the quality of their feedback content beyond user satisfaction surveys (Fernando, Cleland et al. 2008, Harvey, Radomski et al. 2013, Denison, Bate et al. 2016). As discussed above, it is common to use pre-written checklists to guide examiners mark and give feedback formatively. These checklists shape the content of feedback but, most of the time, their validity and the reliability are established only by internal discussions between relatively small numbers of educators at individual sites without following any validation framework (Mavis, Turner et al. 2006, McLaughlin, Gregor et al. 2006, Halaas, Zink et al. 2007, Eva, Munoz et al. 2010, Moineau, Power et al. 2011).

A consensus opinion is that effective feedback is based on observations, given in a conducive environment, characterised by a non-judgmental approach, focused on specific skills and behaviours, and limited in quantity to avoid overwhelming learners (Kelly and Richards 2019). In other words, the quality of the content of feedback depends on its specificity, length, accuracy, and neutrality.

2.5.3.1 Specificity

Feedback being too generic or including no specific suggestions for improvement a common complaint by learners (Rauf, Shamim et al. 2014). The content of feedback should be concise. Shute's review of the literature for formative feedback called for feedback to be kept relatively

simple because vagueness leads to a higher cognitive load for learners, which reduces the likelihood of understanding and learning from it (Shute 2008). Directive feedback, in which detailed guidance is given on what needs to be changed and which strengths should be maintained, is useful for novice or struggling learners, whereas facilitative feedback, which comprises comments and suggestions, is suitable for more experienced or higher performing learners (Shute 2008). Giving short examiner feedback directly to students is in line with theories about cognitive load, which demonstrate that presenting learners with a lot of information in too complex a manner is counter-productive (Sweller, Van Merrienboer et al. 1998).

However, evidence shows that even if feedback includes specific recommendations for improvement, the learners do not always agree with them and are more likely to set their own learning goals according to their pre-existing beliefs (Eva, Munoz et al. 2010). Learners do not like receiving feedback which challenges their self-perception (Mann, van der Vleuten et al. 2011). Low performing students tend to set learning goals focused on their strengths to compensate their weaknesses rather than addressing the weaknesses (Chang, Chou et al. 2011).

The quality of feedback in WBAs is often poor in terms of its specificity (Fernando, Cleland et al. 2008). WBA feedback does not often include suggestions for improvement, action plans or learning goals (Fernando, Cleland et al. 2008, Harvey, Radomski et al. 2013). Using digital technology was found to help improve the specificity of feedback. Denison et al's study compared the quality and quantity of feedback generated by digital marking to that of paper marking in large scale summative and formative OSCEs, involving up to 3802 individual assessments in each exam (Denison, Bate et al. 2016). They developed a five point rating scale

to assess the quality of feedback comments, ranging from judgmental or non-specific to objective appraisal of performance with specific suggestions for improvement. The specificity of feedback generated using digital tablets was significantly higher than that of paper marking.

2.5.3.2 Length

Any feedback during an OSCE is likely to be short as the available time is limited. As discussed above, findings from a literature review (Shute 2008) and the cognitive load theory (Sweller, Van Merrienboer et al. 1998) support the use of short and simple feedback comments. Evidence shows that learners engage well with short written comments (Lipnevich and Smith 2009) while they may lose interest with long complex ones (Shute 2008).

2.5.3.3 Accuracy

Whatever the content of feedback is, it needs to reflect what happened at the exam accurately. Inaccurate feedback is invalid and it could be detrimental to learning. Several studies investigated the accuracy of the non-faculty feedback by having a faculty educator as control (Yudkowsky, Otaki et al. 2009, Burgess, Roberts et al. 2013). Those studies found that both types of assessors were able to produce accurate feedback.

2.5.3.4 Neutrality

Even if feedback is based on accurate facts, it could include positive or negative distortions to the message. As summative feedback is accompanied by marks or grades, it is bound to elicit emotions. Grades and marks elicit strong emotions even in a formative context (Lefroy, Hawarden et al. 2015). Lefroy et al's qualitative study found that grades are concrete, powerful and blunt and they needed to be explained to help students create helpful meaning from them. Believing low grades may reduce the learners' self-efficacy, feedback givers tend to positively distort

their message (Lefroy, Hawarden et al. 2015). Qualitative evidence has shown that students have concerns about passing judgement on their peers' performance (Papinczak, Young et al. 2007). While the students in the Papinczak et al's study felt increased responsibility for others and improved learning due to formative peer assessment, they also felt discomfort and expressed concerns about its negative impact on the cooperative, non-judgmental atmosphere of their tutorial groups. Such perceptions were not confined to peer examiners. Faculty examiners are also known to positively distort their feedback as they felt bad grades were demotivating (Daelmans, Overmeer et al. 2006).

Although both faculty and peer examiners try to avoid causing negative emotions, one study found that any increase in self-efficacy and learner satisfaction due to positively framed feedback faded away after a short time (van de Ridder, Peters et al. 2015). Distortion of the feedback message has no lasting educational impact, provided that the basic facts are accurate (van de Ridder, Peters et al. 2015). Therefore, feedback should be honest and straightforward.

2.5.3.5 Innovations

Several studies developed and evaluated innovative feedback contents. One study explored the impact of feedback after confidence weighted testing (Aoyama, Tamura et al. 2013). Confidence weighted testing involves asking the student how confident they were with their answers and actions, and marking their performance according to the confidence level they stated beforehand. Feedback included the marks, confidence level and written examiner comments. This approach is aimed to help learners become conscious of their competence. Students who received this type of feedback showed a significant improvement in clinical skill performance at the year-end summative exams (Aoyama, Tamura et al.

2013). However, it was impossible to determine if the difference is solely due to the feedback as the paper did not mention any measures for control. Another study followed a similar approach aimed at producing practitioners who are aware of their own competence. In this study, the learners had to propose a differential diagnosis after reading the clinical scenario before assessing the patient (Yudkowsky, Otaki et al. 2009). Feedback included the correct diagnosis and another chance to perform the relevant examination. However, they did not describe any evaluation to assess the educational impact of the intervention.

2.6 Discussion

The researcher reviewed the evidence on the best approaches to timing, mode of communication, feedback giver and content regarding feedback after clinical skill assessment. The body of evidence is mainly composed of cross-sectional or descriptive studies.

In terms of timing and mode of communication, evidence in the summative context supports the use of digital technology for delayed written feedback (Harrison, Konings et al. 2013, Harrison, Molyneux et al. 2015, Denison, Bate et al. 2016). However, it should be noted that the body of evidence on summative feedback is small. Digital marking has several advantages over paper marking as feedback comments on paper are often poorly constructed and illegible due to time constraints, and often required tutors to help the students interpret the comments (Harrison, Molyneux et al. 2015). Typing alleviates problems caused by bad handwriting. Furthermore, the increase in efficiency and cost effectiveness is appealing (Denison, Bate et al. 2016). Although an initial investment is required to buying the devices and the software, they can be reused. The same cannot be said for most of the assets in paper marking. Handling large volumes of confidential documents is labour

intensive both before and after the exams. As the available evidence supported the advantages of digital marking, the researcher decided to use digital tablets to enable examiners to mark and write short feedback comments for students during summative OSCEs and then to deliver written feedback to students online after the OSCEs.

In terms of examiners, faculty examiners are the only acceptable option for marking and feedback in a summative context at the University of Nottingham. As discussed above, non-faculty examiners tend to be too generous with marking for use in high stakes OSCEs. Their credibility and acceptability is questionable in high stakes assessments although evidence supported their ability to provide adequate formative feedback. Adequate examiner training is required for generating specific, accurate and honest feedback with minimal positive or negative distortions.

In terms of content, any written feedback during the OSCEs would likely have to be short as the time for the examiners is limited. Evidence supports the educational value of simple, short feedback (Sweller, Van Merrienboer et al. 1998, Shute 2008, Lipnevich and Smith 2009).

No recent evidence supports the use of immediate verbal feedback in summative clinical skill assessments although it is commonly used in formative assessments. Regarding delayed verbal feedback, it would be a challenge to arrange hundreds of individual meetings and the examiners may not be able to recall what happened at the exam for every student. The researcher was not able to find any evidence supporting this type of feedback. There was no evidence to support the educational benefits of delayed video feedback either. However, digital devices were used to record the examiners' verbal comments and to deliver them to students online successfully in one study (Harrison, Molyneux et al. 2015). It was not possible to replicate this because there was no commercially available

digital software to deliver thousands of voice recordings by dozens of examiners to hundreds of students. Available funding, time and expertise also did not allow for developing a new software for this study.

Once feedback is delivered, it is necessary to evaluate its quality and quantity distribution across the student cohort as evidence showed that these were often inadequate even in low stakes formative assessments. Furthermore, as both digital marking and written feedback during OSCEs were new introductions at Nottingham in 2016, examiners were not familiar with the technology or feedback giving. There had not been any standardised training for the examiners regarding feedback or marking before 2016. Therefore, there is a possibility that their written feedback may be of poor quality: it could be illegible, poorly constructed, unspecific or even unprofessional. There also could be disproportionate inequalities in the distribution of feedback across the student cohort. Analysing the quantity distribution of feedback would help us understand more about the examiners' feedback giving behaviours in a summative context, a previously unexplored topic in medical education.

2.7 Summary

It was concluded that digital technology is the most suitable tool for this study's intervention. Evidence shows that examiners can produce short written feedback comments during summative OSCEs using tablet computers and deliver personalised feedback to all students online after the exams. However, as feedback quality and quantity is often poor, it is necessary to evaluate these factors once the programme has been implemented.

Chapter 3. Interview study

3.1 Introduction

A rationale was presented in Chapter 1 for the need to consider the learners' perspective when designing a new feedback intervention. According to situativity theory, learning related constructs such as knowledge and thinking cannot be separated from the context in which they took place, but in fact, they are dependent upon it (Durning and Artino 2011). As this study was aimed to identify the perceived needs of the learners regarding feedback after high stakes summative assessments, the researcher investigated the learners' perception of the feedback they received and the assessments they were subjected to at the time in 2015 at the University of Nottingham. However, non-assessment related feedback, such as feedback during classroom and clinical workplace based teaching, was not investigated. The purpose of this study is to gain an insight on the student perception in order to inform the design of the feedback intervention. This study forms the qualitative part of this mixed method research project.

3.1.1 Research questions

The research questions for this qualitative study were;

1. What do students think of the feedback they receive in the context of the current assessment programme?
2. What kind of improvements do they want for the future?

3.2 Methodology

Methodology is strategy to frame and guide research. This study follows the tradition of transcendental phenomenology as it was the most appropriate to answer the research questions. A brief description of phenomenology and other qualitative traditions is presented below to explain the rationale of choosing this tradition.

3.2.1 Qualitative traditions

The most common traditions in qualitative research include; phenomenology, which focuses on rich descriptions of the lived experiences of the participants on the topic of interest; narrative, which focuses on the lives of individuals as told through their own stories; grounded theory, which focuses on inductively analysing data to generate a theory; ethnography, which focuses on observing and or interacting with participants in their real-life environment to gain an insight into their culture; and case study, which focuses on a detailed analysis of the individual participant's data (Creswell and Poth 2016).

Each tradition has distinct disciplinary roots and unique features. However, the primary aim of qualitative research regardless of the traditions is to understand the participants' perspective. Consequently, the typical product of qualitative research is rich description.

3.2.2 Phenomenology

Phenomenological research stems from philosophy. Several authors have presented phenomenology as a single approach (Merriam and Tisdell 2015, Bryman 2016). However, three distinct approaches could be identified within the tradition of phenomenology; transcendental, hermeneutic and existential (Eddles-Hirsch 2015). All three follow the four primary phenomenological concepts; description, reduction, imaginative variation and essences (Moustakas 1994). However, they have different assumptions to research.

Transcendental phenomenology forms the basis of all the other phenomenological methods (Eddles-Hirsch 2015). Husserl, a mathematician turned philosopher, widely considered to be the founder of phenomenology, asserted that humans attach meanings to external stimuli. Human's consciousness and the world are one and therefore could

not be studied in isolation from one another. Therefore, a social phenomenon should not be studied as a separate entity. A "lived experience" should be studied instead (Laverty 2003). The focus in phenomenology is on the world as lived by a person, but not the world, the person nor the reality separately on its own (Valle, King et al. 1989, Merriam and Tisdell 2015).

The mental relationship between a human and the world around them is termed as "intentionality". In Husserl's perspective, every intentional experience consists of a "noema" and a "noesis" (Moustakas 1994). The noema represents the objective experience, whereas the noesis represents the subjective experience. In the present study, the noema constitutes the exams and the feedback the students received after exams. The noesis relates to how the students felt about it, their opinions and their wishes. By considering the meanings of intentionality, the underlying structure or essence of a phenomenon could be discovered.

In order to ensure the validity of the findings, the transcendental phenomenologist must purposefully set aside any preconceived knowledge or everyday beliefs that might be used to explain the phenomena being investigated. This is termed "epoche" or "bracketing". This allows the researcher to listen and record the participant's description of an experience in an open and naive manner (Moustakas 1994). In addition, explicitly describing the researcher's background and role in the context of the study helps the reader understand how the researcher came to the conclusions. In summary, transcendental approach is descriptive and it assumes that it is possible to suspend and or clarify the researcher' pre-existing knowledge.

In contrast, Heidegger believed that it is impossible to set aside one's own presuppositions and beliefs (Eddles-Hirsch 2015). Rather than taking a

detached observer position, hermeneutic phenomenology takes a purposeful recognition of both the researcher's and participants' perceptions. Hermeneutic phenomenology takes an interpretive approach, rather than a descriptive one as the goal is to create meaning and achieve a sense of understanding (Laverty 2003). The researcher attempts to make an interpretation from different meanings deduced from the participants' lifeworld experiences. The hermeneutic phenomenologist attempts to understand why participants perceive the phenomenon the way they do while explicitly recognising that the explanation is shaped partly by his or her own pre-existing knowledge.

The existentialist approach differs from others in the interpretation of intentionality, the mental relationship between a human and his or her world, based on the philosophical perspective of Sartre and Merleau-Ponty (Eddles-Hirsch 2015).

3.2.2.1 Rationale for phenomenology

This study's research questions asked primarily the "what" of the learners' experience and perspective. Therefore, transcendental phenomenology, which focused on the contextualised rich descriptions of the phenomenon, rather than the interpretative approaches, was chosen.

3.2.2.2 Analysis steps

There are specific strategies to transcendental phenomenological analysis.

The first step recommended by Moustakas in the phenomenological reduction process is horizontalization of the data (Moustakas 1994). Horizontalization is a process in which the researcher keeps an open mind and be naively receptive to every statement of the participants' experience, granting each comment equal value (Moustakas 1994, p.122). This step is related to the epoche or bracketing process, in which the

researcher purposefully sets aside any preconceived knowledge. To fully inform the readers of how the researcher analysed the data, the researcher's pre-existing knowledge, attitude and position in the context of the study is to be clearly described. All repetitive and or irrelevant statements are then removed. The criteria for determining relevancy of a statement is whether it contains a necessary and sufficient expression for understanding the experience and whether it is possible to abstract and label it. Then the statements that referred to the phenomenon are lifted out of the transcript and recorded separately, and clustered together into similar meaning units. These statements were then referred to as the horizons.

This is followed by the imaginative variation process, in which the researcher identifies the structural themes from the textural descriptions. Imaginative variation requires the researcher to see the phenomenon from a variety of perspectives, so that they can understand the essence of the participants' experiences. It is vital that the themes originate directly from the first hand experiences of the participants, rather than the researcher. To ensure this, the research could have a meeting with the participants to ask if the themes reflected their opinions adequately and make amendments where necessary. This process is termed as member checking. It was not advocated by Moustakas, but many other researchers believe that member checking strengthens the validity of qualitative findings (Eddles-Hirsch 2015).

Finally, a statement was written based on the composite descriptions that reveal the essences of the phenomenon of interest. Moustakas (1994) stated that these essences are never truly exhausted, but simply represent one researcher's perspective at a particular time and place. The researcher followed these steps as closely as possible.

3.3 Methods

3.3.1 Peer interviewing

This study used peer interviewers to collect data from other participants. There is evidence on the use of people with lived experience to conduct interviews with their peers. This approach has been used mainly in research related to hard-to-reach and or disenfranchised groups such as sex workers, refugees and homeless people (Devotta, Woodhall-Melnik et al. 2016).

At the core of the peer interviewing is the concept of "expertise by experience" (Gallacher and Gallagher 2008, Frankham 2009). Peer interviewers, by the virtue of their lived experience, are believed to have information that traditional researchers or outsiders cannot readily access (Nind 2017). Proponents of this concept believe that one's identity creates a particular world view which then produces valid knowledge because it is grounded in the experiences of those being researched (Nind 2017).

This is an approach to envision a more nuanced understanding of complex lives or phenomena (Nind 2017). Peer interviewers have the potential to discover knowledge that may otherwise go unnoticed by researchers without lived experience. Peer interviewers are viewed by their community as 'similar but different': similar because they have lived experience, yet different from the participants in their role as interviewers (Edwards and Alexander 2011). Participants are more likely to respond genuinely to people they know, although this is not a guaranteed outcome (Devotta, Woodhall-Melnik et al. 2016).

Proponents also argue that it provides tangible benefits for both researchers and peer interviewers (Elliott, Watson et al. 2002, Kilpatrick, McCartan et al. 2007, Remple, Johnston et al. 2007, Greene, Ahluwalia et al. 2009, Kelaher, Warr et al. 2010). The former benefits through

increased greater facilitation of access to target populations, increased comfort of participants in the research process and consequently, higher quality and quantity of data. The peer interviewers benefit through skills development and incentives.

However, there are practical and ideological challenges to using this approach. Quality of the data partly depends on the interviewer. Peer interviewers may require comparatively more training for their interviewing skills to ensure adequate data collection and ethical principles (Benoit, Jansson et al. 2005). Another challenge is that researchers and peer interviewers may have different opinions on research questions and the data (Delman and Lincoln 2009). There is also an inherent power imbalance between the researcher and the peer interviewers. These factors contribute to the concerns about trustworthiness of the data (Nind 2017). To address the power imbalance, the researchers need to acknowledge their relationship with the peer interviewers, social positions, possible biases, motivations for conducting research, and how these influence the research process. This is part of bracketing in phenomenology. Bracketing does not reduce the power imbalance but it supports the trustworthiness of the data by informing the reader of the researchers' positions.

Considering the advantages and researcher's confidence to mitigate at least some of the challenges, peer interviews were used in this study.

3.3.2 Participants

The researcher invited student union representatives via email. Two students, R.C. and T.H., volunteered to be interviewed and then act as peer interviewers.

The researcher, M.M.S., interviewed them and trained them to conduct semi-structured interviews by themselves using a pre-written interview guide. Afterwards, R.C. and T.H. recruited the rest of the participants and interviewed them independently. To clarify, the researcher, M.M.S., interviewed R.C. and T.H. only. They went on to interview the rest of the participants.

The main reason for this approach was to gain more honest responses from the participants. R.C. was a foundation year one doctor (FY1) and T.H. was a year three student (Clinical Phase 1 or CP1). They had both completed their summative exams recently at the time of the interviews and recruited their friends who had lived the same experience. Through familiarity and shared lived experiences, it was hoped that the peer interviews would generate evidence which is more reflective of the participants' true feelings than the traditional researcher-led interviews. Another reason was to attract more participation as it was difficult to recruit participants without any monetary incentives. However, this approach has inherent limitations, the most obvious of which are selection bias and interviewer bias. A more detailed reflection on these limitations is presented in this chapter's discussion section.

The training for interview skills for R.C. and T.H. included reading through the semi-structured interview question guide together with M.M.S., and reflecting on the interview they had with M.M.S. They were explained about the rationale and the nature of semi-structured interviews in the context of this study. Interviewers were instructed to get the interviewees' informed written consent for the interview and digital recording. The interviewers were also instructed to inform interviewees that participation is totally voluntary and the participant may choose to stop the interview at any moment. They were encouraged to allow their interviewees to raise

issues and expand on their responses, rather than adhering strictly to the interview question guide. The interviewees were to be allowed to talk about anything they wanted as long as it remained relevant to the original aim of the study.

The purpose of recruiting students from different years was to explore learner perceptions at different stages of training. It was a sample of convenience as the interviewers, R.C. and T.H., approached only those they knew personally. Participants were given background information about the study before informed consent was obtained. All participants gave written consent to attend a one-on-one interview, to allow recording of the discussion on a digital device and to disseminate the research as appropriate. Participation was completely voluntary and anonymous. The peer interviewers received recommendation letters for their active voluntary contribution. The rest of the participants (interviewees) received acknowledgement letters. No other educational or monetary incentives were given.

3.3.3 Data Collection

Approval from the Ethics Committee of the School of Medicine was obtained (reference number R14072015, 18 August 2015). See Figure 22 and Figure 23 in Appendix. One-on-one semi-structured interviews were conducted. Each lasted approximately fifteen minutes. Time and location were planned according to the interviewee's convenience with an emphasis on privacy. Interruptions were kept at minimum. All interviews were recorded digitally and transcribed verbatim. Participants' training year and gender information were tagged to the transcripts. The data were anonymised. Open-ended questions were used and interviewees were encouraged to elaborate their responses as long as they remained relevant to the original area of interest. The interview guide was designed

to allow for flexible exploration of topics and issues most relevant to students. Students were asked what kind of exams they had, what they thought about the exams and feedback they received and what they wanted in the future. None of the interviewers kept reflective diaries during data collection.

3.3.4 Timeline

2015 August	- Ethics approval. - Email invitation via student union - M.M.S. interviewed T.H. and R.C. and trained them to interview others.
2015 August-October	- T.H. and R.C. interviewed others and transcribed
2015 December	- Member checking meeting between M.M.S., T.H. and R.C.

3.3.5 Interview schedule

Four main topics were discussed.

1. Types of exams students had in the past year,
2. Types of feedback they received,
3. How they feel about it and
4. Type of feedback they want in coming year

The following is the question guide.

1. What exams did you have to do last year?
2. Tell me about how you felt about your most recent finals.
    1. What did you think about the difficulty of the exam?
    2. What did you think about the format of the exam?
    3. What did you think about the timing of the exam?
    4. What did you think about the exam location and examiners (if any)?
    5. Do you think the exam was fair?
    6. What is good about the exams?
    7. What could be done to improve the exams?

3. What do you think about the feedback you receive after exams?
    1. How do you get feedback at present?
    2. What is good about the feedback you get at present?
    3. Is the feedback you get useful? How?
    4. How could the feedback you get be improved?
4. How do you think feedback should be delivered?
    1. Who should get feedback?
    2. When should it be delivered?
    3. In what format should it be delivered?

As the purpose of this study was to gather information on what students think of the feedback they receive and what changes they want for the future, the interview questions focused on asking the student opinions of the exams, the feedback and their suggestions for improving feedback.

Firstly, the interviews focused on the context in which feedback was generated, i.e. the exams. The interviewers focused on the learner perception of their recent summative exams and what type of exams they had. It was necessary to do so as participants were recruited from different training years in which different types of exams and feedback were used.

Seven specific questions regarding the exams were devised for the interviewer to prompt the participants in case they were not forthcoming with the answers. When the interviews were conducted in 2015, the OSCEs were still using traditional checklists. Multiple OSCEs were held at several sites within the Nottingham University Hospitals NHS Trust and the Nottinghamshire NHS trust simultaneously, using the same stations. The school was making plans to move all OSCEs to a single site in 2016 and to change the marking scheme to domain marking, in which examiners rate the student performance on a five point scale on broad

domains such as communication and professionalism. In the context of the upcoming changes to the exam format, the interviewers asked students of their opinions on the current exam format, in terms of difficulty, timing, location, examiners and fairness. As domain marking in OSCEs gives examiners more freedom to use their judgement, it could be seen as less objective and consequently, less fair. Therefore, the interviewers asked students what they felt about fairness in the exams while the traditional OSCE checklists were still being used. Then the interviewers included questions on the perceived strengths and the areas for improvements for the exams.

These were followed by the questions about the learners' perception of feedback, the focus of research for this study. Firstly, the participants were asked how they received feedback and what was good and bad about it. Then they were asked for their specific advice on how to improve it.

As mentioned earlier, the items in the interview questionnaire were aimed to be used as prompts when necessary to keep the discourse on track and to elicit more specific responses. The interviewers were allowed to vary the questions depending on the participant. The interviewers were to allow the participants to expand on any issue they wanted as long as it remained relevant to the research questions. The question guide was used as a flexible framework to keep the discussion on topic.

3.3.6 Analysis

Firstly, the researcher read the transcripts repeatedly while listening to the recordings until he was familiar with the data. In line with the transcendental phenomenological concepts of horizontalization and bracketing, the researcher kept an open mind and granted every comment

equal value. The researcher did his best to be aware of and set aside any of his pre-existing knowledge and opinions.

To clarify the researcher's position in this study, the researcher, M.M.S., is a male junior doctor trained in Myanmar. The researcher received a MMEdSci degree in medical education at the University of Nottingham. The researcher was a full time PhD student at the time in 2015, never having worked in the NHS nor studied undergraduate medicine in the UK. The researcher has never been subjected to UK undergraduate medical exams or feedback. The researcher recalls having no personal feelings or opinions about feedback after assessments in the UON undergraduate curriculum, except for the feedback intervention he was planning. Prior to this study, the researcher's sources of information about feedback at UON were the medical school administrative staff, his supervisors and the school website. The two other interviewers were UON undergraduate students. They were not involved in analysis except at member checking. The researcher kept a reflective diary during data analysis to record what he felt every time he read the data. It helped keep track of the researcher's opinions for bracketing.

The researcher made notes on the transcripts and saved them in a separate document along with verbatim quotes. Similar notes were incorporated into codes. Data from the first four transcripts were open-coded. Codes were then revised and compared across the transcripts to generate themes, keeping in mind that the themes should originate directly from the transcripts rather than the researcher's opinions. New codes were incorporated into themes as analysis went on. They were also retroactively applied to the previously coded transcripts. Then the researcher checked for any possible relationships between the themes.

After the themes were generated from the first four transcripts, the interviewees and the researcher had a meeting to perform member checking. The researcher asked the interviewers, T.H. and R.C., if the themes adequately reflected their views to make amendments where necessary. They felt the themes adequately represented their own opinions and their impression of their interviewees' opinions. The researcher then coded subsequent transcripts into the finalised themes.

3.4 Results

Students at different stages of training had similar feelings and thoughts about the exams and feedback. Multiple factors such as the difficulty of the questions, venues, examiners and patients contributed to the students' perception of high stakes summative exams. Three main themes were identified; reasons for not using feedback, concerns about fairness and exam difficulty.

There were sixteen participants in total. Two peer interviewers interviewed fourteen participants they recruited. All sixteen participants belonged to the A100 course (undergraduate entry to medicine, Non-GEM). Students from three different years (Year Three, Final Year and Foundation Year) participated in the study. No one from the A101 GEM course participated in the interviews. Participants were; two males and three females from the year three, two males and three females from the final year, three males and three females from the foundation year one (FY1). Only one of the participants has had to take a resit for failing an exam.

3.4.1 Not using feedback

Students focused on this topic in the interviews. They took more time to discuss this than any other topic. Student opinions on current feedback were similar across all years. Most felt it was too generic and wanted more individualised feedback. The content of the feedback not being relevant for

the learners' future was also discussed. Students appreciated the effort required to produce feedback. Frustration at the individual meetings with the faculty after failing the exams was also discussed.

For written knowledge assessments, students received learning objective based feedback via ROGO. A third year student found it was ***"quite in depth"*** and useful ***"to a certain extent"*** but some others found it useless for them.

> ***"There is the traffic lights, the traffic lighting system for the ROGO exams but that is very difficult to access and not that helpful. And it does get quite confusing because the marks don't really correlate to what your actual mark is and… we've stopped using that to be honest."(Final year female 1)***
>
> ***"There's not really any feedback for the written papers, as far as I know. I think maybe you get some feedback on how well you covered the objectives. I, I think that's next to useless really cos… I know how well I know the objectives that cover… cardiology for example, I find… I need the question and I need to know what the right answer and the wrong answer and the reasoning was."(Foundation year one female 2)***

Most students did not have any suggestions to improve written exam feedback. Third year female 2 suggested letting students see their answer sheets again would be the only way to improve feedback.

Students focused more on OSCE feedback. All students discussed about it at length. For the OSCEs, students received individual station marks and average of the whole the cohort. All students felt it was too generic and not useful for passed students to improve although some felt it might be useful for those who failed. Students wanted more specific individualised feedback.

> ***"You get general pointers as to how a lot of people lost marks. So I like getting the score for each station but er written comments generally aren't that helpful….. The written comments aren't that useful because they're not that directly applied to you." (Third year female 1)***

> ***"Rather than knowing what the main areas of downfall for everybody was, finding out where you specifically went wrong would be a lot more helpful." (Third year male 2)***
> ***"The sort of written feedback was completely er generic and although potentially useful, particularly if you had failed the station, ……. you could compare yourselves to the average. Erm, but the, the sort of written feedback was completely er generic…. again the cranial nerves station I lost a mark and I can't for the life of me think of what I did wrong. Erm, so, what with there being no personal feedback to your erm, your set of exams, you have no way of knowing how to improve." (Foundation year 1 male 3)***

Final year female 2 failed her OSCEs and received individualised feedback. She was convinced she completed her task well and she should have passed. She said the feedback giver told her the examiner may not have heard or seen her completing her task and attributed her failure to bad luck. She found the face to face feedback session frustrating and useless.

> ***"I don't really know how to improve my performance from that. Erm, and I would app… I would have appreciated a more, err constructive answer. Erm and actual points I could work on as opposed to the examiner didn't hear you. Because I can't control that next time either." (Final year female 2)***

Some male students did not use the feedback they received. Their attitude was, as long as they passed, they did not care where they did well or not. They also felt it was irrelevant for the future.

> ***"Well it wasn't useful for me personally because I kind of saw that I passed the stations and then didn't really look into it any further" (Foundation year one male 2)***
> ***"I have to say I'm probably not the best in that if I've passed it I'm not that bothered about how, why, why I've got what I did……I don't know how that's gonna benefit me now sort of thing… Cos they're quite artificial situations, so even if they say oh you did this wrong, I don't know if that's that useful to apply to when you're actually working, a job" (Final year male 1)***

One felt feedback was sometimes not worth the effort. A notable finding was that students sometimes had to put in considerable effort to receive feedback, such as going to Mansfield, a town fifteen miles away from Nottingham, for face to face meetings with the faculty. They mentioned that if the feedback they get there was not satisfactory, it discouraged them to attend future meetings.

> ***"Mansfield's a pain to get to, and I've got to take time out of my current placement to go back to Mansfield. And then when I got there the feedback was just, I dunno, nothing really, it didn't give me any specific pointers; like I knew it wasn't a good exam, but I'd rather have had 'this is specifically why we failed you' it was just your history wasn't complex enough or something like that. And then his advice was come back to Mansfield a few more times and we'll do a bit more practice but yeah I didn't specifically know what had gone wrong from their point of view in that" (Final year male 1)***

3.4.1.1 Suggestions for improvement

All participants felt that every student should receive individualised feedback after exams whether they passed or failed. They did not like the school's policy to give feedback only to the failed students. They also felt new feedback should be delivered along with the results or within a few weeks after the results while ***"it's fresh in people's heads" (Third year female 3)***. They disagreed on the mode of delivery. A few students felt face to face sessions for everyone would be the best ideally. Most thought it would not be feasible in real life.

> ***"Organize a meeting with the module lead for example, erm which I think would be quite useful. And just getting the chance to discuss it with someone really – the way they give it is so generic it doesn't really help you."(Final year male 1)***

Most students wanted short, individualised written feedback to be delivered along with the exam results online. They felt it would be

achievable. Some said ***"even a couple of comments"*** would be better than what they were receiving at the time.

> ***"It would be good if you had individual feedback, but that takes a lot of time. But I think even if it was like a couple of comments that would be quite useful."(Final year female 3)***
>
> ***"For OSCE? I think just have one positive, one negative, maybe or both negatives in the feedback form applied to you. And then, for the other exams I don't think it would be possible really to do better feedback. The only thing you could do is to be able to see the paper again." (Third year female 1)***

Such feedback was given after OSCEs in the second year. Third year students found it useful and more senior students wanted something similar.

> ***"Gave us erm one piece of positive and one piece of negative feedback, like after each station. And that was quite consistent across every station so that worked well."(Third year male 2)***
>
> ***"I think for practicals some verbal feedback from the examiners would be, I think, really valuable; or a couple of written comments and I think they did do that, I seem to remember them doing that for the first and second year OSCEs. …..but then for the computer-based ones, I think some sort of report online that you could go and have a look at. I don't think there's any other way you could do that really."(Final year male 2)***

Students wanted the feedback giver to be someone who knew the subject and the student well.

> ***"I think ideally a face to face discussion,…… with someone who's known you kind of through med school. So obviously that can be difficult if you don't have a close relationship with your tutor, and they don't really know sort of what level you are generally performing at"(Foundation year one female 1)***

In summary, one type of students felt they did well enough by passing the exam and they did not care to improve. They did not reflect on the feedback they received and also did not feel it would be useful for them in

the future. They felt the current feedback programme may be useful only for those who failed. The second type of students wanted to improve. They did not find the current feedback useful and they wanted better feedback. They felt the feedback they received was too generic to be useful for them. They are also unsatisfied with having little feedback other than individual station marks for OSCEs. They wanted more individualised feedback to help them pass or get higher marks in the exams.

3.4.2 Concerns about fairness

Students were concerned about any potential factor that would unfairly put them at a disadvantage at the exams. It was an interesting finding because they were concerned about the possibility of unfairness even though no one experienced any problems first-hand. All students in this study felt that the examiners they had were fair, professional and competent, yet they felt the examiners' mood, competency and behaviour can potentially vary significantly. One student felt the OSCEs were subjective ***"by their nature"*** due to examiner and patient variability. They did not seem to appreciate the oxymoronic nature of their belief.

> ***"Erm examiners, like I said, with the OSCE it's all subjective, which is tricky." (Third year male 1)***
> ***"Examiner were all courteous to us and explained what we needed to do at each station rather well and gave us erm one piece of positive and one piece of negative feedback, like after each station." (Third year male 2)***

Final year students had psychiatry OSCEs. One of the stations required examiners to interact with students. Students worried the examiner may not act consistently. No one had any "bad" examiners but they heard about others having bad experiences.

> ***"For the OSCE I thought your score for a station could vary depending on if you had a harsh or nice examiner and if the examiner was in a good or bad mood. I thought that varied a lot." (Third year female 1)***

> ***"Examiner wise, I think they were fair to me, but other people complained about their examiners, so… "(Final year female 3)***
> ***"They both seemed quite fair… Errr, although I did hear between different sites and different exams there were some discrepancies between how much the examiners contributed to the scenario, I mean how much they prompted students. Erm this apparently did vary so perhaps they need some more training on making that even" (Final year female 1)***
> ***"They were nice. Erm, none particularly stick out in my head for being horrible examiners, which is probably a good thing" (Foundation year one male 2)***

One pointed out the issue where the examiner and student relationship may influence the marks.

> ***"Going into a station where you know the examiner and you've got a good working relationship with them is quite different to going into a station where you don't know them, or where you don't necessarily have such a good relationship with them – makes a big difference about how you feel about it. Erm, and… it it, I think it is variable about how much information an examiner gives you as well. Errr, and I know it's all moderated for, but still makes you… changes the way you feel about them." (Foundation year one female 3)***

The second concern was the patients. The OSCEs used both real and simulated patients. Paediatric OSCEs used real patients. Some students accept that patients vary in real life and it could not be helped in the exam. While students wanted realistic, clinically relevant scenarios, they did not like the variability of real life.

Some students wanted things which conflicted with each other. Final year female 1 wanted longer, more realistic OSCE stations. She was unhappy with time pressure and artificial feeling of the six-minutes-long stations. She was also unhappy with using real paediatric patients because they can vary and it is unfair. It seemed she wanted more realistic exams but she did not like the perceived unfairness of real patients in OSCEs.

> ***"OSCEs run the risk of being too artificial because of the time pressures." "Depending on the child that you get on the day it can, it can go in a completely different direction. You have no control over it. It seemed quite different at the different sites at which people were assessed." Final year female 1.***

Third year male 2 suggested patients should have a more comprehensive script to follow so that they would be more consistent.

Other students also voiced concerns on patient cooperation varied according to age, illness and severity. Seventeen years old patients cooperated better than younger severely disabled ones. While it makes the scenario more realistic, students felt it also makes the OSCEs unfair.

> ***"But having what I had, which was quite a severely disabled baby was, I dunno, to me it felt, it sort of threw me quite a bit so…" Final year male 1.***

Time pressure is another student concern in addition to patient variability.

> ***"I think is the patients obviously vary but that can't be helped. So, erm with the history taking station, I think it depends a lot on the patient, because if you get a patient who wants to really chat and speak to you a lot then you don't have time to ask all the questions. Whereas, if you get a less chatty patient, it's a lot easier." Third year male 2.***

3.4.3 Exam difficulty

Students found the exams difficult but they were ***"achievable", "straightforward"*** and at a ***"reasonable difficulty level".*** They felt being difficult to pass was the nature of high stakes exams. There also was a common belief that the exams were equally difficult for everybody. Final year students seemed to have found exams more challenging and stressful than second year students did. Students wanted more time for revision before exams. They felt it would make things less difficult.

> ***"I think they were quite difficult. They were erm, er it was very, it was not much time to revise…..they er could have been better timetabled so that people had more time to revise" (Third year male 1)***
>
> ***"I thought it would have been nice to have a week or so to revise beforehand. It was quite tough going straight from finishing a placement on a Friday to having a big exam on the Monday, but I suppose that's what it's like in real life so… it was difficult but most of us managed ok." (Final year female 1)***
>
> ***"I thought it was pretty fair… erm, there were some quite straight forward stations, some more difficult stations, erm, some that you'd be expecting, erm so I thought they were pretty fair." (Foundation year 1 female 3)***
>
> ***"They were an intimidating set of exams ….. They were reasonably tough they weren't sort of unnecessarily harsh in any way." (Foundation year 1 male 3)***
>
> ***"Finals were pretty tough, because they were finals" (Foundation year 1 male 2)***

There was also discussion about timetabling of the exams. Third years felt workload for June exams was much higher than that for January. They wanted workload to be distributed more equally. Final year students reported finishing a placement on a Friday and taking finals on Monday caused considerable stress. Final year female 1 wanted one week of revision time. Third year male 2 felt knowledge assessments ending and OSCEs starting was too close.

> ***"So between the theory exams and the practical exam, it was very difficult to try to come into the hospital to practice in the clinical skills room, because everyone was trying to go." (Third year male 2)***

3.5 Discussion

Students felt that they were not able to use feedback because they were not receiving enough personalised specific feedback. They wanted more personalised feedback in addition to learning objective based feedback

after knowledge exams. They wanted feedback after OSCEs. Most participants wanted feedback for everybody, rather than just the failed students, as soon as possible after the exams. However, some students felt that did well enough by passing the exam and they had no interest for further feedback.

While wanting more detailed feedback appeared universal, students had different opinions about what content would be most useful. Some wanted to review their mark sheets with an examiner. Medical schools do not give such feedback, even during remediation for a small number of failed students (Hauer, Teherani et al. 2008, Hauer, Teherani et al. 2009). While this type of feedback will provide closure, there is no evidence to support its educational value. Some other students wanted feedback from the examiners. This is more feasible. These findings resonated with the evidence found in the literature review regarding examiner feedback to students (Harrison, Molyneux et al. 2015). Immediate verbal examiner feedback in high stakes summative OSCEs was not feasible as students were wary of any factor that would interfere with their performance at summative OSCEs.

However, even if personalised feedback could be generated for every student, there remained a question of how receptive the learners would be of the feedback they receive after summative assessments. In this study, some felt no need to use the feedback they received because they have already passed the exam, similar to findings by Sinclair et al (Sinclair and Cleland 2007). Some felt the feedback might be useful for others although they did not personally use it, similar to findings by Winstone et al in their aptly named paper, “It’d be useful but I wouldn’t use it” (Winstone, Nash et al. 2016). This finding is congruent with the evidence of feedback paradox in which students complain about the quality of feedback they

receive, yet also make limited use of it (Withey 2013). This is also congruent with Harrison et al's findings in which the learners were reluctant to engage with feedback that is not directly relevant to passing the next summative exam (Harrison, Konings et al. 2015).

Therefore, the implication is not only to deliver the feedback but also to evaluate how much of it the learners actually read. However, Harrison cautioned that high feedback recipience rates may represent only the learners' curiosity rather than actual learning (Harrison 2019). Therefore, the next step is to explore the learners' attitudes towards feedback. Measuring both learner recipience and attitude towards feedback together could form a more complete picture of feedback engagement.

The one failed student who participated in this study described her frustrating feedback session with an educator. The issue was that the feedback giver said the examiner may not have seen her perform a particular task, to justify why she was not awarded the mark she felt she deserved. This response served no educational or emotional purpose. On the other hand, the feedback givers at the time did not have access to any more information than what was on the mark sheet. This indicated that a better system to record and deliver feedback was required. A possible solution was to cut out the middle man, by conveying examiner feedback directly to students. Training the examiners and providing them with a structure to give feedback appeared to be a suitable approach to address the learners' perceived needs.

Unsatisfactory feedback discouraged students from seeking feedback and spread doubts about examiners. Even though no other participant had personally experienced any problems, some students seemed to have little faith in the fairness of the exams in terms of examiners, locations and patients. While the participants mentioned wanting more realistic OSCEs,

their fear of unfairness appeared to be a larger concern. Students were wary of being assigned with patients who would put them off, such as the severely disabled, very young, uncooperative or chatty.

Making the assessments more standardised, such as using simulated patients, could reinforce their validity and fairness. However, more standardisation could lead to a larger disconnection between the summative assessment and the authentic clinical practice, which is a barrier to learner engagement with feedback produced from such assessments (Harrison, Konings et al. 2015). It is challenging to develop reliable measurements of student performance which not only have predictive value for subsequent clinical competence but also have a formative, educational role (Wass, Van der Vleuten et al. 2001).

Students were often discouraged to putting in effort to receive feedback when they did not get what they expected from the feedback. The educators should appreciate that students have to put in both mental and physical effort, for example by going to Mansfield fifteen miles away, to receive feedback.

3.5.1 Implications

To address the learners' perceived feedback needs, individualised feedback should be given after summative OSCEs to everyone as soon as possible. To improve the perception of fairness, examiners should be trained to mark and give feedback consistently. The quantity and quality of the delivered feedback should be evaluated to address concerns about fairness and lack of specific information. As students elaborated not using feedback for various reasons despite thinking it may be useful for them, further research should investigate whether they engage with feedback they receive and their feedback orientation. The construct of feedback orientation covers the learners' awareness of the utility of feedback and

their sense of responsibility to use the feedback they receive, both of which appeared to be weak in some students in this study.

Furthermore, students should be better informed about measures for fairness and standard setting in the exams to ease undue stress. This can potentially improve learners' trust, perception of fairness and optimism.

3.5.2 Limitations

Having a detailed interview question guide, using peer interviewing and having only one researcher for analysis led to several limitations.

The interview question guide contained multiple items under each topic. The interviewers were told to only use it to keep the discussion on topic and allow the participants to express their views freely. However, the interviewers were not experienced and the interviewer training was brief. Consequently, all the items were asked in all the interviews and most answers were brief and specific. This led to short interviews which appeared to take a positivist approach, similar to a verbally conducted survey. Short responses to multiple questions made data collection appear incongruent with phenomenological methodology's aim to collect a rich contextualized description of lived experience.

Peer interviewing and convenience sampling raised the possibility of participant bias, which is only those who had a positive attitude towards the topic or the interviewer participated in the interviews and selection bias, as it reduced the chance of obtaining a diverse selection of interviewees. The peer interviewers (T.H. and R.C.) approached their friends to volunteer for the interviews and consequently, the sample was largely homogeneous. All participants belonged to the same university and had similar academic performance levels. Only one participant has failed a summative exam while the rest passed. As with any other face to face

interviews, there was also a likelihood of social desirability bias, in which the participants tend to say things that they think would be liked and accepted.

Transcendental phenomenology approach assumes that the analyst is capable of bracketing. However, having only one analyst, M.M.S., increased the risk of confirmation bias, in which the researcher tends to interpret the results to confirm to their pre-existing beliefs.

A reflective diary should have been kept during both data collection and analysis. The reflective diary was kept by the researcher, M.M.S. only during analysis. The interviewers, T.H. and R.C. did not keep one during data collection. Consequently, themes which emerged in earlier interviews were not followed up in subsequent interviews.

Member checking was done between M.M.S. and the two interviewer-participants only. Including more participants would have further reduced the risk of confirmation bias.

3.6 Summary

This chapter presented the interviews on the learners' perceptions to the feedback and assessment program they were subjected to. This study found several factors influencing the students' perception of feedback. Students felt the assessment programme at the time did not generate the kind of feedback they needed. The learners wanted detailed individualised feedback to not be limited only for those who failed but to be offered to everyone. Therefore, the next step was to design and deliver a feasible and acceptable feedback programme for all students after summative assessments. However, questions remained on whether the students will use the summative feedback once they get it as students elaborated various reasons for not using the feedback they received.

Chapter 4. Digital marking and feedback

4.1 Introduction

In Chapter 1, a rationale was presented for providing personalised feedback to all students after summative OSCEs, and the need for evaluating the intervention in terms of the quality and quantity of feedback produced. A need for generating more evidence on how learners from different academic performance levels engage with feedback they receive was also presented.

In Chapter 2, evidence on the best approaches to timing, mode of communication, feedback giver and content regarding feedback to students after summative OSCEs was considered. The evidence supported the feasibility, acceptability and efficiency of digital tablet computers to record written examiner comments during the OSCEs and delivering the feedback online after the exams.

In Chapter 3, interviews were conducted with medical students at the University of Nottingham regarding their perceived feedback needs. The interviews showed that they appreciated receiving learning objective based feedback after summative knowledge exams. The students were not receiving any feedback after the OSCEs and they wanted short feedback comments directly from the examiners. They wanted this kind of feedback not to be limited only to those who failed but to be offered to everyone.

This study's aim was to pilot, implement and evaluate summative examiner feedback to students using digital marking technology. The objectives were presented under objective 3 under subheading "1.8 Research question, aim and objectives" in Chapter 1. The researcher designed a feedback intervention according to the best evidence and findings from the student interviews in Chapter 2 and Chapter 3 respectively.

4.1.1 Domain marking

There was a change in the OSCE marking scheme from detailed checklists to domain marking in 2016. This change in marking format was not part of this study's intervention but the context in which this study took place. Until 2015, clinical skills assessments at UON used the traditional detailed OSCE checklists. The school decided to move away from detailed checklists to domain checklists in 2016. The school's Assessment Unit generated eight domains; clinical communication, professionalism, information gathering, information giving, physical examination, procedural skills, clinical reasoning, and management. Each domain was marked with a five point scale. This reduced the number of marked items from dozens to about six per station. There was no validation work for the new marking structure apart from internal discussions between the dean, the module leaders and the Assessment Unit. The GMC had been aware of the change since its implementation and expressed no concerns.

Domain marking is a valid tool for clinical skill assessment. The use of domain marking and feedback in formative workplace based assessments (WBAs) showed high acceptability and increasing quality of feedback and student satisfaction over time (Coulby, Hennessey et al. 2011, Haffling, Beckman et al. 2011). This marking approach relies on expert judgement, which is seen to be potentially subjective and unfair. However, there has been recent examinations about the role of subjectivity in assessment. Epstein called for caution regarding the assumption that conclusions drawn from quantitative data are always more valuable than qualitative expert judgment (Epstein 2007). Eva and Hodges pointed out that while the literature is replete with critiques of the dangers of subjectivity in assessment, there is evidence showing that many fallible judgments, summed together, create value (Eva and Hodges 2012). The value of subjective judgement was shown to increase with the diversity of

perspectives, and the number and independence of judgments (Suriowiecki 2004).

Reliability and validity are the principal markers of the acceptability of assessment instruments (Tavakol and Dennick 2012). To boost inter-rater reliability, traditional OSCEs use detailed checklists which typically required the examiner to check dozens of items. Detailed checklists measure the various aspects of the candidates' clinical competency in terms of specific observable actions (Black and Harden 1986). This is meant to minimise the examiner's bias. The role of the examiner is to mark according to the pre-written checklist, while consciously minimising their own bias. Ideally, this should result in perfect inter-rater reliability no matter who the examiner is. In this approach, the experience of the examiner and their construct of competence are less important than their ability to observe and complete the checklist. Consequently, standardisation of the questions, and training examiners and patients became vital in delivering an OSCE. However, these tasks are required for both domain marking and traditional detailed checklist marking OSCEs.

The detailed checklist approach is based on the assumption that competence is a fixed observable construct (Hodges 2013). However the construct of competence can vary according to context (Hodges 2013). Competency assessment in medical education shifted its focus from measuring small fixed dimensions of a trainee to a holistic notion of their character (Whitehead, Hodges et al. 2013). Whitehead argued that while there are many advantages to identifying individual actions of competence, to place sole reliance on measurements of small dimensions is to risk losing the art of judging character. Similarly, Schuwirth used an analogy to assert the value of expert judgement in assessment: when diagnosing a patient, an expert clinician's spot diagnosis is more valuable

than a novice's (Schuwirth, van der Vleuten et al. 2017). Furthermore, the clinicians are trusted to consciously minimise their biases and make a diagnosis to the best of their abilities. Domain marking applies the same logic to assessment.

Domain feedback informs the students of their patterns of behaviour. Therefore, this feedback was not only a tool for closure about what happened at the exam but also a tool for continued learning. Furthermore, by reducing the number of checklist items, the examiners are likely to have more time for feedback. Therefore, the researcher viewed domain marking as a valid assessment approach which created more opportunities for feedback.

4.2 Methods

4.2.1 Developing digital marking

The UON School of Medicine purchased sixty Samsung Galaxy Tab™ digital tablet computers and a one year subscription for MaxExam™ software. The school intended to use these assets in all its clinical skills assessments.

The researcher initially considered modifying ROGO, the assessment management software developed by UON. It was used for knowledge assessments but it could be modified for use in OSCEs. However, the university policy did not allow modifications to ROGO. It was not possible to develop a new software from scratch using the available expertise, time and resources. Therefore, the researcher used MaxExam™, a commercially available software for assessments. The main reason for choosing MaxExam™ was that the software company expressed their willingness to modify their software to meet UON requirements for free

and it was expected that extensive customisation will be needed for any chosen software to meet local needs. Furthermore, MaxExam™ was already in use at the UON Veterinary School and the Royal College of Dentistry exams.

MaxExam™ software was written only for digital tablet computers which used an android operating system. Therefore, Samsung android tablets were purchased. The system had three components; the client on the tablets, also known as an "app", a website and a secure server. Firstly, mark sheets were created using the website and saved on a secure server. Examiners could then access the mark sheets using the tablets on the app during the exam. The marks and feedback for each student were saved on the tablet's removable memory card and the online server simultaneously and continuously throughout the assessment. The marks on the server could then be accessed via the website for standard setting. Initially, both the pass-fail results and feedback were delivered to students via email. After a software update in January 2018, it became possible to publish the feedback via the website. Any access to the website and the app required usernames, passwords and the website administrator's approval. A Wi-Fi internet connection is required at the beginning of an exam to enable examiners to log in and download mark sheets and student lists. During the exam, the system is able to function offline if necessary. At the end of the exam, the internet connection is again required to upload the marks to the server. The pass marks were set after the exam by the assessment board.

The researcher approached the staff from other departments and schools within the university to learn from their experience in using digital tablets as similar systems were in use for teaching and formative assessments. The researcher gathered practical information regarding issues with

charging multiple tablets simultaneously, safely, storage, transportation, internet connection, updating software, and digital security for the data and physical security for the tablets.

4.2.1.1 Designing the mark sheets

Every mark sheet included a marking checklist followed by a feedback section. The OSCE scenarios and marking checklists were developed by the Assessment Unit without any input from the researcher. Once the scenarios and the marking checklists were written, the Assessment Unit piloted the stations using Foundation Year One (FY1) doctors and simulated patients. FY1 doctors are recently graduated junior doctors working in the National Health Service. This was to assess and modify the stations' difficulty. Once the scenarios and mark sheets were finalised, the Assessment Unit gave the researcher access to develop the feedback sections for each station. The researcher alone was responsible for developing the feedback section which did not contribute to the pass-fail decisions or the marks.

4.2.1.2 Sample mark sheet

Figure 2 and Figure 3 show a mark sheet used in a station in the pilot, December 2016 Healthcare of the Elderly (HCE) Resit.

Figure 2 Sample OSCE sheet marking section

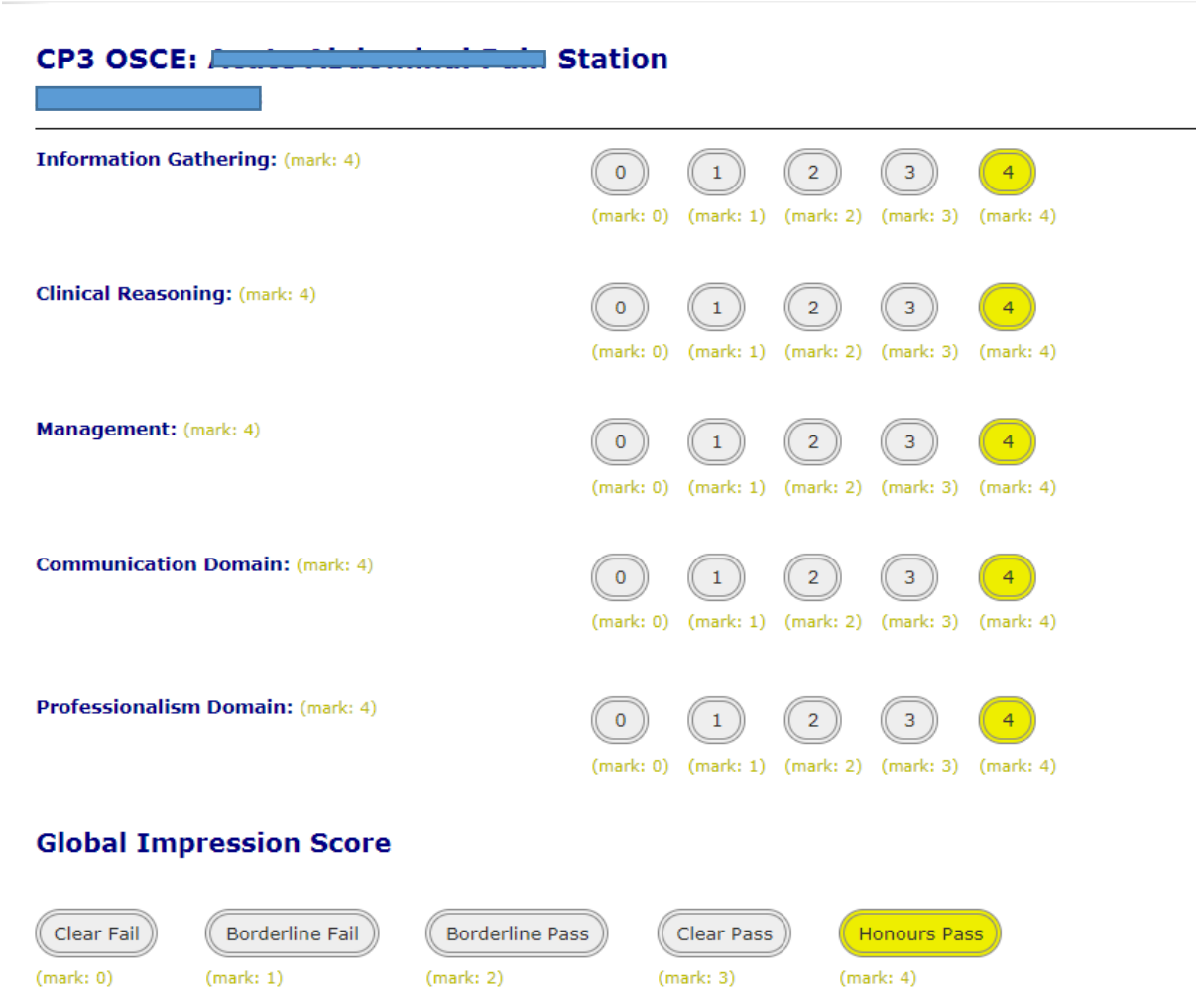
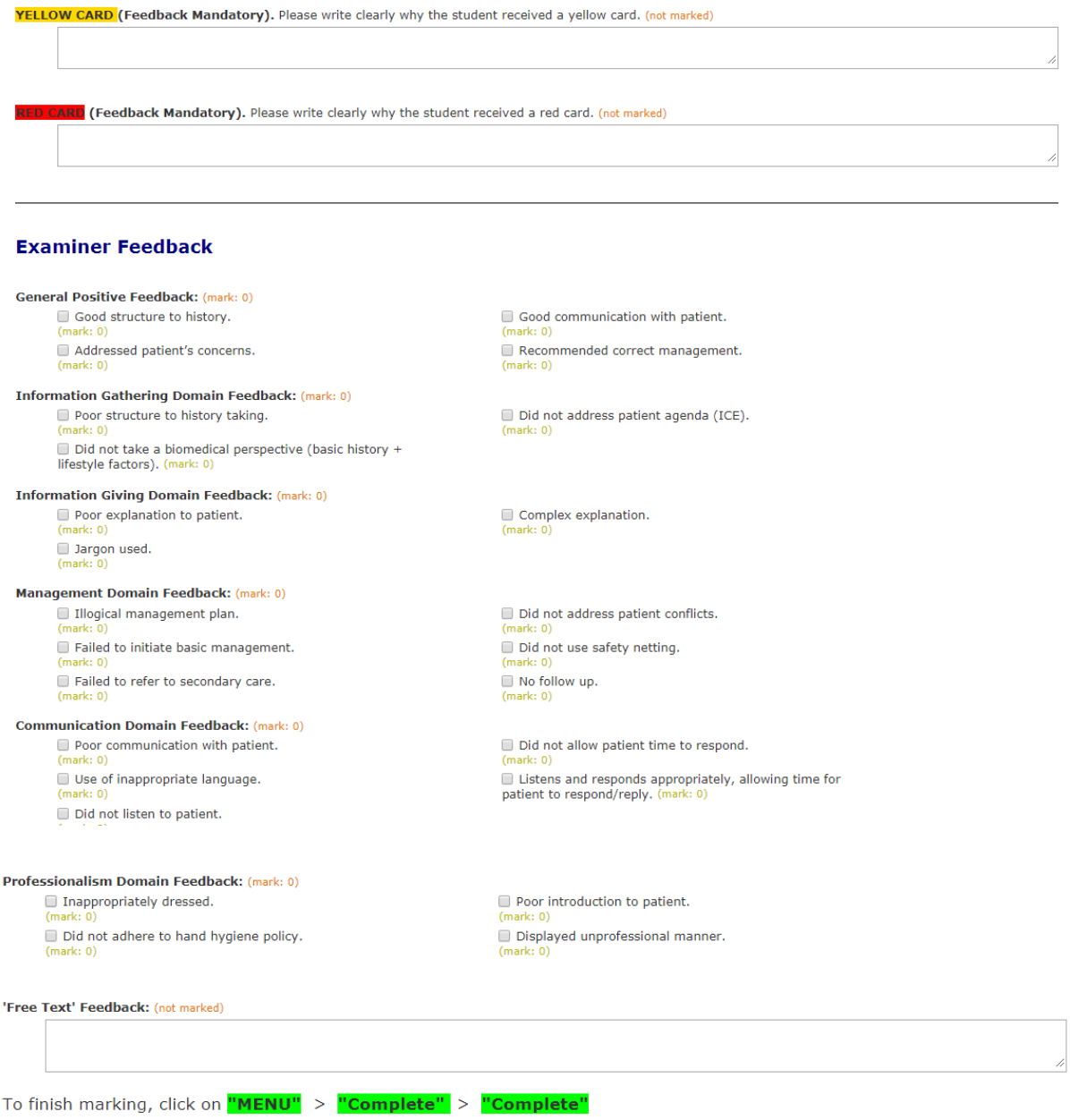


Figure 3 Sample OSCE sheet feedback section



Designing the on-screen format of the mark sheet was an iterative process. The Assessment Unit, the school's staff and the researcher had multiple discussions regarding the visual design. The discussions led to designing an intuitive mark sheet for the tablet screen, rather than trying to make it look like the paper version. Spacing and positioning of the items were refined during the course of development to aid the examiner navigate easily on the tablet screen. To minimise the chance of examiner error, the on-screen letters and buttons were made as big, clear and far

apart from each other as possible. Until the time of the writing in 2018, the mark sheet's appearance had not changed significantly from this example although multiple updates had been installed to improve functionality.

The marking section was written by the assessment unit. The feedback section for the pilot was written by the researcher. Examiners could access all the sections by scrolling up and down on their tablet screens. Each mark sheet contained up to six domain items on a five point scale, with zero being the lowest and four the highest. The domain checklist is followed by Global Impression Score which ranged from "Clear Fail" to "Honours Pass". Five domains were assessed in this example station. The domain items contributed to the student's mark while the global impression score was used for standard setting. This marking format remained unchanged until the time of writing. All the domain checklist items and the global impression score were mandatory: the software did not allow examiners to submit their marks unless all those items were completed. This prevented missing data.

The feedback section included red and yellow card feedback, pre-written feedback items and free text feedback. The Assessment Unit reviewed and agreed to it. This format also remained unchanged since the pilot. If the candidate behaved unprofessionally or dangerously, a yellow card or a red card was to be issued, depending on the potential severity of the action. The examiners were instructed to give written feedback to explain why if a card was issued. Later in 2017, a software update ensured that the examiners had to give written feedback once they issued cards. Otherwise, all feedback was optional. All other items in the feedback section were not mandatory. Pre-written feedback items were aimed to cover common mistakes that students made in each station. Examiners

could tick a box next to the relevant pre-written item. This was named "tickbox feedback". The researcher alone generated these items for the pilot. New items were added according to suggestions by the Assessment Unit and the examiners in the subsequent exams. The text box at the bottom of the mark sheet was for the examiners to write any "free text feedback" at their discretion.

4.2.1.3 Pilot

Examiner training sessions were held a few weeks before the pilot. It was a half-day session to train the examiners to use the tablets and familiarise them with domain marking and feedback formats. The examiners were instructed to focus on marking. It was also made clear that they did not need to give reasons for their decisions in the feedback and that all feedback was optional, to be given only if they were willing and able to do so. Similar sessions were held for subsequent exams.

Digital marking and feedback was piloted in a small OSCE. It was a resit OSCE in December 2016 for final year (CP3) students who failed the finals. It was an actual high stakes summative assessment. The students would have had to either repeat the year or be terminated from the course if they failed this exam.

Students were informed several weeks before the exam that digital marking would be piloted at their exam and that this change would have no effect on their marks in any way. Preparations were made to switch to paper backups with minimal interruption in case there were issues with digital marking.

Immediately after the exam, students and examiners were explicitly asked to raise concerns or report incidents. Examiners and students did not

experience any problems with digital marking and feedback. The results and feedback were delivered to students two weeks after the exam.

4.2.1.4 Large-scale implementation

In February 2016, digital marking was used in a large scale examination. Before the pilot, up to five different sites hosted OSCEs simultaneously. In 2016, the medical education facility at Derby became the only site for all the OSCEs. Using a single large site allowed a relatively small number of administrators and examiners to accommodate a large number of students. Wi-Fi internet, loudspeaker systems, equipment for storage, transport and charging cabinets for the digital tablets were installed. Moving all assessments to a single site was not part of this study's intervention but it was a change in the school policy that coincided with the intervention.

An arrangement of "circuits" and "rounds" was used to assess a large number of students in a short time. See Table 1. A circuit is composed of eight examiners in eight stations. A round is the time it takes for students to rotate though all the eight stations.

Table 1 Example circuit arrangement

Station	Yellow Circuit	Blue Circuit	Red Circuit	Purple Circuit
1	Examiner1	Examiner9	Examiner17	Examiner25
2	Examiner2	Examiner10	Examiner18	Examiner 26
3	Examiner3	Examiner11	Examiner19	Examiner27
4	Examiner4	Examiner12	Examiner20	Examiner28
5	Examiner5	Examiner13	Examiner21	Examiner29
6	Examiner6	Examiner14	Examiner22	Examiner30
7	Examiner7	Examiner15	Examiner23	Examiner31
8	Examiner8	Examiner16	Examiner24	Examiner32

The assessment facility had the capacity to accommodate 32 assessments at the same time. Assessing each station lasted ten minutes and there were two minute breaks between stations. For eight students to rotate through eights stations to complete a round took approximately one hour and forty minutes. Up to four rounds could be completed per day to assess 128 students. See Table 1.

The researcher, M.M.S., developed an intuitive coding system to categorise and order the candidate list to minimise the risk of mix-ups. The code is a set of three alphanumeric characters for the day of the OSCE, time of the OSCE and the colour of the circuit. For example, if an examiner typed “11Y” into the app on the tablet, the app would show the examiner only eight candidates who were scheduled to take their OSCE on the first round of the first day of the OSCE in the yellow circuit. The coding system was aimed to help examiners find their relevant candidates quickly and also to “filter out” the irrelevant ones. The code was explained to the

examiners during training and also during the briefing just before the exam when they received their assigned codes and assigned student lists. The examiners were also instructed to check their candidate's identity before starting the assessment.

After the exam, the administrators ran student debriefing sessions in which they explicitly asked the students to voice their incidents, concerns and complaints. They were told they could approach the administrators for private discussions at the end if they were not comfortable complaining publicly. Separate debriefing sessions were held for the examiners. At the same time, the administrators checked if all the examiners have submitted their assessments onto the server.

As mentioned earlier, the marks were downloaded afterwards for standard setting. Once standard setting was finished, the administrators initially had to "mail merge" the feedback and the pass-fail results using Microsoft Outlook™. It was impossible to know whether the students read the feedback they received or not. However, a software update in January 2018 made it possible to deliver the feedback via the website. Students were then able to log in and view their own feedback. It became possible to see which students read their feedback. The pass-fail results continued to be delivered via email. The staff and the examiners expressed during training sessions and meetings that they were intimidated by unfamiliar technology in high stakes assessments at first. However, confidence and optimism appeared to improve with time.

4.2.1.5 Timeline

December 2016	- Pilot 2016 CP3 Resit OSCE, feedback via email
January 2018	- Software update for student log-in
February 2018	- CP3 OSCE, feedback via website
June 2018	- CP3 Resit OSCE, feedback via website

4.2.2 The criteria for feedback quality assessment

It was assumed that evaluating the feedback quality, quantity and user engagement would generate a higher level of evidence than a user satisfaction survey targeted at the most basic level of evaluations (Kirkpatrick 1996).

Only one validated scale was found in the literature to assess the quality of written examiner feedback in OSCEs. Denison et al developed a five point scale to appraise the written feedback in paper marking and digital marking in OSCEs as they were unable to find a similar scale (Denison, Bate et al. 2016). Denison scale was validated for construct validity by four independent experts and had a high inter-rater reliability (Kappa value = 0.625). However, Denison et al did not define clearly what "objective appraisal of performance" was, although they provided verbatim examples. It was one of the key criteria in their scale. Furthermore, Denison et al's scale did not include a point for assessments with no feedback or those with illegible comments. Therefore, a more simplified seven point scale was developed.

Feedback quality appraisal criteria

0 No examiner comment given

1 Illegible, e.g.,

"abd / loin pain not", "I", "ogd", "Patient had a low respiratory rate"

2 Non-specific judgement, e.g.,

"excellent.",
"overall good performance",
"excellent interaction with patient, lovely bedside manner"

3 General description of performance or description of only one action, e.g.,

"Excellent history taking",
"did not examine abdomen",
"neck pulse check during ascultation appeared very lateral in the supraclavicular fossa."

4 Suggestion for improvement, general or specific, without description of action, e.g.,

"adviee on driving needed",
"chunk and check more",
"focus your examination on excluding or confirming differentials show consideration for patient comfort and avoiding further harm during examination"

5 Description of two or more actions without suggestion for improvement, e.g.,

"did not examine abdomen otherwise did well",
"Lets just move on to a bit more about you was good signposting. Well done for getting the patient's consent to discuss what you think the problem is. Very clear guidance re driving",
"Addressed lifestyle risk factors. Initially appeared uncertain re referral. Then decided to refer but not clear how urgently would refer, had to be

prompted by patient... just said 'quickly'. Did give aspirin and advised not to drive",

6 Description of performance AND suggestion for improvement, e.g.,

"Asked patient if they had an aura, most patients wont know what this is. Ambulance not needed here, refer to TIA clinic",
"Very thorough history. You checked for red flags including weight loss and night sweats. Good use of summarising. Could have approached the possibility of the fit being related to previous cancer with more of a breaking bad news structure eg warning shot, but overall a very empathic consultation. Correct management including advice about driving but missed out safety netting"

These example comments were copied verbatim from data. The researcher developed this scale independently. There was no attempt for establishing its construct validity or determining inter-rater reliability as it was a simple criteria.

4.2.3 Data collection

The feedback data were collected from the February 2018 CP3 OSCEs. It included tickbox and free text feedback. This was the first exam for which students could log in to the MaxExam™ website and view their own feedback online. The exam took place over three days. The feedback generated by the examiners in the whole exam was compiled on an Excel™ spreadsheet. It was arranged according to student ID, student's performance at the exam (Honours, Pass, Fail), examiner ID and stations. Excel™ formulas were used to count the number of ticked boxes and words.

Failed students from February 2018 CP3 OSCEs took the resit in June 2018. The website log-in data were collected from both the main exam

and the resits to see how many students viewed their feedback. No other data were collected from the resits.

The researcher collected emails from students regarding the digital marking and feedback intervention from December 2016 to July 2017. During this period, the exam result emails to students included the researcher's email address and a request for the students' feedback regarding the intervention along with their results.

4.2.4 Analysis

Firstly, the researcher read all the free text comments and graded them according to the appraisal scale. Percentages of the different grades were calculated and reported using histograms.

Secondly, student data were categorised into three groups in terms of their performance at the February 2018 CP3 exam; Honours, Pass and Fail. The feedback quantity distribution was analysed between the eight stations and the three performance groups. One–way ANOVA tests were performed to compare the number of tickbox ticks and free text word counts between the performance groups.

Thirdly, feedback view rates in terms of website log-ins were compared between these performance groups in the two exams (the main in February 2018 and the resit in June 2018). SPSS™ 20 and Excel™ to were used to generate the statistical reports.

Furthermore, the researcher read the emails from students to consider the acceptability of the intervention.

4.3 Results

Seventy examiners assessed 287 students in eight stations in February 2018 CP3 OSCEs, resulting in a total of 2296 individual assessments. Thirty eight students failed this exam and took the resit in June 2018.

4.3.1 Incidents

At the time of writing in 2019, approximately three hundred examiners have participated in 26 large scale OSCEs. The exams were delivered without any serious disruptions or incidents. Since the pilot, there had been five incidents during the exams in which the examiners had to switch to paper marking because they had difficulty with digital marking. There was one incident after an exam in 2017 in which the administrators mixed up the yellow card feedback data and sent wrong yellow card information to students. This mistake was detected immediately as the students emailed about their confusion with the yellow cards feedback. The students were asked to ignore the emails and correct data were sent out on the same day. This incident involved only the yellow/red card feedback but not any other mark or feedback data. One more incident in 2018 occurred when the administrators allowed the students to log in and see their feedback before their results were published. Students were able to see their feedback while not knowing whether they passed or failed. They immediately contacted the school and asked to publish their pass-fail results. The results were delivered two hours after the feedback. Neither the student representatives nor the faculty raised concerns regarding any of these incidents at learning community forums. There had not been any complaints or appeals by individual students related to digital marking or feedback at the time of writing.

4.3.2 Efficiency

Paper marking in 2015 required at least two experienced administrators to print, transport, distribute, collect and store the physical mark sheets securely. These tasks for each large-scale exam took at least two working days. Digital marking required only one administrator to complete equivalent logistic tasks, such as the preparation or editing of checklists, setting up the examination online, charging tablet batteries and processing the data. Sixty tablets were enough for all the exams instead of thousands of paper sheets for each exam. The raw marks were available for standard setting immediately after the exams. Initially after standard setting, it took one administrator about five hours to mail-merge the feedback to students using Microsoft Outlook™. After the software update that enabled students to log in and view their own feedback on the website was released, it took the administrator less than an hour to deliver feedback on the website and mail-merge the results to the students.

4.3.3 Quality

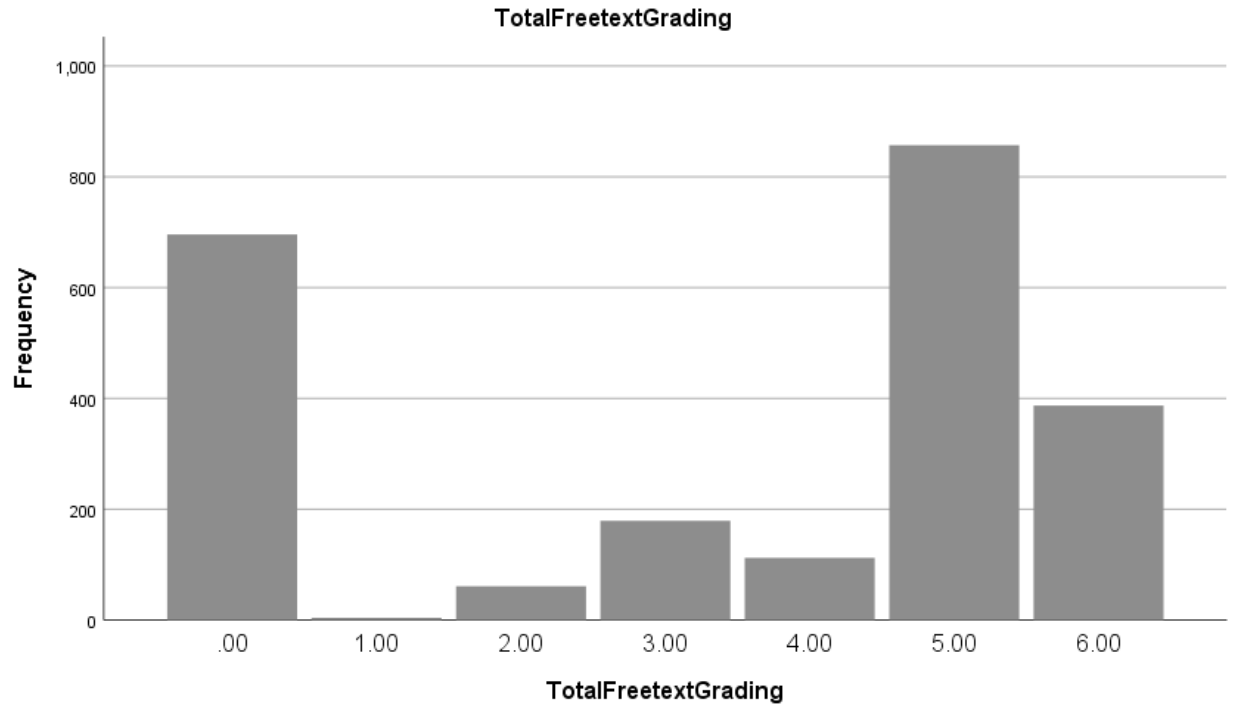
There were only four illegible comments in February 2018 CP3 OSCEs. No derogatory or abusive comments were found. Approximately 30% of the assessments did not produce any free text feedback comments (N=696, 30.31%). Of the assessments in which free text feedback was given (N=1600), 240 (14.89%) generated only a general judgement of the student's overall performance or addressed only one action by the student. About a quarter of the comments included both a description of the student's actions and a suggestion for improvement (N = 387, 24.18%). See Table 2 and Figure 4.

In all stations except station one, Grade 5 feedback in which the examiners highlighted two or more actions of the student without any explicit advice for improvement was the most common type of feedback.

Table 2 Total feedback quality distribution table

GRADING	FREQUENCY	PERCENT OF THE TOTAL (N= 2296)	PERCENT OF WRITTEN COMMENTS (N=1600)
0	696	30.31	NA
1	4	0.17	0.25
2	61	2.66	3.70
3	179	7.80	11.19
4	112	4.94	7.00
5	857	37.33	53.56
6	387	16.86	24.18
TOTAL	2296	100	100

Figure 4 Total feedback quality distribution histogram (N=2296)



4.3.4 Quantity

4.3.4.1 Overall

There were 2296 assessments for 287 students across eight stations. Either free text or tickbox feedback was given in 2266 assessments (98.69%).

In terms of student numbers, all students received tickbox feedback in at least five stations. Nearly all students (N=284, 98.96%) received tickbox feedback in at least six stations. See Table 3. Station two and station four produced feedback for most students (N=283, 98.61%) while station eight did so for the least (N=220, 76.66%). See Table 4.

All students received free text feedback in at least two stations. Nearly all students (N=271, 94.43%) received free text feedback in at least four stations. Station three produced free text feedback for most students (N=257, 89.55%) while station two did so for the least (N=139, 48.43%).

Table 3 Feedback by station count

TOTAL NUMBER OF STATIONS IN WHICH STUDENTS RECEIVED FEEDBACK	STUDENT COUNT FOR FREE TEXT	STUDENT COUNT FOR TICKBOX
	N (%)	N (%)
0	0 (0%)	0 (0%)
1	0 (0%)	0 (0%)
2	5 (1.74%)	0 (0%)
3	11 (3.83%)	0 (0%)
4	43 (14.98%)	0 (0%)
5	77 (26.83%)	3 (1.05%)
6	81 (28.22%)	39 (13.59%)
7	47 (16.38%)	124 (43.21%)
8	23 (8.01%)	121 (42.16%)

Table 4 Quantity distribution by station in terms of whether or not feedback was generated (N = 287 for all stations)

	ST1 N = 287	ST2 N = 287	ST3 N = 287	ST4 N = 287	ST5 N = 287	ST6 N = 287	ST7 N = 287	ST8 N = 287	TOTAL (N = 2296)
FREE TEXT	206 (71.78%)	139 (48.43%)	257 (89.55%)	213 (74.22%)	184 (64.11%)	210 (73.17%)	220 (76.66%)	171 (59.58%)	1600 (69.69%)
NO FREE TEXT	81 (28.22%)	148 (51.57%)	30 (10.45%)	74 (25.78%)	103 (35.89%)	77 (26.83%)	67 (23.34%)	116 (40.42%)	696 (30.31%)
TICKBOX	234 (81.53%)	283 (98.61%)	263 (91.64%)	283 (98.61%)	267 (93.03%)	255 (88.85%)	280 (97.56%)	220 (76.66%)	2085 (90.81%)
NO TICKBOX	53 (18.47%)	4 (1.39%)	24 (8.36%)	4 (1.39%)	20 (6.97%)	32 (11.15%)	7 (2.44%)	67 (23.34%)	211 (9.19%)
NO FEEDBACK	4 (1.39%)	0 (0%)	0 (0%)	1 (0.35%)	2 (0.70%)	4 (1.39%)	1 (0.35%)	18 (6.27%)	30 (1.31%)

4.3.4.2 Tickbox count

Average tickbox count was 32.45 for each student. The tickbox count distribution followed closely to the normal distribution curve with a standard deviation of 5.37 and skewness of 0.12. See Figure 5.

There was no statistically significant difference between the means of the three performance groups according to the one way ANOVA test. See Figure 6.

Station two generated the largest number of ticks (mean = 5.01) while station one generated the smallest (mean = 2.18). See Table 5.

Figure 5 Overall tickbox count distribution

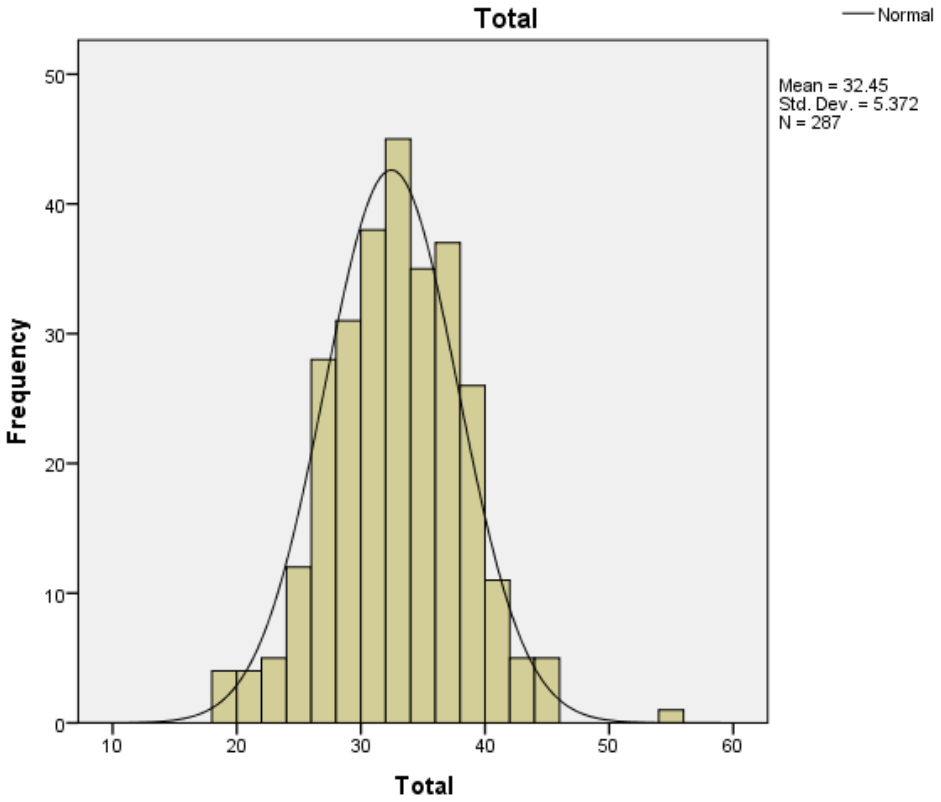


Figure 6 Tickbox count distribution by performance boxplot

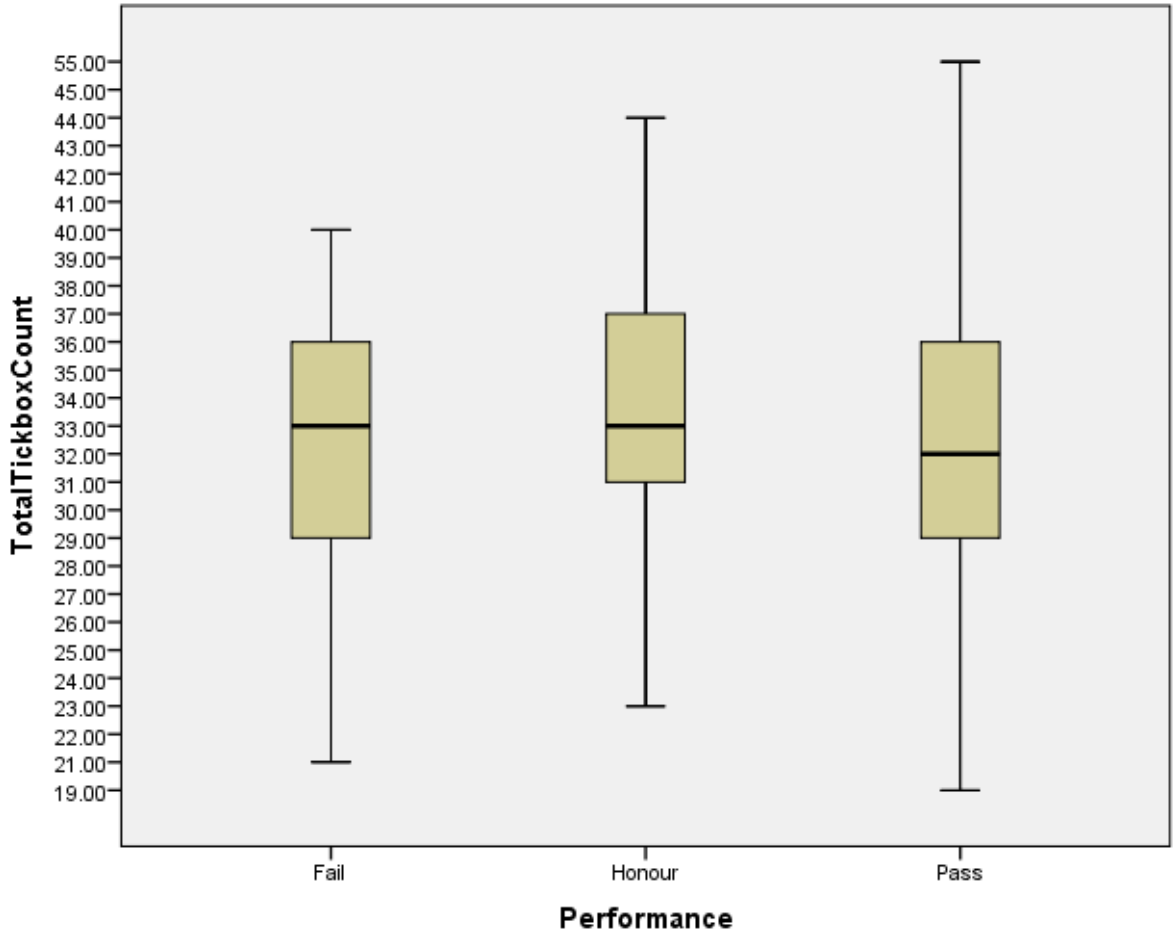


Table 5 Tickbox count distribution by performance and station

	N	MINIMUM	MAXIMUM	MEAN	MEDIAN	STD. DEVIATION
ST1	287	0.00	7.00	2.18	2.00	1.52
ST2	287	0.00	9.00	5.01	5.00	1.78
ST3	287	0.00	10.00	4.16	4.00	2.19
ST4	287	0.00	10.00	4.77	5.00	1.57
ST5	287	0.00	8.00	4.55	5.00	1.89
ST6	287	0.00	12.00	4.85	5.00	2.69
ST7	287	0.00	8.00	4.24	5.00	1.41
ST8	287	0.00	19.00	2.69	2.00	2.27
HONOURS	54	23.00	44.00	33.63	33.00	4.56
PASSED	195	19.00	55.00	32.11	32.00	5.66
FAILED	38	21.00	40.00	32.53	33.00	4.78
TICKBOX COUNT TOTAL	287	19.00	55.00	32.45	32.00	5.37

4.3.4.3 Free text word count

On average, examiners gave 125.73 words as free text feedback to each student with a standard deviation of 50.88 and skewness of 0.67. See Figure 7.

Honours students (N=54) received the lowest amount of free text feedback (mean=111.85, median=106) while the failed students (N=38) received the largest amount (mean=154.05, median=151). One way ANOVA test shows that failed students received significantly more feedback than passed students (P = 0.00, mean difference = 32.39, SE = 8.77), and honours students (P = 0.00, mean difference = 44.2, SE = 10.47) but there was no significant difference between the means of the Pass and Honours groups. See Figure 8.

Stations three and seven generated the largest word counts with each having means over 22. Station five produced the smallest word count with a mean of 9.76. See Table 6.

Figure 7 Overall word count distribution

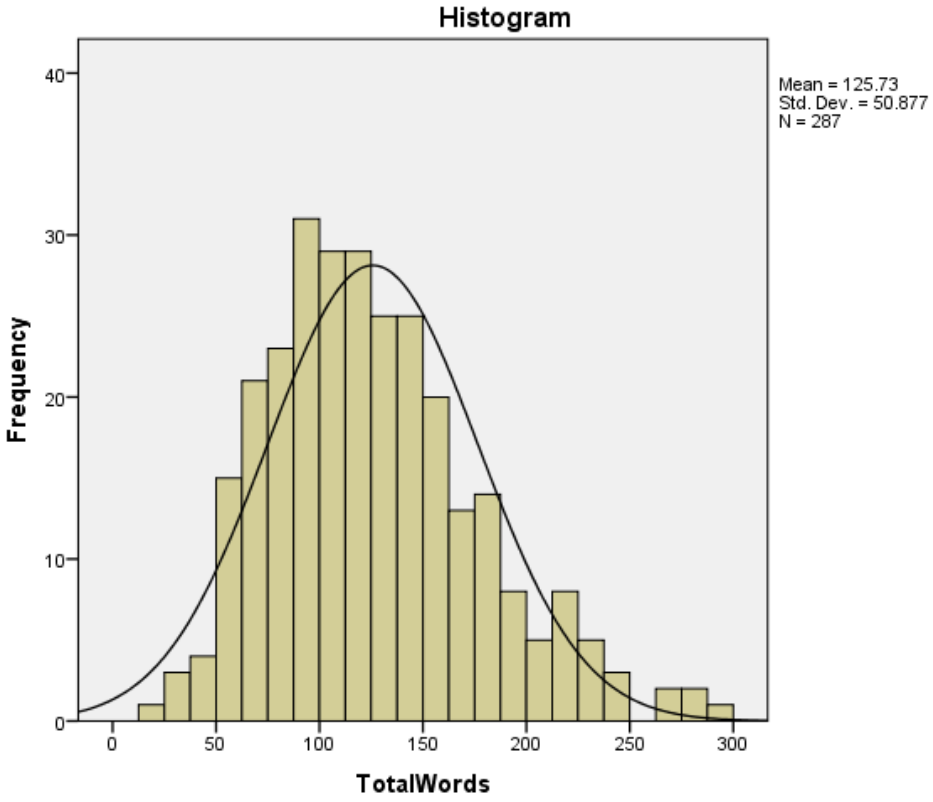


Figure 8 Word count distribution by performance boxplot

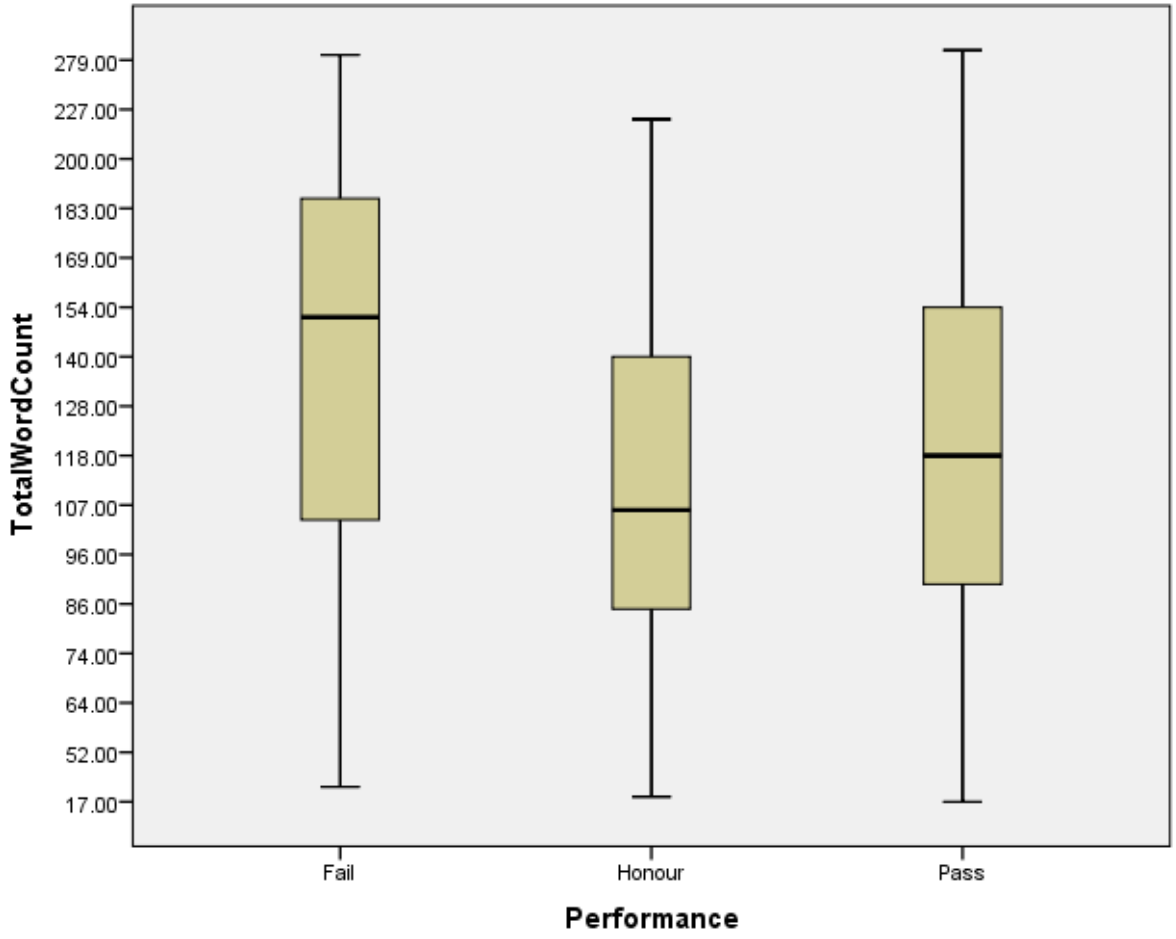


Table 6 Word count distribution by performance and station

	N	MINIMUM	MAXIMUM	MEAN	MEDIAN	STD. DEVIATION
ST1	287	0	113	18.98	13	21.06
ST2	287	0	91	10.93	0	16.32
ST3	287	0	149	22.47	16	22.41
ST4	287	0	95	15.13	14	14.33
ST5	287	0	85	9.76	6	14.54
ST6	287	0	76	14.79	11	14.22
ST7	287	0	91	22.13	12	24.82
ST8	287	0	51	11.54	10	12.18
HONOURS	54	28	219	111.85	106	40.87
PASSED	195	17	288	123.66	118	48.61
FAILED	38	36	284	156.05	151	63.09
FREETEXT WORD COUNT TOTAL	287	17	288	125.73	119	50.88

4.3.5 Feedback engagement

A large majority (N=263, 91.64%) of students viewed their feedback after the main exam. Of those, all the failed students viewed their feedback. Honours students' view rate (92.59%) was higher than that of the passed students (89.74%). The log-in rate was lower after the resit at 78.95%. The majority of the passed students at resit (N=26, 83.87%) viewed their feedback while only three out of seven failed students did so. See Table 7.

Table 7 Website log-in rates by number of students after main and resit OSCEs

	VIEWED FEEDBACK N (%)	DID NOT VIEW FEEDBACK N (%)
HONOURS (N = 54)	50 (92.59%)	4 (7.41%)
PASS (N = 195)	175 (89.74%)	20 (10.26%)
FAIL (N = 38)	38 (100%)	0 (0%)
MAIN TOTAL (N = 287)	263 (91.64%)	24 (8.36%)
RESIT PASS (N = 31)	26 (83.87%)	5 (16.13%)
RESIT FAIL (N = 7)	4 (57.14%)	3 (42.86%)
RESIT TOTAL (N= 38)	30(78.95%)	8 (21.05%)

4.3.6 Feedback emails from students

Six emails from students were received. Two of them were presented in the appendix at the end of the thesis as verbatim examples. See Email 1 and Email 2 under "Sample feedback emails from students" in Appendix.

While they were happy about getting more feedback than before, some students felt they should have received more because they knew some of their classmates received more feedback than they did. This feeling is congruent with the students' concern for fairness as identified in the interviews in Chapter 3. In one OSCE with sixteen stations, one student received free text feedback in only three stations while another received free text in twelve. Both students felt they should have been given more feedback.

One expressed concerns about the accuracy of the examiner comments as some of the feedback comments did not reconcile with the student's recollection. The students also wanted more critical feedback for the stations in which they had barely passed.

A failed student wanted his marks checked again, because he had heard from others that *"some (feedback) bits were jumbled up"*. He felt the examiners needed to explain in the feedback why they gave him a fail and that not doing so was unfair.

4.4 Discussion

Digital marking and summative feedback was feasible, reliable, efficient and acceptable. The exams were delivered smoothly. All students received both types of feedback in multiple stations. The quality of feedback was adequate overall. Failed students received more free text feedback than the rest. Nearly all learners engaged with feedback after summative OSCEs. High performing learners engaged with feedback more than lower performers did. Designing a feedback intervention based on the best approaches to feedback in literature and the UON students' opinions was successful in producing adequate quality and quantity of feedback and attracting learner engagement.

4.4.1 Feasibility, acceptability, efficiency

Improvement in efficiency was clear, considering the reduced workload. In addition, as the technology ensured there was no missing data, there was no more need to award full marks for missing data. This consequently improved the reliability of the assessment data. While spelling mistakes and grammatical errors were common in examiner comments, they were all legible enough to contain coherent information except for the four completely illegible comments. Furthermore, as the tablets, software, server and other equipment were reused, and it required less workload, this intervention was likely to have been more cost effective than paper marking. However, it was not possible to compare pre and post-

intervention costs because there was no exam-specific financial accounting at the UON School of Medicine at the time.

As this study found digital marking useful and practical in a summative context, the use of digital marking could be extended to formative assessments also. Student-led mock OSCEs could be conducted using digital marking and feedback. As discussed in Chapter 2, most students do not mind the extra workload to conduct peer or near-peer led formative assessments (Young, Montgomery et al. 2014). They provide both the peer examiners and candidates learning opportunities (Burgess, Roberts et al. 2013).

4.4.2 Quality and quantity

Despite the time constraint and being informed that feedback giving was an optional task, the examiners were willing and able to provide feedback. All students received both tickbox and free text feedback in multiple stations. The quality of the free text feedback was largely adequate. Only four comments were illegible and there were no unprofessional or abusive comments. Furthermore, generic, unspecific comments formed only a small minority. A generally accepted and simple criteria for good quality feedback is that it should address what the student did and provide advice for improvement (Denison, Bate et al. 2016). Nearly a quarter (24.18%, N = 387) of the free text comments (N = 1600) fit this criteria. However, it was impossible to verify how accurate these comments were. One student raised concerns that their recollection was different from the examiner's comments. Accuracy of the feedback comments can and should be tested (Moineau, Power et al. 2011). As found in Chapter 2, evidence supported the feasibility and acceptability of video recording in formative OSCEs and, therefore, it could be used to assess the accuracy of feedback.

In terms of quality, more than half of the comments (53.56%) belonged to grade five: they addressed two or more actions of the student without offering any explicit advice. Some of these comments were long and detailed. They may have contained more useful information than some of the short comments that met the grade six criteria for good quality feedback. It should be noted that neither the free text feedback section on tablets nor the examiner training provided any guidance for examiners to give students specific high quality feedback.

The quantity distribution of both types of feedback followed closely to the normal distribution with small positive skews and standard deviations. Nearly all students received both kinds of feedback in most stations. It indicates that most students received feedback in quantities close to the average of the whole cohort. As it is easier to tick off pre-written items than type on the tablet, examiners produced a larger number of tickbox ticks than words.

The eight stations produced varying amounts of feedback. As each station contained a unique scenario, they may have offered different opportunities for tickbox and free text feedback. There was no single station that produced significantly more or less feedback overall.

Denison et al's study compared the quality and quantity of feedback generated by digital marking to that of paper marking (Denison, Bate et al. 2016). It took place in a similar context to the present study. Both studies took place in large scale OSCEs. Denison's study collected data from 3802 assessments while the present study did so from 2296. Denison's study collected data from both summative and formative OSCEs while the present study was concerned with only summative OSCEs. Denison's study reported producing an average of 13 words per station. In comparison, the present study produced an average of 15.72 words per

station. Denison's study did not involve tickbox feedback. As the quality appraisal criteria were different between the two studies and Denison et al only reported the mean of the quality grading on a five point scale, it was impossible to make a meaningful comparison.

No other study was similar enough for direct comparison in terms of feedback quality. However, it was possible to make general comparisons to some of the studies identified in the literature review, Chapter 2. One study involving 173 students instructed their examiners to include suggestions for improvements and action plans in their feedback comments after formative mini-CEX workplace based assessments (Fernando, Cleland et al. 2008). A much larger proportion of the comments (71.8%) in Fernando et al's study included suggestions for improvement than that of the present study (grade 4 and grade 6, 31%). However, examiners in the present study produced better quality feedback than those in a smaller study involving 33 students on formative mini-CEX feedback in which very few feedback comments expressed forward-focused learning goals or suggestions for improvement (Harvey, Radomski et al. 2013).

Failed students received significantly more free text feedback than the rest of the cohort while the Pass and Honours students commonly received shorter, unspecific comments such as "excellent". The examiners did not know for certain during the exam who was going to pass or fail. As such, it is clear that examiners put in more effort to help students who made mistakes than to help competent students excel. This is in line with a fundamental principle in medicine to do no harm and a common sentiment in summative assessment culture to prioritise passing the exam over perfecting one's practice as discussed in Chapter 1. However, it could also be argued that all learners should receive equal attention and feedback.

As evidence shows that feedback givers tend to positively distort their comments as low ratings are considered to be demotivating (Daelmans, Hoogenboom et al. 2004, Daelmans, Hoogenboom et al. 2005), feedback for the failed students may have contained positively distorted messages. This highlights the implication to test the accuracy of the examiner comments in future research.

All tickbox feedback items were phrased negatively except those under "General Positive Feedback" section as shown in Figure 3. As most of the tickbox items were negative and failed students should have merited more negative feedback, it is curious why the examiners did not give them significantly more tickbox feedback although they gave them significantly more free text feedback. Future research should extend the evidence on the examiners' feedback giving behaviour: what type of examiners give what kind of feedback to what type of students?

Examiners should be informed of the criteria for good quality feedback and discouraged from writing unspecific short comments. Giving honest constructive feedback should be encouraged. Quantity of the feedback may be improved if longer gaps could be accommodated between the stations. However, a similar study which recorded verbal feedback had only a 90 second gap between the stations, which was an increase from a 60 second gap they used previously (Harrison, Molyneux et al. 2015). The present study had two minute gaps. Lengthening the gaps would increase the overall exam time and add workload, physical and mental stress for all involved and may not result in producing more useful feedback.

Evidence supports that specific feedback is more helpful than generic feedback, as discussed under subheading "2.5.3.1. Specificity" in Chapter 2. However, the purpose of the "General Positive Feedback" sections in the present study was to provide non-specific feedback. Furthermore, as

mentioned previously, there was no guidance for examiners on how to give specific feedback. These shortcomings were due to the researcher trying to comply with the school's directive to avoid "giving away" their summative OSCE scenarios via feedback. Quality of feedback could potentially be improved by removing the "General Positive Feedback" sections and making all domain tickbox sections a more balanced mixture of positive and negative items. Future free text sections could be divided to subsections which specifically addressed what the student did, what went well and what improvements were required. This may improve the specificity of free text feedback.

4.4.3 Feedback engagement

The present study's finding on feedback engagement was similar to that of the Harrison et al' study (Harrison, Konings et al. 2013). Harrison et al reported a high log-in rate at 95.7% for the whole sample. However, as the Harrison study's website contained a numerical breakdown of scores by each station but not the examiner comments, the students may have been viewing the feedback only to see which stations they performed well. The present study's website contained only the examiner feedback: all those who logged in were there to access only the examiners' comments. Still, more than 90% of students logged in to view the feedback in the present study.

Harrison et al had three performance categories for those who passed (Excellent, Good, Just pass) while the present study had only two (Honours, Pass). Honours students were the keenest engagers in both studies. Harrison et al reported that borderline performance, "Just pass" group viewed feedback significantly less than the rest of the cohort. In the present study, the Pass student group engaged with feedback less than the Honours group. Declining log-in rates after the resit in the present

study further indicate that lower performance students did not engage with feedback as much as the higher performance ones did. Everyone in the Fail group in both studies had to read their feedback. As discussed in Chapter 2, further evidence in the literature shows that lower performing students tend to be less enthusiastic about feedback engagement or even avoid it (Sinclair and Cleland 2007, Winston, Van der Vleuten et al. 2010). For such students, only giving more feedback will not address their learning needs. A better solution is to help students understand what the feedback is aimed for and how to best utilise what is available. Learners with low feedback orientation would simply demand for more feedback and put all the responsibility for learning regarding feedback on the educators whereas those with strong feedback orientation have a sense of responsibility to learn from the feedback they received and proactively seek feedback (London and Smither 2002, Linderbaum and Levy 2010). To improve the learners' feedback orientation, their current level of feedback orientation needs to be measured first.

The researcher was unable to determine how deep the engagement was. This study's feedback website had only one page for the students to log in and see their feedback. As such, it was not possible to see how students read it and what they learned from it. A qualitative study in the future could investigate how deep the learner engagement with feedback was and what lasting changes in learning it led to.

The feedback emails from the students did not mention any specific problems with the intervention but they focused on wanting more feedback or clarification about why they failed. The students were made aware that the purpose of feedback in this context was not to justify why they failed the exam but only to improve their learning. As discussed above, the distribution of feedback across the cohort was adequate. There

were outliers who received significantly more or less feedback than others but this is an inherent phenomenon in normal distribution. The examiners also gave significantly more feedback to failed students. As the examiners were producing adequate quality feedback in adequate amounts for all students, the researcher concluded that there is no need to produce more feedback.

Similar to a finding in Chapter 3 in which some students were wary of the fairness of the examiners although they did not personally encounter any problems, one failed student wanted to have his marks rechecked because he heard some other students' feedback comments were jumbled up. While this request is not directly relevant to feedback, it does reinforce the findings from the interviews about the learners' concerns for fairness in the exams based on hearsay.

4.4.4 Limitations

The face validity of the quality appraisal criteria was not established using independent experts. Its inter-rater reliability was also not assessed. Only one researcher appraised all the comments using this criteria. This carried a risk of bias and error.

Although relatively large, being a single-site cross sectional study may limit the generalisability of its findings.

The tickbox feedback items were generated either by the researcher alone or by small group meetings within the medical school only without following a strong framework for validation.

4.4.5 Summary

This study found that the examiners were able and willing to use digital marking to generate individualised feedback of acceptable quality and quantity for all students during summative OSCEs. As nearly all students

read this feedback, feedback after summative OSCEs was found to be a valid learning opportunity. It generated original evidence showing that failed students received significantly more free text feedback than those who passed. Nearly all of the students read the feedback they received after the main exam. Lower performing students did not read their feedback as much as higher performing students did.

Chapter 5. A Validity Argument for the use of Feedback Orientation Scale in Medical Education

5.1 Introduction

In Chapter 1, an argument to define and measure the learners' feedback-specific constructs to gain a better understanding of their perception to and engagement with feedback was presented. The researcher proposed to use the Feedback Orientation Scale (FOS) to measure the construct of feedback orientation. The definition and the four aspects of feedback orientation were also presented.

In Chapter 3, it was found that the medical students' awareness of the utility of feedback and their sense of responsibility to learn from the feedback they received might be weak. The researcher judged that this was an indication for the need to measure their feedback orientation.

In Chapter 4, it was found that the higher performing students read their feedback more than the lower performing ones did. This indicated that there may be significant differences in feedback orientation between students.

This study's aim was to construct a validity argument for the use of Feedback Orientation Scale in medical education and to assess the learners' feedback orientation. The objectives were presented under objective 4 under subheading "1.8 Research question, aim and objectives" in Chapter 1.

5.2 Methods

This is a quantitative cross sectional study. Approval was obtained from the University of Nottingham Faculty of Health Sciences research ethics committee (reference number - EV200317, 4 July 2017). See Figure 24 in Appendix.

5.2.1 Context

The overall context regarding the University of Nottingham, School of Medicine's (UONSOM) curriculum, training years and exams was presented in Chapter 1. This study was conducted at UONSOM in April 2017, two months after the large scale implementation of personalised summative OSCE examiner feedback to students. By this time, students were receiving learning objective based feedback for summative knowledge assessments and examiner comments for summative OSCEs. The researcher assumed that collecting data on feedback orientation at this time would be more valid than doing so when students were receiving only feedback for knowledge assessments.

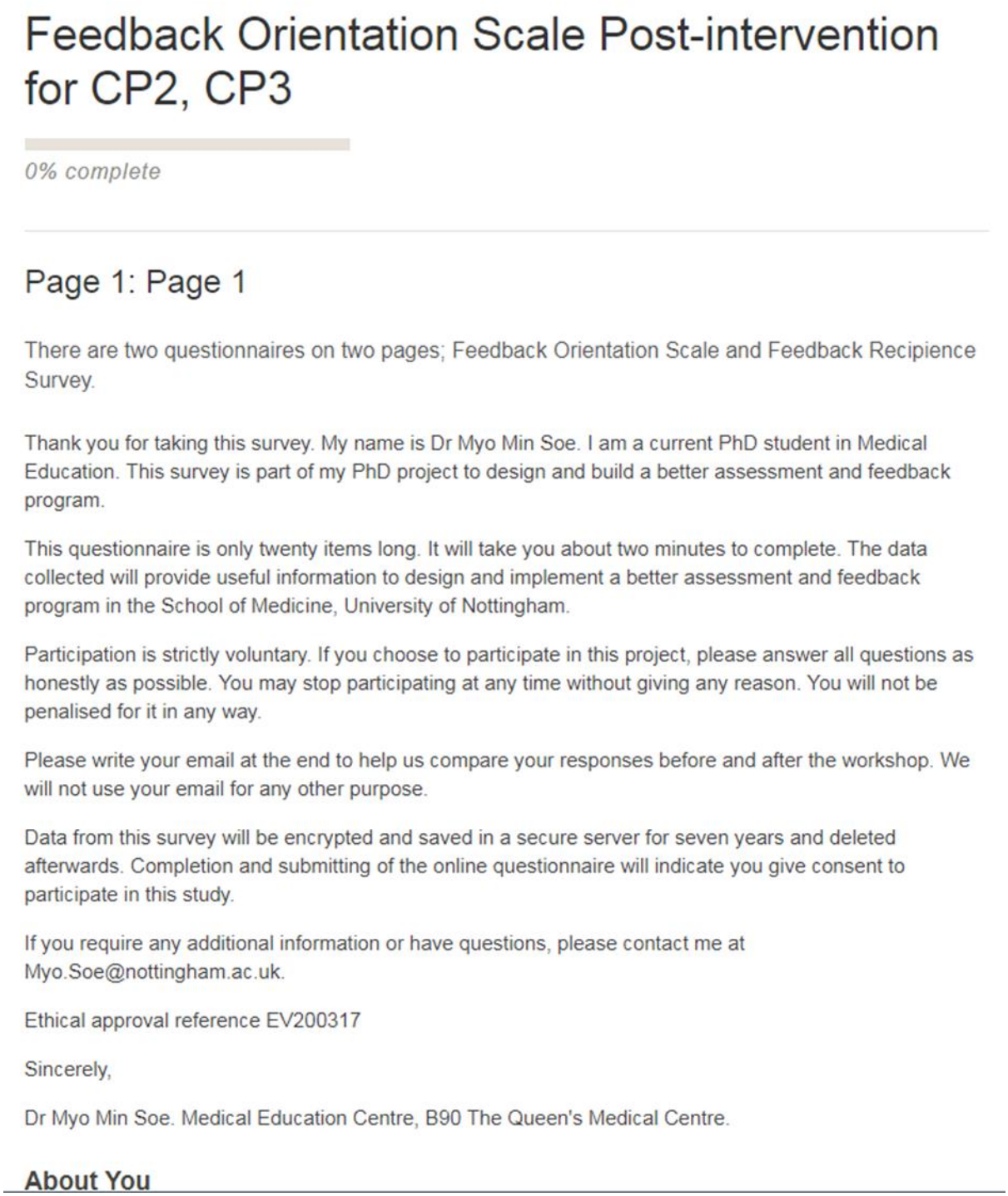
5.2.2 Data collection

Feedback Orientation Scale (FOS) was used. It is a twenty item questionnaire which measures four constructs, Utility, Accountability, Social Awareness and Feedback self-efficacy, which belong to the overall construct of feedback orientation. The origin, rationale and previous evidence regarding FOS was presented in Chapter 1. See Table 8 for the FOS items. See Figure 9, Figure 10 and Figure 11 for its online appearance.

Table 8 Feedback Orientation Scale Items and Domains

Utility
1. Feedback contributes to my success at work.
2. To develop my skills at work, I rely on feedback.
3. Feedback is critical for improving performance.
4. Feedback from supervisors can help me advance.
5. I find that feedback is critical for reaching my goals.
Accountability
6. It is my responsibility to apply feedback to improve my performance.
7. I hold myself accountable to respond to feedback appropriately.
8. I don't feel a sense of closure until I respond to feedback.
9. If my supervisor gives me feedback, it is my responsibility to respond to it.
10. I feel obligated to make changes based on feedback.
Social Awareness
11. I try to be aware of what other people think of me.
12. Using feedback, I am more aware of what people think of me.
14. Feedback helps me manage the impression I make on others.
14. Feedback lets me know how I am perceived by others.
15. I rely on feedback to help me make a good impression.
Feedback self-efficacy
16. I feel self-assured when dealing with feedback.
17. Compared to others, I am more competent at handling feedback.
18. I believe that I have the ability to deal with feedback effectively.
19. I feel confident when responding to both positive and negative feedback.
20. I know that I can handle the feedback that I receive.

Figure 9 FOS online view Top



Feedback Orientation Scale Post-intervention for CP2, CP3

0% complete

Page 1: Page 1

There are two questionnaires on two pages; Feedback Orientation Scale and Feedback Recipience Survey.

Thank you for taking this survey. My name is Dr Myo Min Soe. I am a current PhD student in Medical Education. This survey is part of my PhD project to design and build a better assessment and feedback program.

This questionnaire is only twenty items long. It will take you about two minutes to complete. The data collected will provide useful information to design and implement a better assessment and feedback program in the School of Medicine, University of Nottingham.

Participation is strictly voluntary. If you choose to participate in this project, please answer all questions as honestly as possible. You may stop participating at any time without giving any reason. You will not be penalised for it in any way.

Please write your email at the end to help us compare your responses before and after the workshop. We will not use your email for any other purpose.

Data from this survey will be encrypted and saved in a secure server for seven years and deleted afterwards. Completion and submitting of the online questionnaire will indicate you give consent to participate in this study.

If you require any additional information or have questions, please contact me at Myo.Soe@nottingham.ac.uk.

Ethical approval reference EV200317

Sincerely,

Dr Myo Min Soe. Medical Education Centre, B90 The Queen's Medical Centre.

About You

Figure 10 FOS online view Middle

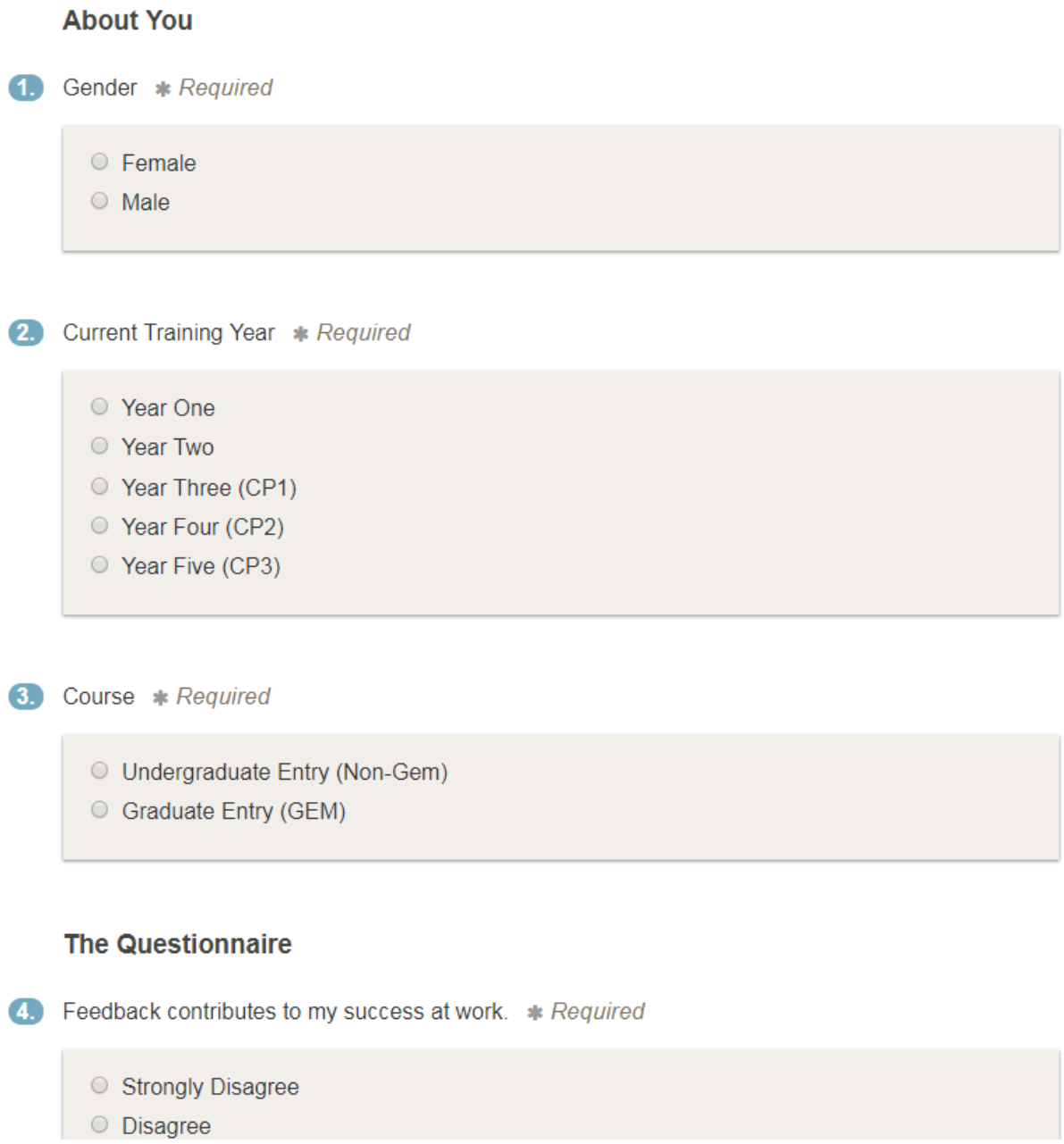


Figure 11 FOS online view Bottom

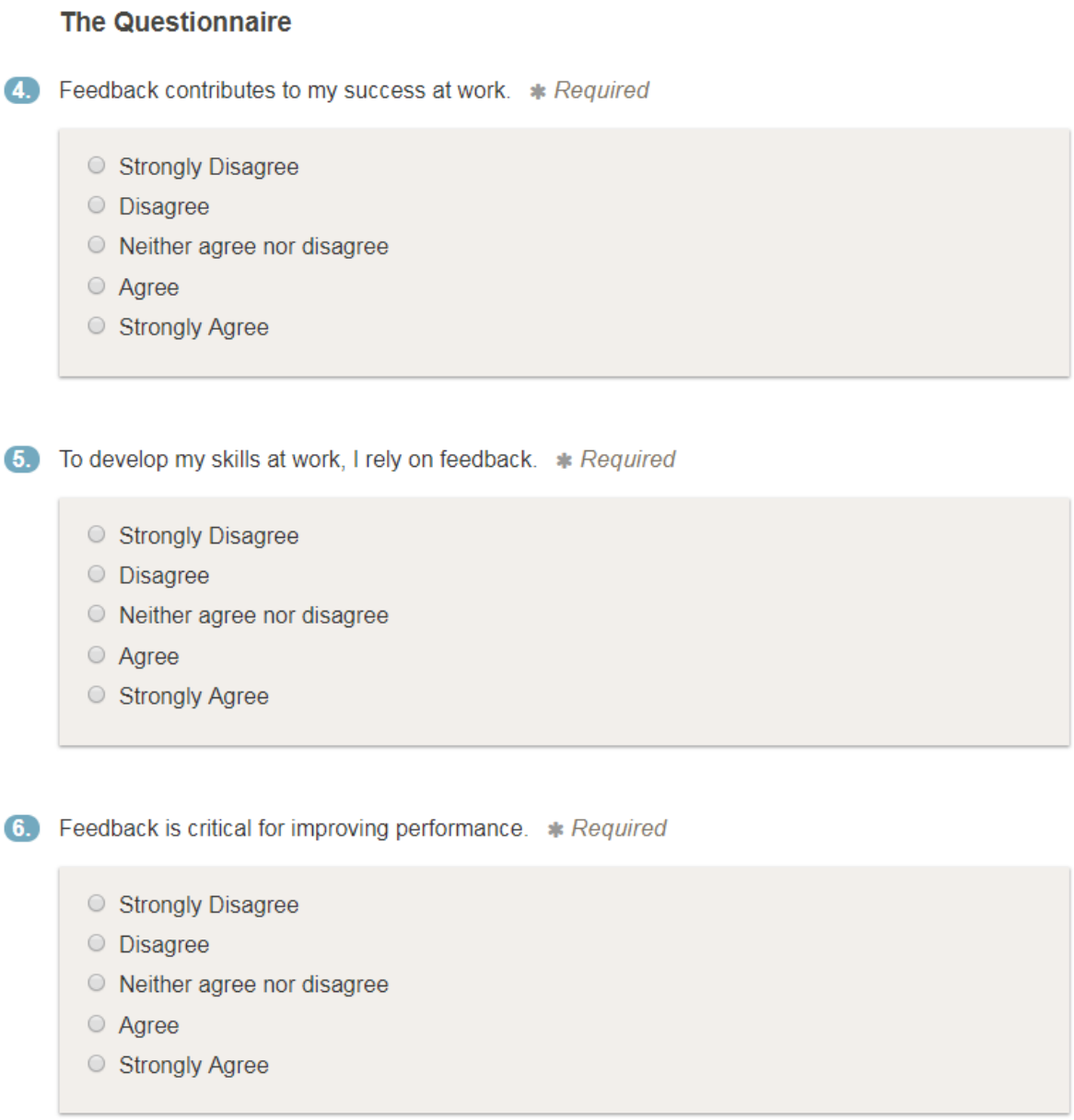


Figure caption - Above is the sample view of Items 1, 2 and 3 of the FOS. The rest of the items are not shown.

The researcher aimed to recruit as many undergraduate medical students as possible to complete the FOS questionnaire. E-mail invitations containing a link to the online FOS questionnaire were sent to all medical students. A data collection desk was also set up in the Medical School foyer. Data on the participants' gender, training year and course were collected. All data were collected digitally using the Bristol Online Survey™ software, which ensured, by design, that there was no missing data. Participation was completely voluntary and anonymous. It took students approximately three minutes to complete the FOS. It used a five point Likert scale; five response alternatives on a continuum from "Strongly Agree" to "Strongly Disagree". The questionnaire did not contain any negative items.

Only one item, Item 4 from the original FOS used in the Linderbaum and Levy study, was modified. The original wording for Item 4 was, "Feedback from supervisors can help me advance in a company." For this study, it was changed to "Feedback from supervisors can help me advance." No changes were made to the rest of the questionnaire. The researcher and the supervisors felt that the rest of the items retained their face validity without any need for modification in the present study's context. No face validity assessment using external experts was done.

5.2.3 Analysis

There were three stages of analysis in this study. The first was Confirmatory Factor Analysis to test the four factor model of FOS using data collected in a medical education context. The second was item analysis using descriptive statistics. The third was an analysis of the effects of demographic variables on the students' FOS scores. All parametric values were approximated to two decimal points.

5.2.3.1 Confirmatory Factor Analysis

Factor analysis is a powerful statistical procedure for scrutinising the relationship between observed variables (item scores) and latent variables termed as factors, constructs, or domains. Confirmatory Factor Analysis (CFA) is used to test how well a known model fits the data (Tavakol, Dennick et al. 2011, Tabachnick and Fidell 2013). Notwithstanding data collected from different groups of participants, items should consistently measure the same constructs if the chosen model is correct. The parameters used in CFA indicate how well the model fits the data. A good fit indicates that the questionnaire reliably measures the factors shown in the model.

Linderbaum and Levy (2010) reported undertaking Confirmatory Factor Analysis (CFA) using Mplus™. Lilford et al used LISREL™ 8.72. This study used AMOS 24. Responses to the questionnaire were coded and entered into SPSS® 22. A score of five was given for strongly agree, four for agree, and so on. Descriptive analyses were performed on all items. Coefficient alpha or Cronbach's alpha > 0.70 was considered to be an acceptable reliability coefficient for determining the internal consistency of the scale (Tavakol and Dennick 2011). More on the coefficient alpha was described below under item analysis. Confirmatory Factor Analysis was done using AMOS 24, a statistical software by IBM™ for structural equation modelling. The researcher performed confirmatory factor analysis on the variance-covariance matrixes. Maximum likelihood procedure was applied to calculate parameters (standardised estimates). Five goodness-of-fit indices were proposed to assess the model fit: Chi squared (X^2) and its subsequent ratio with degrees of freedom (X^2/df); comparative fit index (CFI) and root mean square error (RMSE); standardised root mean square residual (SRMSR) and Tucker Lewis Index (TLI). Since each index has its own assumptions, strengths and

weaknesses, multiple indices are routinely used to enable a comprehensive assessment in CFA.

The chi-square statistic was calculated to assess the fit between the hypothesised statistical model and the set of observed variables. A statistically significant X^2 ($p \leq 0.05$) indicates that the model does not fit the data. Complex models and larger data sets result in larger X^2. Relying on X^2 alone would lead to type 2 error. Comparative Fit Index (CFI) compares the fit of a null model with the fit of the researcher's model using X^2 and degree of freedom. In the null model, latent variables, or factors, are assumed uncorrelated and independent. A value of 1 indicates a perfect fit, while values ≥0.95 show good fit (Babyak and Green 2010). The Tucker-Lewis index (TLI) is a more conservative index with a similar formula. It tends to reject more models than CFI. A TLI ≥0.95 indicates a good fit (McDonald and Marsh 1990). The Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). The error value should be ≤0.06 to show a good fit. All of the above indices penalise for model complexity by including the degree of freedom (df) in the calculations. Therefore they are known as incremental fit indices or parsimonious fit indices (Tabachnick and Fidell 2013). In comparison, Standardised Root Mean Squared Residual (SRMSR) is an absolute fit index which does not penalise for the complexity of the model. It is the standardised difference between the proposed correlation matrix (model) and observed correlation matrix. The residual would be 0 for an identical fit while a value <0.08 is acceptable for a good fit (Hooper, Coughlan et al. 2008).

Using these indices, three hypothesised models were tested. The one factor model hypothesised that all twenty items measured the same factor directly. The four factor model hypothesised the four factors were

measured by five items each, as described above. The second-order model with an overarching construct on top of the four factors hypothesised that all four factors belonged to one overarching factor, overall feedback orientation. Linderbaum and Levy proposed that the second-order model was the best model to represent the constructs measured by the FOS questionnaire (Linderbaum and Levy 2010). The purpose of testing three different hypothesised models was to compare the goodness of fit indices between them.

5.2.3.2 Item analysis

Each item was analysed using its mean, 95% confidence interval of the mean, 5% trimmed mean, corrected item-total correlation, standard deviation, skewness, and its effect on the whole questionnaire's coefficient alpha. These are descriptive statistics to summarise and represent the data in various ways.

An item's mean is its average score. To clarify the terminology in this paper, any reference to the "mean" refers to the sample's mean unless specified otherwise. The mean, like any statistical parameter, comes from the data gathered from a sample. As such, the sample's mean may not be the same as the true mean, mean of the whole population. 95% confidence interval of the mean is a range of values in which one could be 95% sure that the true mean lies. As it is calculated from sample size and variations in the data, large samples with small variations produce small 95% confidence intervals while small samples with large variations would do the opposite. 95% confidence interval of the mean shows how close the sample's mean is to the true mean.

5% trimmed mean is the mean calculated after removing the highest and the lowest 5% of the data. On its own, it shows what the mean would be if outliers are not present. Comparing 5% trimmed mean with the mean

show where these outliers are, the minority which gave extremely different scores than the majority, pulling the mean towards them. If 5% trimmed mean is not different from the mean, it indicates either the absence of outliers or the presence of two equally extreme outlier groups of the same size.

Item total correlation represents the relationship between an individual item and the whole questionnaire. Values between 0 and 0.19 indicate poor correlation: values between 0.2 and 0.39 indicate good correlation: and values 0.4 and above indicate very good correlation (Howard and Forehand 1962). The value becomes negative when there is no correlation at all. It could be because the item was vague, the participants were insincere or there was not enough time to complete the questionnaire, among other possible reasons.

Standard deviation shows the variation of data from the mean. A small standard deviation value indicates that most of the data are close to the mean while the opposite is true for a large standard deviation.

Skewness is a measure of the extent the data "lean" to one side of the mean, in comparison to the normal distribution curve. A negative skew indicates that the bulk of the data lean to the right side of the mean or most of the data have larger values than the mean. A positive skew indicates that the data lean to the left side of the mean. Zero skew indicates that the data are distributed symmetrically to both sides of the mean.

Coefficient alpha or Cronbach’s alpha of a questionnaire is a measure of its internal consistency. Its value ranges between 0 and 1. For questionnaires on attitudes without right or wrong answers, an alpha value of 0 indicates there was no consistency at all to how students responded to the questionnaire while an alpha value of 1 indicates every

single participant gave the exactly the same response to the questionnaire. These extreme values could exist when the items are either too easy or too difficult. A coefficient alpha value between 0.70 and 0.90 is generally considered to show acceptable internal consistency (Tavakol and Dennick 2011). All items contribute to the whole questionnaire's alpha together. This study calculated how deleting each item affects the coefficient alpha of the whole questionnaire. If deleting an item lowers or makes no difference to the alpha of the questionnaire, the item is not redundant as its presence improves or sustains the internal consistency of the questionnaire. If deleting an item increases the alpha value of the whole questionnaire, the item is redundant, and it should be removed as long as doing so would not make the alpha value exceed 0.90.

5.2.3.3 Effect of demographic variables on FOS scores

Afterwards, univariate ANOVA tests were undertaken to determine the individual and combined effects of gender, course and training year on the observed variables. R, a free software environment for statistical computing and graphics was used to generate coloured box plots.

5.3 Results

Four hundred and twenty two students (N = 422) responded to the questionnaire representing a response rate of 29.72% out of a total population of 1420 registered undergraduate medical students at the time. There was no missing data. No GEM students in preclinical years participated in this study. Table 9 shows the demographic data of the whole student population and this study's participants.

Post hoc sample size calculation showed that 302.49 participants would have been required from the total population (N = 1420) if the intended standard deviation was set at 0.5, margin of error at 5% (e = 0.05) and confidence level at 95% (z-score = 1.96). Instead, having 422

participants resulted in a standard deviation of 0.81 and a standard error of 0.03.

Table 9 A description of the participants' demographics

Demographic data	Participants (N =422) N (%)	Total student body (N = 1420) N (%)
Gender		
Female	269 (64%)	814 (57%)
Male	153 (36%)	606 (43%)
Course		
GEM	65 (15%)	170 (12%)
Non-GEM	357 (85%)	1250 (88%)
Year of Study		
Year 1	77 (18%)	240 (17%)
Year 2	65 (15%)	227 (16%)
Year 3	82 (19%)	328 (23%)
Year 4	59 (14%)	290 (20%)
Year 5	139 (33%)	335 (24%)

Note. GEM stands for Graduate Entry Medicine. Non-GEM stands for undergraduate entry to medicine

5.3.1 Reliability

The standard error was 0.03 in all domains. The coefficient alpha for Utility, Accountability, Social Awareness, Feedback self-efficacy and overall feedback orientation fell within acceptable limits (between 0.7 and

0.9 as referenced in the analysis section), the internal reliability of the questionnaire was demonstrated to be good. See Table 10.

Table 10 Reliability of the FOS constructs

Domain	Coefficient alpha	95% CI Interval	No. of Questions
Utility	0.86	0.84-0.88	5
Accountability	0.71	0.67-.076	5
Social awareness	0.81	0.79-0.84	5
Feedback self-efficacy	0.84	0.81-0.86	5
FOS Overall	0.89	0.88-0.91	20

5.3.2 Model Fit

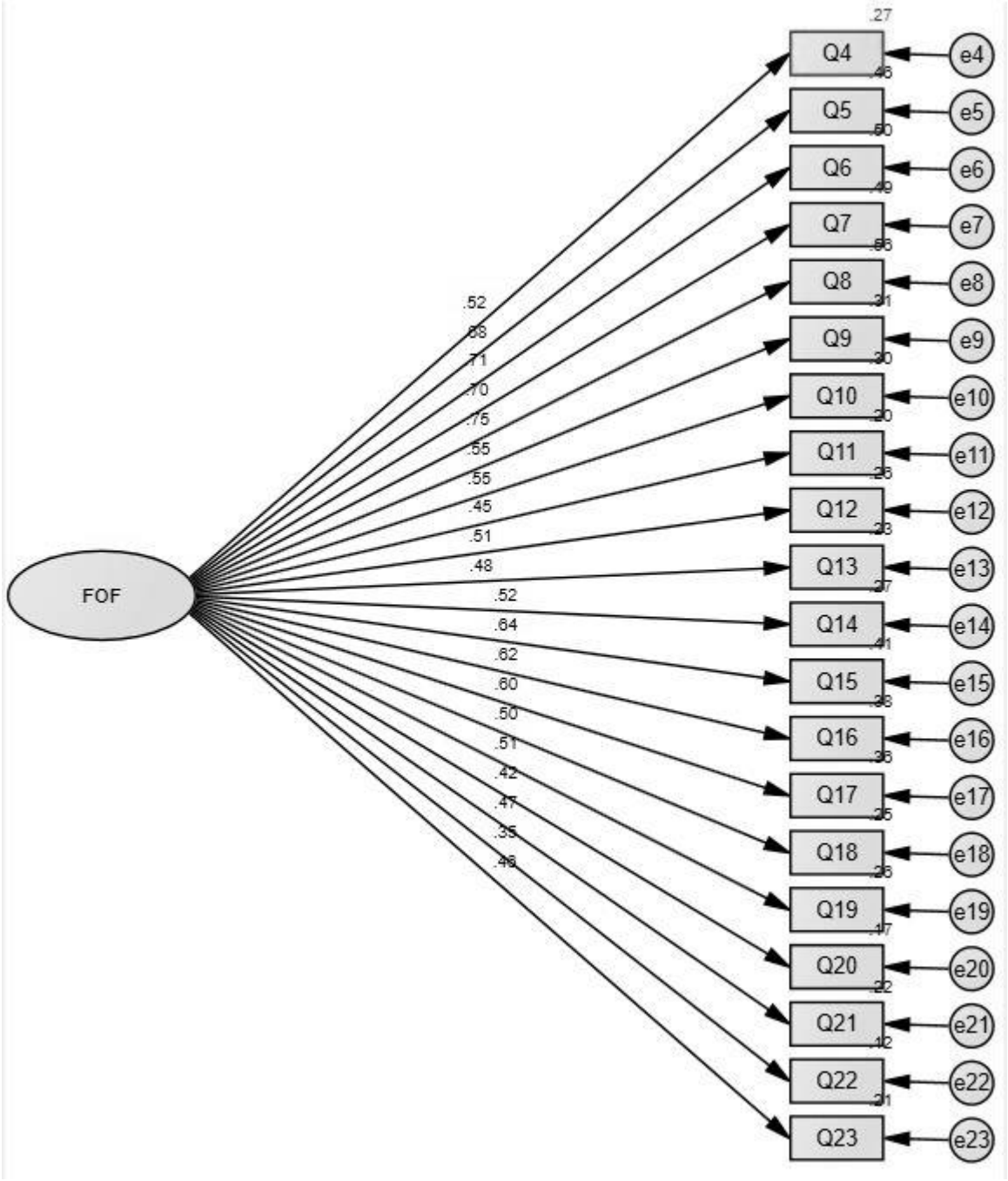
In the following figures, rectangles represent observed variables (Q4 through Q23); ellipses represent latent variables and circles represent measurement errors. Q4 through Q23, the observed variables, represent FOS Items 1 through 20 respectively. One headed arrows between ellipses and rectangles represent regression paths and their numbers represent standardised regression weights. One headed arrows between circles and rectangles represent measurement error terms. Double-headed arrows between ellipses represent factor covariance or correlations between unobserved variables. Table 11 shows fit indices of the models.

Table 11. Results of the CFA for the hypothetical CFA models

Model Name	X^2	df	X^2/df	SRMSR	RMSE	CFI	TLI
One factor model	1465.76	170	8.622	0.10	0.13	0.63	0.59
Four factor model (first order)	437.15	164	2.66	0.05	0.06	0.92	0.91
Hypothesised second-order model	443.25	166	2.67	0.05	0.06	0.92	0.91

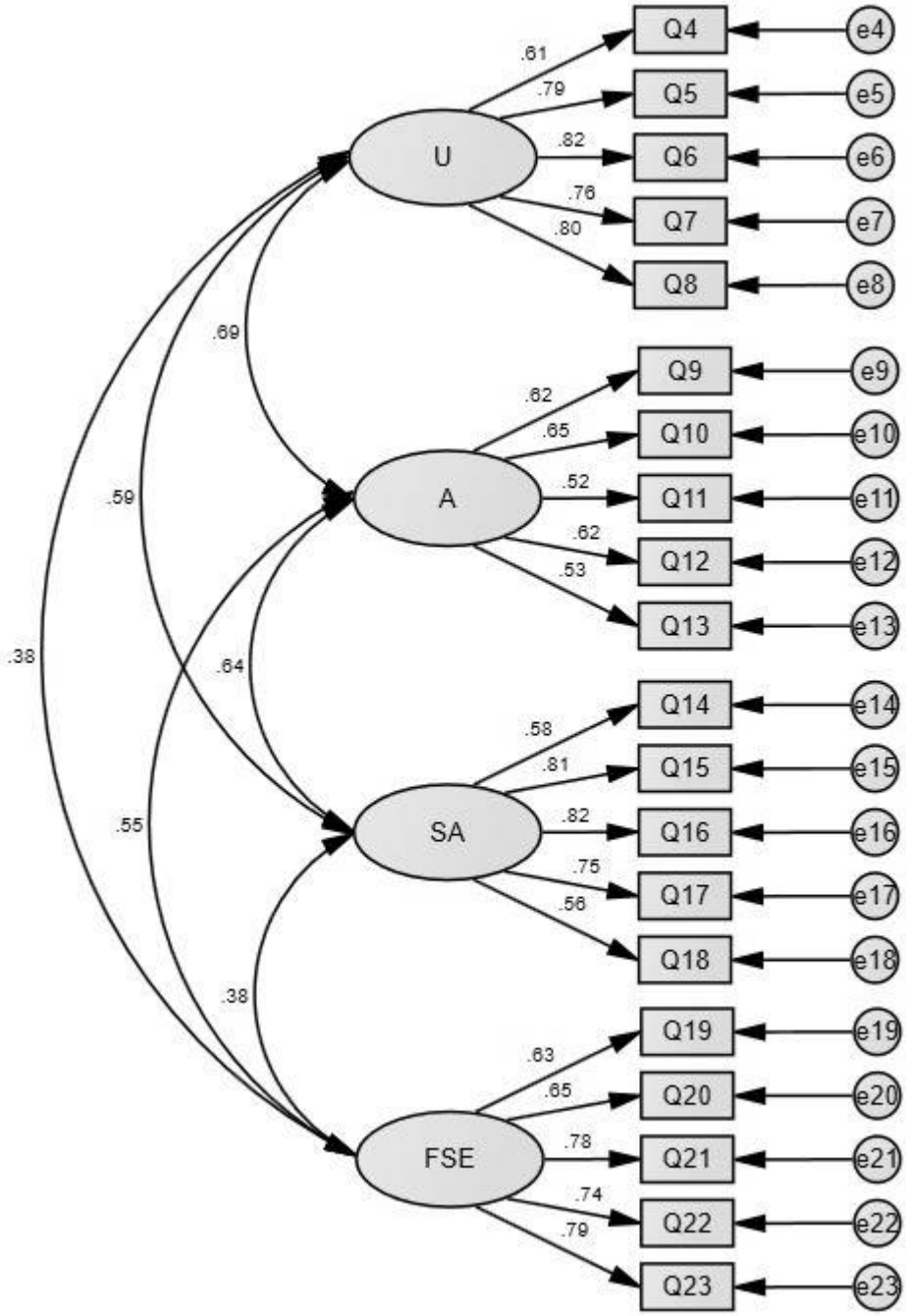
Note - X^2= Chi Squared; df = Degree of Freedom; SRMSR= Standardised Root Mean Square Residual; RMSE= Root Mean Square Error; Comparative Fit Index; TLI= Tucker-Lewis Index

Figure 12 Path diagram of the one-factor CFA model



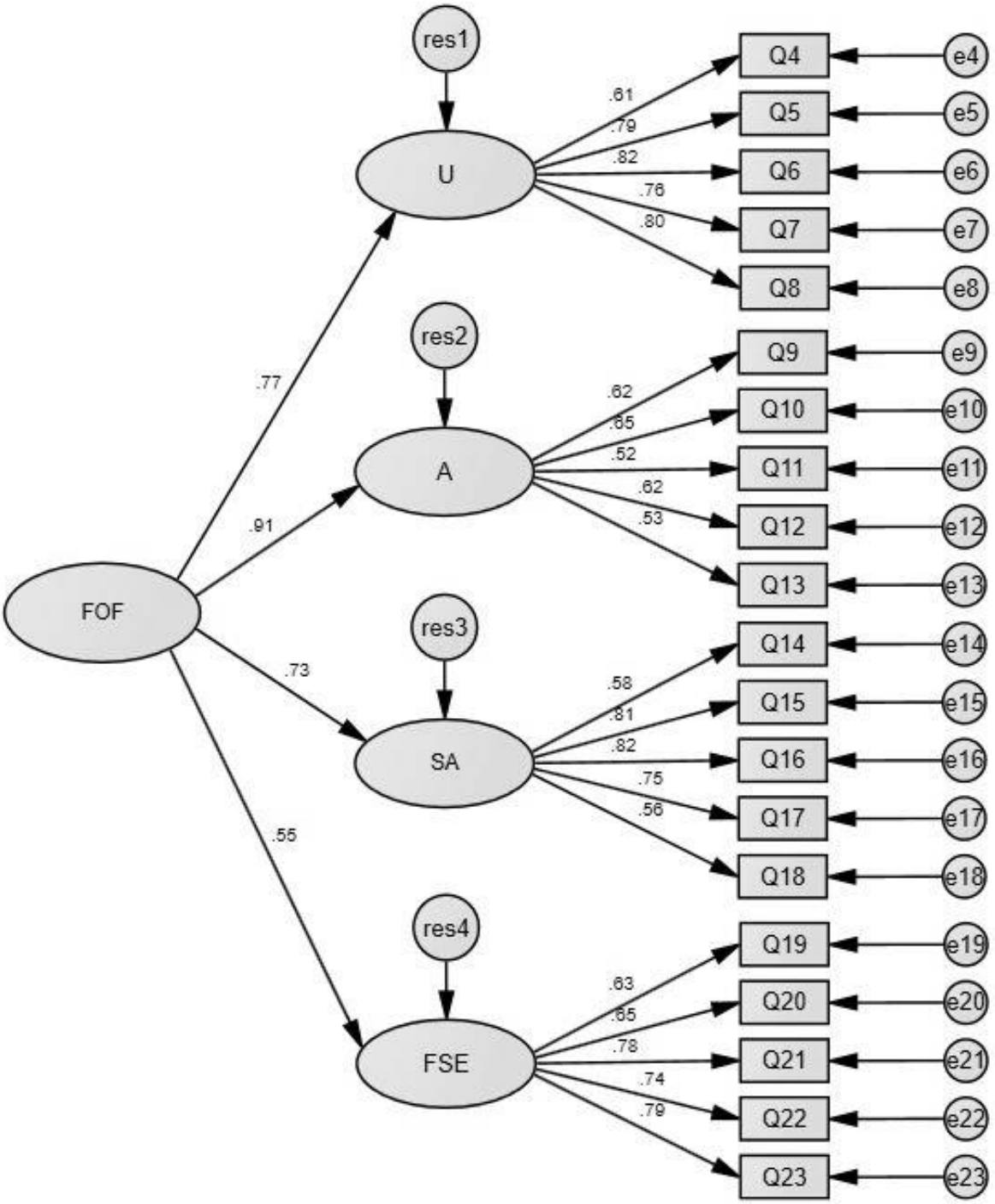
Note - U= Utility; A= Awareness; SA= Self-Awareness, FSE= Feedback self-efficacy, FOF = Feedback Orientation Factor, e = Error, res = Residual

Figure 13 Path diagram of the four factor CFA model (First order)



Note - U= Utility; A= Awareness; SA= Self-Awareness, FSE= Feedback self-efficacy, FOF = Feedback Orientation Factor, e = Error, res = Residual

Figure 14 Path diagram of the second-order CFA model



Note - U= Utility; A= Awareness; SA= Self-Awareness, FSE= Feedback self-efficacy, FOF = Feedback Orientation Factor, e = Error, res = Residual

One factor model had poor fit indices. See Figure 12. Details about the interpretation of the fit indices and their acceptable values are presented in the analysis section above. The four factor model and the second-order model, in Figure 13 and Figure 14 respectively, had acceptable fit indices which were almost identical. CFI and TLI values were smaller than the 0.95 threshold (by 0.03 and 0.04 respectively) in both models. Considering the rest of the indices were within acceptable parameters it was concluded that both models were good fits to the data.

5.3.3 Item analysis

Table 12 presents all the items' descriptive statistics. Items with the largest and smallest parameters were discussed in more detail.

Table 12 Item analysis

ITEM	MEAN	95% CONFIDENCE INTERVAL		5% TRIMMED MEAN	CORRECTED ITEM-TOTAL CORRELATION	STD. DEVIATION	SKEWNESS	ALPHA IF ITEM DELETED
1	4.18	4.10	4.25	4.26	.45	.81	-1.15	.89
2	4.18	4.09	4.26	4.26	.57	.86	-1.07	.89
3	4.44	4.37	4.51	4.52	.61	.72	-1.51	.89
4	4.56	4.50	4.62	4.61	.61	.61	-1.89	.89
5	4.14	4.06	4.22	4.22	.66	.84	-1.04	.89
6	4.36	4.29	4.42	4.43	.51	.70	-1.41	.89
7	4.22	4.15	4.29	4.28	.53	.71	-.90	.89
8	3.17	3.07	3.26	3.15	.44	1.00	.15	.89
9	4.17	4.10	4.24	4.23	.49	.74	-1.09	.89
10	3.88	3.80	3.96	3.93	.48	.87	-.75	.89
11	4.22	4.15	4.29	4.29	.46	.74	-1.18	.89
12	4.13	4.06	4.20	4.19	.59	.74	-1.03	.89
13	4.00	3.92	4.07	4.06	.59	.81	-.96	.89
14	4.12	4.04	4.19	4.18	.56	.76	-1.00	.89
15	3.57	3.47	3.67	3.60	.49	1.03	-.39	.89
16	3.67	3.58	3.76	3.70	.55	.92	-.51	.89
17	3.29	3.21	3.37	3.29	.46	.86	.06	.89
18	3.95	3.88	4.01	4.00	.52	.72	-.86	.89
19	3.73	3.64	3.81	3.77	.39	.89	-.84	.89
20	3.87	3.79	3.95	3.92	.51	.84	-.84	.89

Note - For every item, N = 422, Minimum score = 1, Maximum score = 5, whole questionnaire's standard error = 0.03

Eleven items had means higher than four and the rest had means higher than three. Item 8 (I don't feel a sense of closure until I respond to feedback.) and Item 17 (Compared to others, I am more competent at handling feedback.) had the smallest means at 3.17 and 3.29 respectively while Item 4 (Feedback from supervisors can help me advance.) had the

highest mean at 4.56. These findings showed that students did not feel a strong compulsion to do something after receiving feedback to achieve a sense of closure. They were not strongly confident about their own capacity to learn from feedback. However, they strongly believed that feedback is important for learning or passing exams.

Item 15 (I rely on feedback to help me make a good impression.) had the widest 95% confidence interval of the mean at 0.20 while Item 4 had the smallest 95% confidence interval at 0.11. The size of the 95% confidence interval is the difference between its higher bound and lower bound. This showed that Item 4's mean was closer to its true mean of the population than the rest. As the sample size was the same for all items, variations in the data were responsible for the sizes of these 95% confidence intervals.

The largest difference between 5% trimmed mean and the mean was 0.08 and it was found in Item 2 (To develop my skills at work, I rely on feedback.) and Item 3 (Feedback is critical for improving performance.). This indicates that there were small minorities with extremely different opinions from the majority on these items. Both items had means larger than 4.0. It indicates that while most students rely on feedback for improvement and believe it was critical for improving their performance, a small minority had strong opinions to the opposite. 5% trimmed mean was larger than the mean in eighteen items. 5% trimmed mean was lower than the mean by 0.02 in item 8. These two values were the same in item 17 (Compared to others, I am more competent at handling feedback.).

Item total correlation was strong in all items; the lowest value being 0.39 in Item 19 (I feel confident when responding to both positive and negative feedback.).

Item 8 and Item 15 had the largest standard deviations at 1.00 and 1.03 respectively. Item 4 had the smallest standard variation at 0.61. This

finding is congruent with the 95% confidence interval sizes for these items.

Item 17 and Item 8 had positive skews but all other items had negative skews. Ten of the items had skews bigger than -0.10, the largest being -1.89 for Item 4. This indicated that most of the data were higher than the mean for all items except item 17 and item 8.

The questionnaire's coefficient alpha value was 0.89. This value remained exactly the same no matter which item was deleted. It was an interesting finding. Data and calculations were rechecked and they were found to be correct. This finding conforms to the high item total correlation values described above.

Item 4, "Feedback from supervisors can help me advance", is remarkable. It had the largest mean with the smallest 95% confidence interval, largest negative skew and the smallest standard deviation. Its 5% trimmed mean was also larger than its mean. These findings indicate that a large majority of students strongly and consistently agreed with this item's statement while a small minority strongly disagreed with it.

5.3.4 Overall scores

Students' feedback orientation scores were high. The overall mean score for all students for all items was 3.99 out of 5. See Figure 15. Utility domain had the highest mean at 4.3 and the Feedback self-efficacy domain had the lowest mean at 3.7. Parameters for the Social Awareness and Accountability domains were almost identical. See Figure 16. This indicated that students were well aware of the utility of feedback for learning but they were not as confident about their capacity to use it effectively.

Figure 15 Distribution of the overall FOS scores (N= 422)

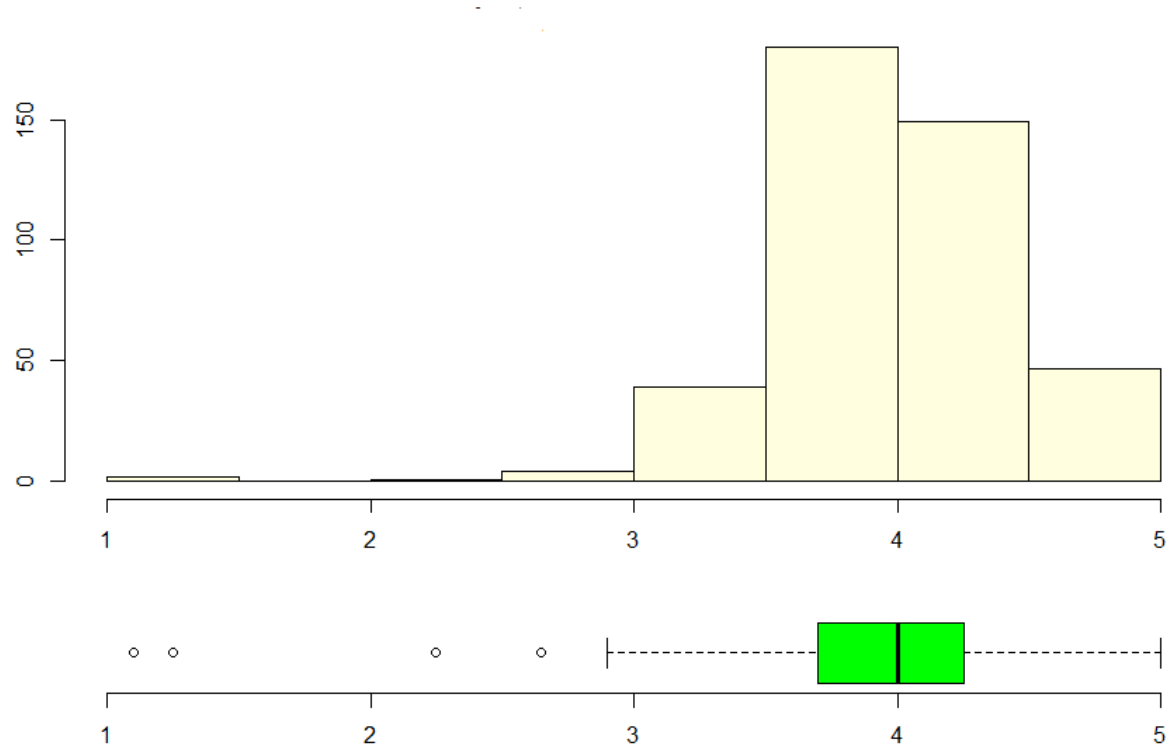
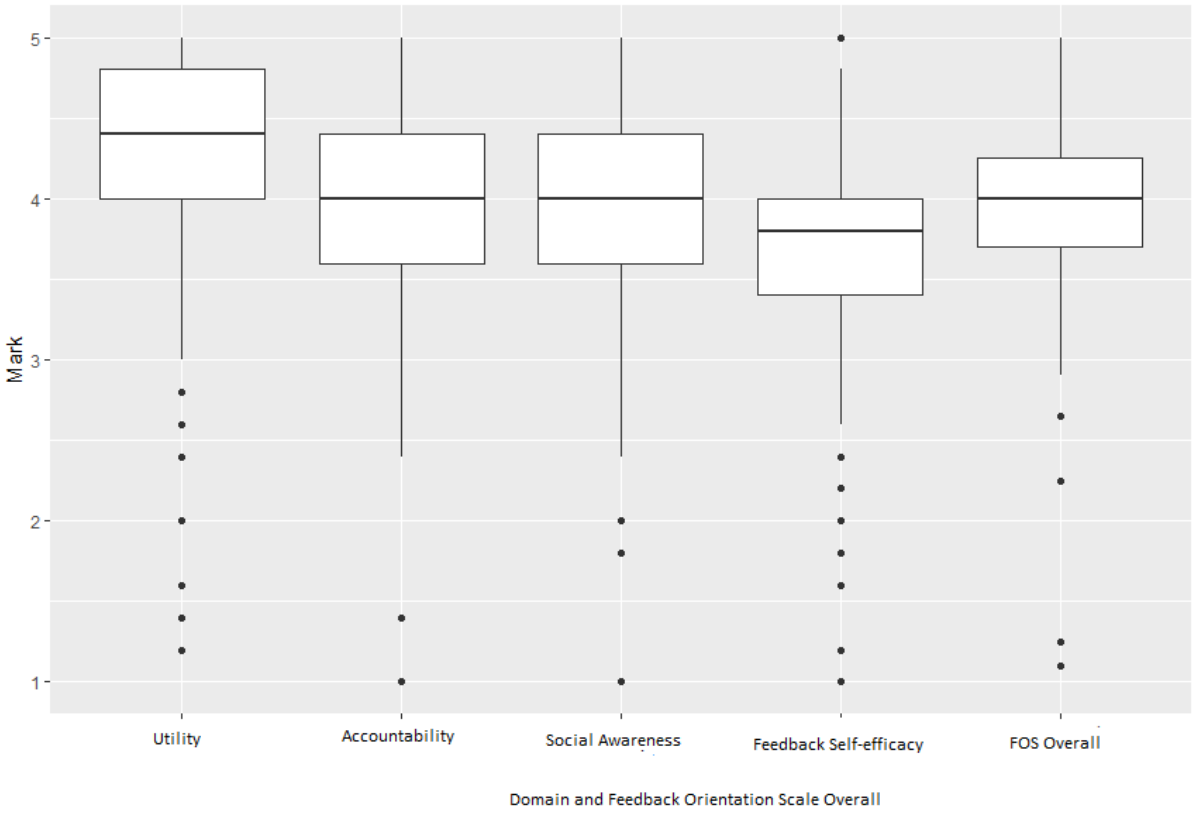


Figure 16 Distribution of the FOS scores according to domains (N= 422)



5.3.5 Relationships between student variables and scores

The overall scores for each of the four factors showed a decreasing trend from year one to year three, but the scores peaked in years four and five ($p = 0.02$). See Figure 17 and Table 13. Males had a significantly higher self-efficacy score than females ($p = 0.02$). See Figure 18 and Table 14. GEM students had significantly higher scores in Utility and Social Awareness domains than non-GEMs ($p = 0.02$ and 0.03 respectively). See Figure 19 and Table 15. The Utility scores increased in year four and five significantly, relative to earlier years ($p = 0.02$), while Social Awareness scores were significantly lower in the first year ($p = 0.00$). There was no other significant difference in any domain according to student variables.

Two-way ANOVA tests showed a significant interaction between the effects of gender and course status on the Utility domain scores, $F_{(1, 418)} = 6.38$, $p = 0.01$. See Figure 20. Females were significantly stronger in Utility than males when they were GEM students ($p = 0.02$). There was no significant interaction between gender and year of study.

Three-way ANOVA to analyse the interaction between all three student variables showed no significant three way interaction in any of the domains. See Figure 21.

Figure 17 Distribution FOS scores by the year of study (N=422)

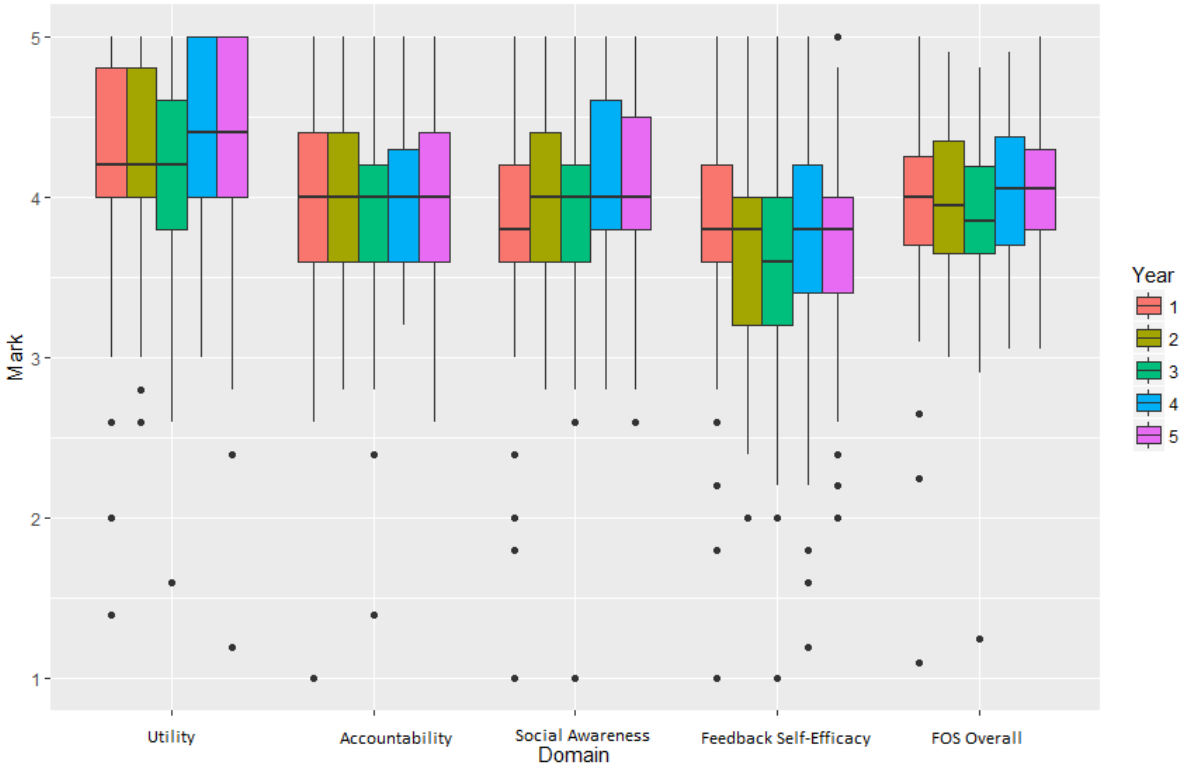


Table 13 Domain means comparison by the year of study (N=422)

	Year of Study						
	Year 1	**Year 2**	**Year 3**	**Year 4**	**Year 5**	**F**	**P-Value**
Utility	4.21	4.24	4.17	4.40	4.41	3.02	0.02
Accountability	3.94	4.00	3.88	3.98	3.99	0.67	0.62
Social Awareness	3.79	4.00	3.90	4.15	4.13	5.33	0.00
Feedback self-efficacy	3.79	3.63	3.55	3.69	3.77	2.09	0.09
FOS Overall	3.93	3.97	3.88	4.06	4.08	3.06	0.02

Note – Significantly higher utility and overall scores in years 4 and 5 than the rest. Significantly lower social awareness scores in year 1 than the rest.

Figure 18 Distribution of FOS scores by gender (N=422)

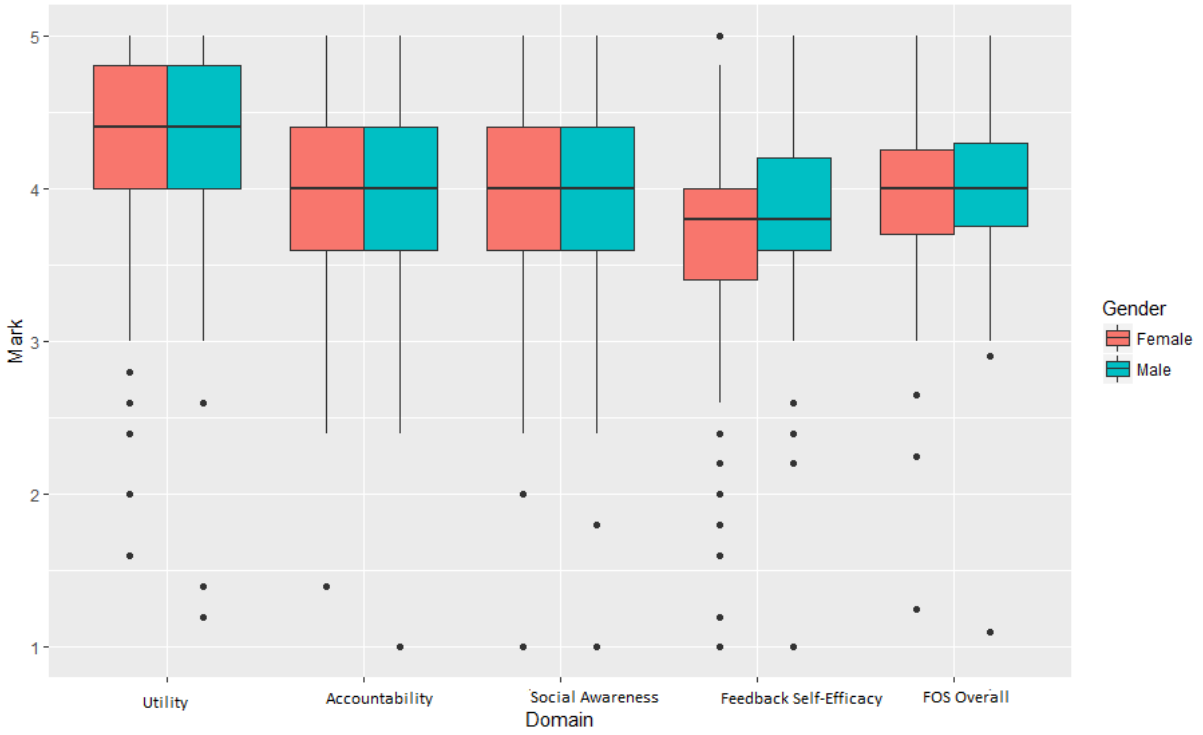


Table 14 Domain means comparison by gender (N=422)

	Gender				
	Females	**Males**	**T**	**DF**	**P-Value**
Utility	4.29	4.31	-0.33	311.289	0.70
Accountability	3.93	3.99	-1.06	303.65	0.28
Social Awareness	4.02	3.97	0.80	301.88	0.42
Feedback self-efficacy	3.62	3.83	-3.19	331.80	0.02
FOS Overall	3.96	4.02	-1.24	297.151	0.21

Note – Significantly higher Feedback self-efficacy in males.

Figure 19 Distribution of FOS scores by course (N=422)

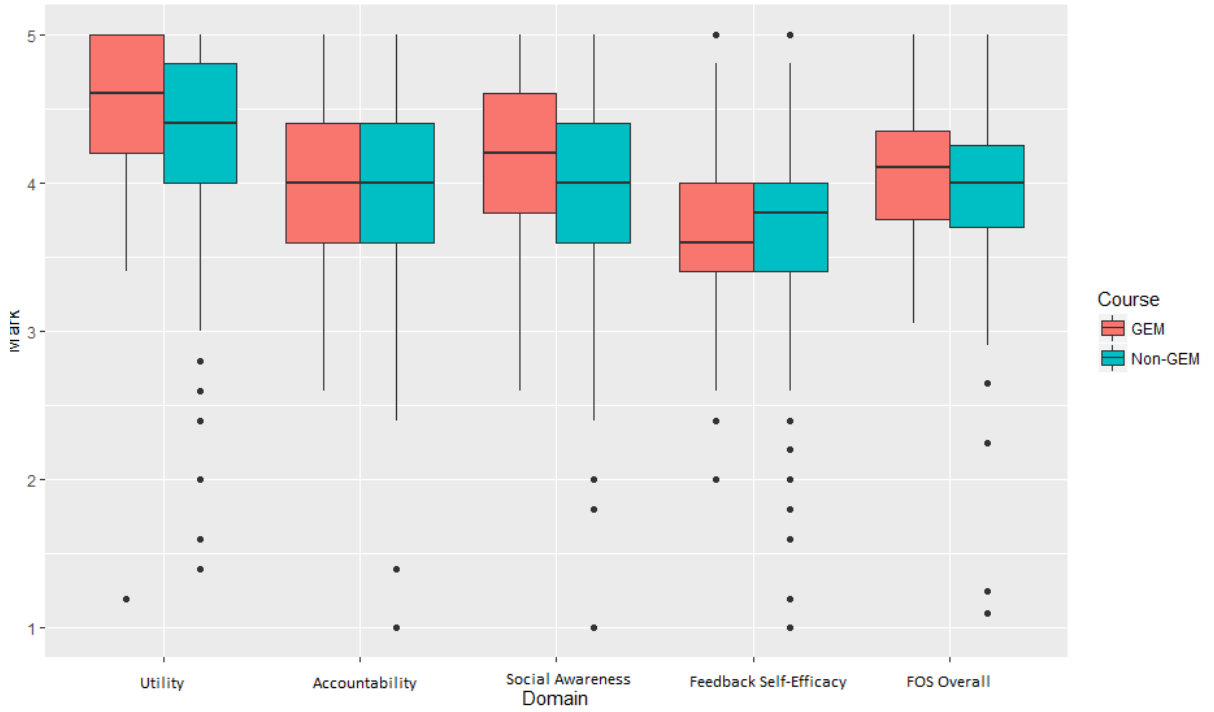
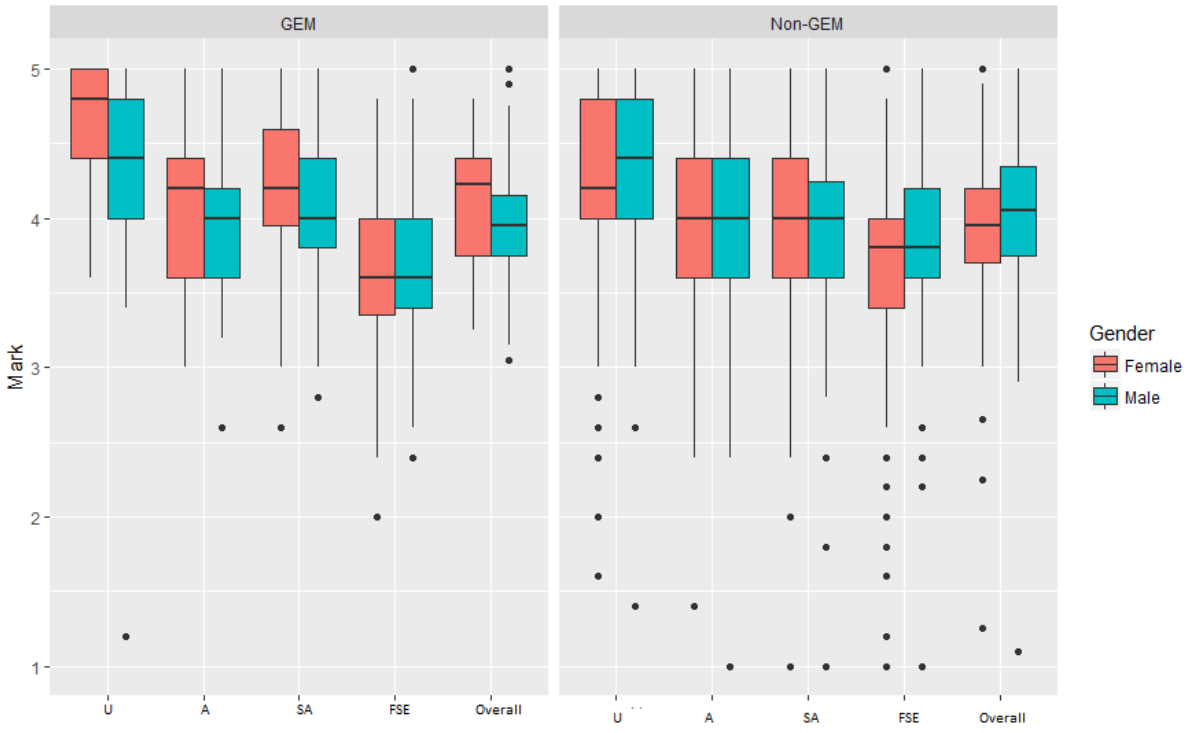


Table 15 Domain scores and the FOS overall means for course (N=422)

	Course				
	GEM	**Non-GEM**	**T**	**DF**	**P-Value**
Utility	4.47	4.27	-2.38	88.85	0.02
Accountability	4.01	3.95	-0.87	90.76	0.39
Social Awareness	4.16	3.98	-2.26	94.57	0.03
Feedback self-efficacy	3.67	3.70	0.42	95.89	0.67
FOS Overall	4.08	3.98	-1.72	95.16	0.09

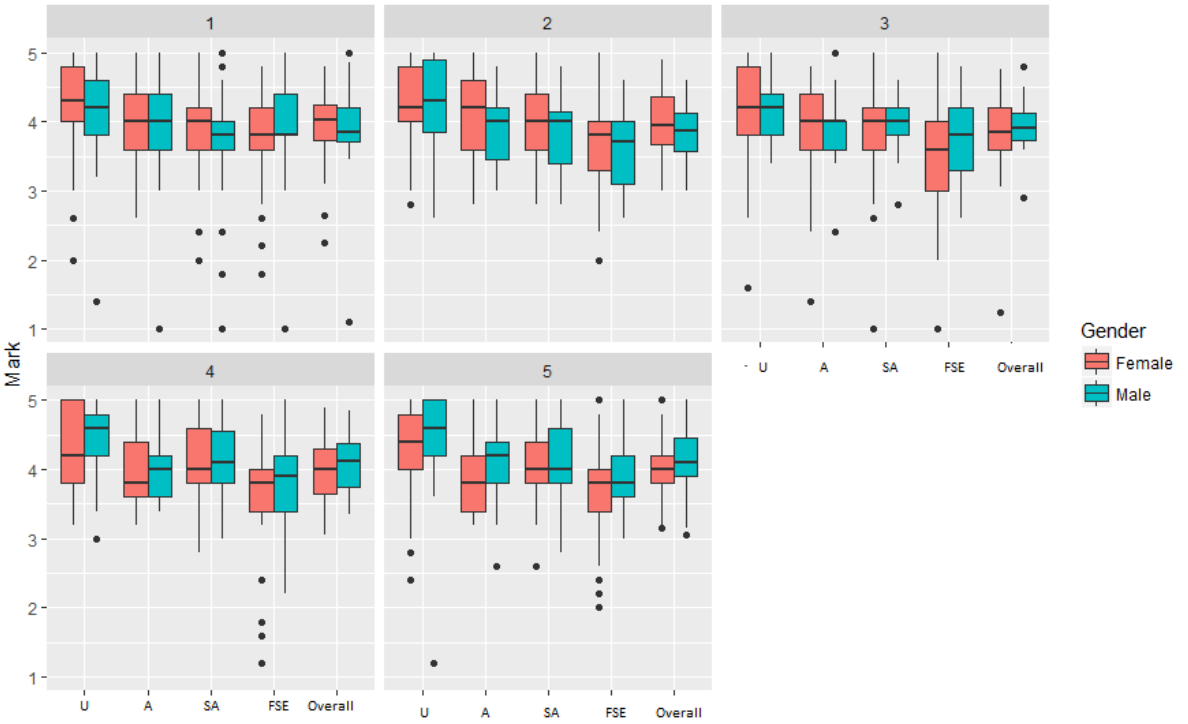
Note – Significantly higher Utility and Social Awareness in GEMs.

Figure 20 Interaction of gender and course (N=422)



Note – Significantly higher utility in female GEMs than the rest.

Figure 21 Interaction of gender and year of study (N=422)



Note – No significant difference.

5.4 Discussion

As the sample was large and included participants from all years and demographic backgrounds in similar ratios to the whole population, it was assumed that the sample represented the existing study body adequately. The reliability and internal consistency of FOS is strong. Every item had high reliability parameters: none of them were redundant.

Confirmatory factor analysis showed that the first and the second-order models had almost identical fit indices, both indicating a good fit to the observed variables. This supports the validity of the second-order model Linderbaum and Levy proposed which represented FOS measuring the four factors belonging to an overarching feedback orientation factor. In comparison, a prior study by Lilford et al in business management conducted on salespersons in South Africa showed closely similar fit indices values to the present study (Lilford, Caruana et al. 2014). The present study's X^2, SRMSR and RMSEA values are marginally stronger but CFI and TLI values are marginally weaker than the Lilford et al's study. Similarity in fit indices despite the difference in contexts showed that FOS was able to consistently measure the same constructs in different contexts.

The overall mean for the whole questionnaire in this study was 3.99. As the maximum possible score was five, the researcher considered 3.99 a high score although it was lower than Lilford et al's mean, 4.18 (Lilford, Caruana et al. 2014). However, while comparing the fit indices between the studies in different contexts is valid, it is less meaningful to compare the feedback orientation scores between medical students in Nottingham and salespersons in South Africa. There was no other similar study in medical education or any healthcare education context for comparison. Most of the items had a negative skew and their 5% trimmed means were

higher than the means. This indicates that there is a small group of students who have very different views to feedback orientation than the majority.

Senior students were stronger in overall feedback orientation. No student group in terms of gender, course and training year had an obviously low feedback orientation score. Older students, especially female GEMs, appreciated the value of feedback more than the rest according to Utility scores. GEM students also showed higher Social Awareness while this did not differ by the year of study. It may be because of GEMs are generally older and had more diverse life experiences. It was not surprising to find that males were more confident of their own ability to handle feedback than females while Feedback self-efficacy was the lowest rated factor overall. This finding was similar to the findings from a previous study on gender differences to reaction and decision making based on the feedback that participants received (Berlin and Dargnies 2016).

As FOS was valid and reliable and the scores were variable between student groups, the implication for practice was to use FOS as a diagnostic tool when giving feedback to individuals or groups. The present study did not attempt to identify individuals with weak feedback orientation. Doing this as part of the feedback process would help an educator gain an insight into the degree of a learner's receptivity to feedback, as well as his or her standing on the four constructs. The feedback giver can then make an informed decision on how to provide feedback and the construct-specific support the learner may need in responding to feedback. Future research is required to design and test the impact of interventions for improving feedback orientation, such as coaching to increase the learners' sense of accountability to be open to and use the feedback they receive, small group discussions to improve social awareness and information

resources to improve the learners' awareness of the utility of feedback. It would also be interesting to see whether the students would be open to feedback regarding their openness to feedback. As a learner's attitude changes over time due to environmental and personal changes, longitudinal research could provide information on the development and shaping of feedback orientation.

A methodological limitation of self-rated questionnaires is that a participant's response may not be the same as how they actually think or act. The results may also be affected by response bias: those who were willing to participate were those who had positive attitudes about the issue in question. The extent to which the responses accurately reflect reality due to social desirability and inaccurate recall cannot be determined. For a more comprehensive understanding of complex factors that determine social attitudes such as feedback orientation, mixed methods research can be used in the future.

More evidence is required on the relationship between a person's feedback orientation and their actual learning behaviour, at least in terms of whether they read the feedback they receive. The researcher missed this opportunity to do so using the website log-in data and FOS. The researcher measured feedback engagement rates and FOS scores separately from two different student groups. It is feasible to use only one participant group to identify the relationship between the learners' feedback orientation and their actual feedback engagement behaviour.

In Chapter 4, it was reported that one student emailed about wanting more feedback despite receiving feedback in thirteen out of sixteen stations. It would be interesting to see whether this student has high feedback orientation. Future research could study the relationship between demanding feedback, reading feedback and feedback orientation.

5.4.1 Limitations

Generalisability of these results may be limited as data were gathered at a single institution.

This study did not calculate the required sample size before collecting the data. Some researchers have argued that sample size should always be calculated a priori since post-hoc power calculations have no value once the study has concluded (Schulz and Grimes 2005, Hickey, Grant et al. 2018).

5.5 Summary

This study generated psychometric evidence supporting the validity of Feedback Orientation Scale in medical education. While the students' feedback orientation is generally strong, a few areas for improvement were identified, such as feedback self-efficacy and accountability. Further research should focus on the relationship between FOS scores and actual learning behaviours of individual students.

Chapter 6. Summary

This chapter presents a summary of the three studies over the course of this PhD project and the researcher's reflections on important findings, implications and shortcomings.

6.1 Summary

The researcher viewed personalised feedback after summative OSCEs to all students as a potential learning opportunity. However, evidence showed that the learners' engagement with feedback was often poor. To deliver effective feedback, it was necessary to firstly understand the learners' perspective, secondly know whether the learners read the feedback they receive and thirdly measure specific constructs that influence their receptivity to feedback.

Therefore, a literature review on the best approaches to timing, mode of communication, feedback giver and content regarding feedback after summative clinical skill assessments was conducted. Digital technology was found to be the most suitable tool for delivering written examiner feedback to large numbers of students after summative OSCEs.

Secondly, the researcher conducted a qualitative study regarding the learners' perception of their assessment and feedback programme at the time and their opinions for improvement. It was found that the students felt they were not receiving the kind of feedback they needed. The students wanted detailed individualised examiner feedback to be offered to everyone. However, they were not sure whether or how they will use the feedback although they wanted it. This further indicated a need to evaluate whether the students read their feedback and also to define and measure specific constructs that influence a learner's receptivity to feedback.

Findings from the literature review and the interviews informed the design of the intervention. The researcher designed and delivered a feedback intervention after summative OSCEs using digital tablets for marking and recording written feedback and delivered it online at the same time as the results. This improved the efficiency of delivering OSCEs, compared to paper marking. The examiners were found to be able and willing to generate individualised feedback of acceptable quality and quantity for all students despite having limited time using new technology and knowing that feedback was an optional task. There were few illegible feedback comments. Failed students received significantly more free text feedback than those who passed. Nearly all of the students read the feedback they received after the main exam. Lower performing students did not read their feedback as much as higher performing students did. These findings supported the feasibility, efficiency, acceptability and learner engagement regarding written examiner feedback after summative OSCEs.

After the feedback intervention was implemented, the learners' feedback orientation was measured. Confirmatory factor analysis generated psychometric evidence to support the validity of Feedback Orientation Scale in medical education. Feedback Orientation Scale was found to be a reliable questionnaire with high internal consistency. The students' feedback orientation was found to be generally strong. Senior students showed significantly higher overall feedback orientation than the rest. Female GEM students scored higher in Utility and Social Awareness while males showed higher Feedback self-efficacy. First year students were significantly weaker in Social Awareness.

6.2 Reflections on findings

The overarching research question was; "What are the elements of a feasible and acceptable approach to the feedback process to students following high stakes summative OSCEs?"

Evidence generated by mixed method research indicated that utilising digital technology, giving written examiner feedback along with results and being aware of the learners' perspective, feedback engagement and feedback orientation are the elements of a feasible and acceptable feedback process to students following high stakes summative OSCEs.

To the best of the researcher's knowledge, only one previous study has objectively appraised the quality and quantity of examiner feedback produced in a summative context (Denison, Bate et al. 2016). The present study contributed to this small body of evidence. The present study's examiners produced a slightly higher feedback word count per station than Denison et al's examiners. Both studies found that the examiners were able and willing to produce enough good quality feedback for all students. There were no unprofessional comments. Only four out of 1600 comments were illegible. Nevertheless, two students expressed their desire for more feedback.

This study also contributed to the small body of evidence on learner engagement with summative feedback. Similar to a previous study by Harrison et al (Harrison, Konings et al. 2013), more than 90% of students read their feedback online. The difference was that the present study's students were able to view only the examiners' comments on the website. Similar to Harrison et al's findings, higher performing students were more engaging with feedback than lower performing ones: website log-in rates declined after the resit OSCE. This was not surprising because it is logical to assume that higher performing students are more motivated learners

who would engage with feedback or any learning opportunity more than unmotivated or lower performing students would.

Furthermore, this study produced original evidence showing that examiners produced more free text feedback for failed students, indicating their tendency to prioritise helping poor performing students correct their mistakes over helping competent students perfect their practice. As the purpose of summative assessments is to make sure the learners have achieved minimum acceptable competency, learning is directed mainly to pass the exam, rather than excel (Harrison, Konings et al. 2015). The present study's finding showed that the educators shared this mentality to prioritize achieving minimum acceptable competency over excellence. However, it is not appropriate to infer too much from this finding. More evidence should be attained on the examiners' feedback giving behaviour in a summative context.

This study also generated original evidence of feedback orientation in medical education. The researcher was able to collect data from a large number of students and generate strong psychometric evidence to support the four factor model of feedback orientation in medical education. Despite the variations, no particular group's feedback orientation was deficient enough to cause concern. It was not surprising to find more senior students and graduate entry to medicine students having stronger feedback orientation or males having more self-efficacy than females or first years having the least social awareness. Longitudinal studies would be useful to visualise how one's feedback orientation evolves over time and how interventions affect it.

Overall, the present study's feedback intervention was feasible, efficient and successful in attracting the educators and learners to engage with the feedback process. Examiner training should include specific information

regarding what constitutes good feedback in order to help examiners generate better quality feedback. Future research should explore the relationship between the learners' feedback orientation and their actual engagement with feedback. Knowing which aspects of feedback orientation the learners are weak at would help identify what type of support they need. Future research should focus on developing targeted interventions to improve each aspect of feedback orientation.

6.3 Reflection on limitations

The limitations to each study were presented in detail in each chapter. Most of these shortcomings could have been avoided with better planning.

All three studies were conducted by only one PhD student without peer collaboration at a single site using samples of convenience. An additional independent reviewer could have improved the quality appraisal of the studies and examiner comments in the literature review (Chapter 2) and the feedback intervention (Chapter 4) respectively. Involving a peer collaborator would have enabled testing inter-rater reliability of the quality appraisal tools and consequently reduced the risk of bias and error.

The interviews in Chapter 3 appeared to take a positivist approach because peer interviewers were given a detailed interview guide. Better training of the peer interviewers, keeping a reflective diary and having an additional researcher to independently analyse the data according to the established approaches to phenomenological approach could have improved the credibility of the findings.

Despite having a large sample, being a cross-sectional single site study weakened the generalisability of the findings from the FOS study (Chapter 5).

6.4 Conclusion

In conclusion, this study generated evidence on feedback after summative OSCEs and the construct of feedback orientation in medical education. The researcher found that summative feedback is a viable learning opportunity and awareness of the learners' feedback orientation enabled a better understanding of their engagement with feedback.

References

Ajzen, I. and M. Fishbein (1977). "Attitude-behavior relations: A theoretical analysis and review of empirical research." Psychological Bulletin **84**(5): 888.

Anders Ericsson, K. (2008). "Deliberate practice and acquisition of expert performance: a general overview." Academic emergency medicine **15**(11): 988-994.

Aoyama, M., Y. Tamura, Y. Ishikawa, M. Yada and I. Miyawaki (2013). "Confidence-weighted testing: A descriptive study of Japanese nursing students." Nursing & Health Sciences **15**(4): 504-509.

Babyak, M. A. and S. B. Green (2010). "Confirmatory factor analysis: an introduction for psychosomatic medicine researchers." Psychosomatic Medicine **72**(6): 587-597.

Benoit, C., M. Jansson, A. Millar and R. Phillips (2005). "Community-academic research on hard-to-reach populations: Benefits and challenges." Qualitative Health Research **15**(2): 263-282.

Berlin, N. and M.-P. Dargnies (2016). "Gender differences in reactions to feedback and willingness to compete." Journal of Economic Behavior & Organization **130**(Supplement C): 320-336.

Berliner, D. C. (2002). "Comment: Educational research: The hardest science of all." Educational Researcher **31**(8): 18-20.

Bierer, S. B., E. F. Dannefer, C. Taylor, P. Hall and A. L. Hull (2008). "Methods to assess students' acquisition, application and integration of basic science knowledge in an innovative competency-based curriculum." Medical Teacher **30**(7): e171-177.

Bing-You, R., V. Hayes, K. Varaklis, R. Trowbridge, H. Kemp and D. McKelvy (2017). "Feedback for learners in medical education: What is known? A scoping review." Academic Medicine **92**(9): 1346-1354.

Black, N. M. and R. M. Harden (1986). "Providing feedback to students on clinical skills by using the Objective Structured Clinical Examination." Medical Education **20**(1): 48-52.

Bord, S., R. Retezar, P. McCann and J. Jung (2015). "Development of an objective structured clinical examination for assessment of clinical skills in an emergency medicine clerkship." Western Journal of Emergency Medicine **16**(6): 866.

Boud, D. and E. Molloy (2013). "Rethinking models of feedback for learning: the challenge of design." Assessment & Evaluation in Higher Education **38**(6): 698-712.

Braddy, P. W., R. E. Sturm, L. E. Atwater, J. W. Smither and J. W. Fleenor (2013). "Validating the Feedback Orientation Scale in a Leadership Development Context." Group & Organization Management **38**(6): 690-716.

Brett, J. F. and L. E. Atwater (2001). "360° feedback: Accuracy, reactions, and perceptions of usefulness." Journal of Applied Psychology **86**(5): 930.

Bryman, A. (2016). Social research methods. Oxford, Oxford University Press.

Burdick, W. P., J. R. Boulet and K. E. LeBlanc (2018). "Can We Increase the Value and Decrease the Cost of Clinical Skills Assessment?" Academic Medicine **93**(5).

Burgess, A., T. Clark, R. Chapman and C. Mellis (2013). "Senior medical students as peer examiners in an OSCE." Medical Teacher **35**(1): 58-62.

Burgess, A. W., C. Roberts, K. I. Black and C. Mellis (2013). "Senior medical student perceived ability and experience in giving peer feedback in formative long case examinations." BMC Medical Education **13**: 79.

Chalmers, I., L. V. Hedges and H. Cooper (2002). "A brief history of research synthesis." Evaluation & The Health Professions **25**(1): 12-37.

Chang, A., C. L. Chou, A. Teherani and K. E. Hauer (2011). "Clinical skills-related learning goals of senior medical students after performance feedback." Medical Education **45**(9): 878-885.

Conrad, C. F. and R. C. Serlin (2005). The SAGE Handbook for Research in Education: Engaging Ideas and Enriching Inquiry. Thousand Oaks, CA, Sage.

Coulby, C., S. Hennessey, N. Davies and R. Fuller (2011). "The use of mobile technology for work-based assessment: the student experience." British Journal of Educational Technology **42**(2): 251-265.

Creswell, J. W., V. P. Clark and A. Garrett (2003). Advanced Mixed Methods Research Designs. Handbook of Mixed Methods in Social and Behavioural Research. Sage Publications. A. Tashakkori and C. Teddlie. Thousand Oaks, CA, Sage**:** 209-240.

Creswell, J. W. and C. N. Poth (2016). Qualitative inquiry and research design: Choosing among five approaches. London, Sage publications.

Daelmans, H. E., R. J. Hoogenboom, A. J. Donker, A. J. Scherpbier, C. D. Stehouwer and C. P. van der Vleuten (2004). "Effectiveness of clinical rotations as a learning environment for achieving competences." Medical Teacher **26**(4): 305-312.

Daelmans, H. E., R. J. Hoogenboom, A. J. Scherpbier, C. D. Stehouwer and C. P. van der Vleuten (2005). "Effects of an in-training assessment programme on supervision of and feedback on competencies in an undergraduate Internal Medicine clerkship." Medical Teacher **27**(2): 158-163.

Daelmans, H. E., R. M. Overmeer, H. H. van der Hem-Stokroos, A. J. Scherpbier, C. D. Stehouwer and C. P. van der Vleuten (2006). "In-training assessment: qualitative study of effects on supervision and feedback in an undergraduate clinical rotation." Medical Education **40**(1): 51-58.

Dannefer, E. F. and L. C. Henson (2007). "The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine." Academic Medicine **82**(5): 493-502.

Delman, J. and A. Lincoln (2009). Service users as paid researchers. Handbook of Service User Involvement in Mental Health Research. J. Wallcraft, B. Schrank and M. Amering. Singapore, Wiley-Blackwell**:** 139-151.

Denison, A., E. Bate and J. Thompson (2016). "Tablet versus paper marking in assessment: feedback matters." Perspectives on medical education **5**(2): 108-113.

Devotta, K., J. Woodhall-Melnik, C. Pedersen, A. Wendaferew, T. P. Dowbor, S. J. T. Guilcher, S. Hamilton-Wright, P. Ferentzy, S. W. Hwang and F. I. Matheson

(2016). "Enriching qualitative research by engaging peer interviewers: a case study." Qualitative Research **16**(6): 661-680.

Dochy, F., M. Segers and D. Gijbels (2007). Assessment engineering: Breaking down barriers between teaching and learning, and assessment. Rethinking Assessment in Higher Education, Routledge**:** 97-110.

Durning, S. J. and A. R. Artino (2011). "Situativity theory: A perspective on how participants and the environment can interact: AMEE Guide no. 52." Medical Teacher **33**(3): 188-199.

Eddles-Hirsch, K. (2015). "Phenomenology and educational research." International Journal of Advanced Research **3**(8).

Edwards, R. and C. Alexander (2011). Researching with peer/community researchers–ambivalences and tensions. The SAGE Handbook of Innovation in Social Research Methods. . London, Sage**:** 269-293.

Elliott, E., A. J. Watson and U. Harries (2002). "Harnessing expertise: involving peer interviewers in qualitative research with hard-to-reach populations." Health Expectations **5**(2): 172-178.

Ende, J. (1983). "Feedback in clinical medical education." Journal of American Medical Association **250**(6): 777-781.

Epstein, R. M. (2007). "Assessment in medical education." New England Journal of Medicine **2007**(356): 387-396.

Eva, K. W., H. Armson, E. Holmboe, J. Lockyer, E. Loney, K. Mann and J. Sargeant (2012). "Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes." Advances in Health Sciences Education **17**(1): 15-26.

Eva, K. W. and B. D. Hodges (2012). "Scylla or Charybdis? Can we navigate between objectification and judgement in assessment?" Medical Education **46**(9): 914-919.

Eva, K. W., J. Munoz, M. D. Hanson, A. Walsh and J. Wakefield (2010). "Which factors, personal or external, most influence students' generation of learning goals?" Academic Medicine **85**(10 Suppl): S102-105.

Evans, D. J., P. Zeun and R. A. Stanier (2014). "Motivating student learning using a formative assessment journey." Journal of Anatomy **224**(3): 296-303.

Fenigstein, A., M. F. Scheier and A. H. Buss (1975). "Public and private self-consciousness: Assessment and theory." Journal of Consulting and Clinical Psychology **43**(4): 522.

Ferenchick, G. S., D. Solomon, J. Foreback, B. Towfiq, K. Kavanaugh, L. Warbasse, J. Addison, F. Chames, A. Dandan and A. Mohmand (2013). "Mobile technology for the facilitation of direct observation and assessment of student performance." Teaching and Learning in Medicine **25**(4): 292-299.

Fernando, N., J. Cleland, H. McKenzie and K. Cassar (2008). "Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments." Medical Education **42**(1): 89-95.

Fowell, S. L. and J. G. Bligh (1998). "Recent developments in assessing medical students." Postgraduate Medical Journal **74**(867): 18.

Frankham, J. (2009). Partnership research: A review of approaches and challenges in conducting research in partnership with service users. ESRC National Centre for Research Methods Review Paper.

Fredricks, J. A., P. C. Blumenfeld and A. H. Paris (2004). "School engagement: Potential of the concept, state of the evidence." Review of Educational Research **74**(1): 59-109.

Gallacher, L.-A. and M. Gallagher (2008). "Methodological immaturity in childhood research? Thinking throughparticipatory methods'." Childhood **15**(4): 499-516.

Garry, A. and K. Stirling (2012). "Achieving 360 degrees student feedback using SPaCE." The Clinical Teacher **9**(4): 222-227.

Gibbs, G. (2006). How assessment frames student learning. Innovative Assessment in Higher Education. London and New York, Routledge**:** 43-56.

Grant, M. J. and A. Booth (2009). "A typology of reviews: an analysis of 14 review types and associated methodologies." Health Information & Libraries Journal **26**(2): 91-108.

Greco, M., A. Brownlea and J. McGovern (2001). "Impact of patient feedback on the interpersonal skills of general practice registrars: results of a longitudinal study." Medical Education **35**(8): 748-756.

Greene, S., A. Ahluwalia, J. Watson, R. Tucker, S. B. Rourke, J. Koornstra, M. Sobota, L. Monette and S. Byers (2009). "Between skepticism and empowerment: the experiences of peer research assistants in HIV/AIDS, housing and homelessness community-based research." International Journal of Social Research Methodology **12**(4): 361-373.

Haffling, A. C., A. Beckman and G. Edgren (2011). "Structured feedback to undergraduate medical students: 3 years' experience of an assessment tool." Medical Teacher **33**(7): e349-357.

Halaas, G. W., T. M. Zink, K. D. Brooks and J. Miller (2007). "Clinical Skills Day: Preparing Third Year Medical Students for the Rural Rotation." Rural and Remote Health **7**: 788.

Hall, A. T., D. D. Frink and M. R. Buckley (2017). "An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability." Journal of Organizational Behavior **38**(2): 204-224.

Handley, K., M. Price and J. Millar (2011). "Beyond 'doing time': investigating the concept of student engagement with feedback." Oxford Review of Education **37**(4): 543-560.

Harrison, C. (2019). "Can we redesign the MRCGP assessment to support lifelong learning?" Education for Primary Care **30**(1): 9-12.

Harrison, C. J., K. D. Könings, E. F. Dannefer, L. W. T. Schuwirth, V. Wass and C. P. M. van der Vleuten (2016). "Factors influencing students' receptivity to formative feedback emerging from different assessment cultures." Perspectives on Medical Education **5**(5): 276-284.

Harrison, C. J., K. D. Konings, A. Molyneux, L. W. Schuwirth, V. Wass and C. P. van der Vleuten (2013). "Web-based feedback after summative assessment: how do students engage?" Medical Education **47**(7): 734-744.

Harrison, C. J., K. D. Konings, L. Schuwirth, V. Wass and C. van der Vleuten (2015). "Barriers to the uptake and use of feedback in the context of summative assessment." Advances in Health Sciences Education Theory & Practice **20**(1): 229-245.

Harrison, C. J., K. D. Könings, L. W. T. Schuwirth, V. Wass and C. P. M. van der Vleuten (2017). "Changing the culture of assessment: the dominance of the summative assessment paradigm." BMC Medical Education **17**: 73.

Harrison, C. J., A. J. Molyneux, S. Blackwell and V. J. Wass (2015). "How we give personalised audio feedback after summative OSCEs." Medical Teacher **37**(4): 323-326.

Harvey, P., N. Radomski and D. O'Connor (2013). "Written feedback and continuity of learning in a geographically distributed medical education program." Medical Teacher **35**(12): 1009-1013.

Hattie, J. and H. Timperley (2007). "The power of feedback." Review of Educational Research **77**(1): 81-112.

Hauer, K. E., A. Teherani, D. M. Irby, K. M. Kerr and P. S. O'Sullivan (2008). "Approaches to medical student remediation after a comprehensive clinical skills examination." Medical Education **42**(1): 104-112.

Hauer, K. E., A. Teherani, K. M. Kerr, D. M. Irby and P. S. O'Sullivan (2009). "Consequences within medical schools for students with poor performance on a medical school standardized patient comprehensive assessment." Academic Medicine **84**(5): 663-668.

Hickey, G. L., S. W. Grant, J. Dunning and M. Siepe (2018). "Statistical primer: sample size and power calculations-why, when and how?" European Journal of Cardio-Thoracic Surgery **54**(1): 4-9.

Higgins, R., P. Hartley and A. Skelton (2001). "Getting the message across: the problem of communicating assessment feedback." Teaching in Higher Education **6**(2): 269-274.

Hill, D. A., A. I. Guinea and W. H. McCarthy (1994). "Formative assessment: a student perspective." Medical Education **28**(5): 394-399.

Hodder, R. V., R. N. Rivington, L. E. Calcutt and I. R. Hart (1989). "The effectiveness of immediate feedback during the objective structured clinical examination." Medical Education **23**(2): 184-188.

Hodges, B. (2013). "Assessment in the post-psychometric era: Learning to love the subjective and collective." Medical Teacher **35**(7): 564-568.

Hoefer, S. H., J. Sterz, B. Bender, M.-C. Stefanescu, M. Theis, F. Walcher, R. Sader and M. Ruesseler (2017). "Conveying practical clinical skills with the help of teaching associates—a randomised trial with focus on the long term learning retention." BMC Medical Education **17**(1): 65.

Hojat, M., M. J. Vergare, K. Maxwell, G. Brainard, S. K. Herrine, G. A. Isenberg, J. Veloski and J. S. Gonnella (2009). "The Devil is in the Third Year: A Longitudinal Study of Erosion of Empathy in Medical School." Academic Medicine **84**(9): 1182-1191.

Hoonpongsimanont, W., M. Feldman, N. Bove, P. K. Sahota, I. Velarde, C. L. Anderson and W. Wiechmann (2018). "Improving feedback by using first-person video during the emergency medicine clerkship." Advances in Medical Education and Practice **9**: 559.

Hooper, D., J. Coughlan and M. Mullen (2008). "Structural equation modelling: Guidelines for determining model fit." The Electronic Journal of Business Research Methods **6**(1): 53 - 60.

Howard, K. I. and G. A. Forehand (1962). "A method for correcting item-total correlations for the effect of relevant item inclusion." Educational and Psychological Measurement **22**(4): 731-735.

Hughes, J., W. J. Wilson, N. MacBean and A. E. Hill (2016). "Simulated patients versus seminars to train case history and feedback skills in audiology students: A randomized controlled trial." International Journal of Audiology **55**(12): 758-764.

Hulsman, R. L. and J. van der Vloodt (2015). "Self-evaluation and peer-feedback of medical students' communication skills using a web-based video annotation system. Exploring content and specificity." Patient Education & Counselling **98**(3): 356-363.

Kalish, R., M. Dawiskiba, Y. Sung and M. Blanco (2011). "Raising medical student awareness of compassionate care through reflection of annotated videotapes of clinical encounters." Education for Health **24**(3): 490.

Kanfer, R., M. Frese and R. E. Johnson (2017). "Motivation related to work: A century of progress." Journal of Applied Psychology **102**(3): 338.

Kataoka, H., N. Koide, K. Ochi, M. Hojat and J. Gonnella (2009). "Measurement of empathy among Japanese medical students: psychometrics and score differences by gender and level of medical education." Academic Medicine **84**.

Kelaher, M., D. J. Warr, P. Feldman and T. Tacticos (2010). "Living in 'Birdsville': Exploring the impact of neighbourhood stigma on health." Health & Place **16**(2): 381-388.

Kelly, E. and J. B. Richards (2019). "Medical education: giving feedback to doctors in training." BMJ **366**: l4523.

Kilpatrick, R., C. McCartan, S. McAlister and P. McKeown (2007). "'If I am brutally honest, research has never appealed to me…'The problems and successes of a peer research project." Educational Action Research **15**(3): 351-369.

Kirkpatrick, D. (1996). "Great Ideas Revisited. Techniques for Evaluating Training Programs. Revisiting Kirkpatrick's Four-Level Model." Training and Development **50**(1): 54-59.

Kneebone, R., F. Bello, D. Nestel, N. Mooney, A. Codling, F. Yadollahi, T. Tierney, D. Wilcockson and A. Darzi (2008). "Learner-centred feedback using remote assessment of clinical procedures." Medical teacher **30**(8): 795-801.

Kramer, M. (2009). Deciphering the future of learning through daily observation. 3rd WLE mobile learning symposium: mobile learning cultures across education, work and leisure, WLE Centre, IOE, London.

Laverty, S. M. (2003). "Hermeneutic Phenomenology and Phenomenology: A Comparison of Historical and Methodological Considerations." International Journal of Qualitative Methods **2**(3): 21-35.

Lee, W. S., J. Y. Hwang, J. E. Lim, S.-Y. Suh, K. H. Park and N.-J. Sung (2013). "The effect of videotaping students' interviews with patients for interview skill education." Korean Journal of Family Medicine **34**(2): 90.

Lefroy, J., A. Hawarden, S. P. Gay, R. K. McKinley and J. Cleland (2015). "Grades in formative workplace-based assessment: a study of what works for whom and why." Medical Education **49**(3): 307-320.

Levy, P. E., M. D. Albright, B. D. Cawley and J. R. Williams (1995). "Situational and Individual Determinants of Feedback Seeking: A Closer Look at the Process." Organizational Behavior and Human Decision Processes **62**(1): 23-37.

Lilford, N., A. Caruana and L. Pitt (2014). "Psychometric properties of the feedback orientation scale among South African salespersons." Psychological Reports **114**(1): 126-133.

Lind, J. (1753). A Treatise of the Scurvy. Edinburgh, Sands, Murray and Cochran.

Linderbaum, B. A. and P. E. Levy (2010). "The development and validation of the Feedback Orientation Scale (FOS)." Journal of Management **36**(6): 1372-1405.

Lipnevich, A. A. and J. K. Smith (2009). "Effects of differential feedback on students' examination performance." Journal of Experimental Psychology: Applied **15**(4): 319.

London, M. and J. W. Smither (2002). "Feedback orientation, feedback culture, and the longitudinal performance management process." Human Resource Management Review **12**(1): 81-100.

Mann, K., C. van der Vleuten, K. Eva, H. Armson, B. Chesluk, T. Dornan, E. Holmboe, J. Lockyer, E. Loney and J. Sargeant (2011). "Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict." Academic Medicine **86**(9): 1120-1127.

Maudsley, G. (2011). "Mixing it but not mixed-up: mixed methods research in medical education (a critical narrative review)." Medical Teacher **33**(2): e92-e104.

Mavis, B., J. Turner, K. Lovell and D. Wagner (2006). "Developments: Faculty, students, and actors as standardized patients: expanding opportunities for performance assessment." Teaching and Learning in Medicine **18**(2): 130-136.

McDonald, R. P. and H. W. Marsh (1990). "Choosing a multivariate model: Noncentrality and goodness of fit." Psychological Bulletin **107**(2): 247.

McLaughlin, K., L. Gregor, A. Jones and S. Coderre (2006). "Can standardized patients replace physicians as OSCE examiners?" BMC Medical Education **6**(1): 12.

McLeod, R., G. Mires and J. Ker (2012). "Direct observed procedural skills assessment in the undergraduate setting." The Clinical Teacher **9**(4): 228-232.

Merriam, S. B. and E. J. Tisdell (2015). Qualitative research: A guide to design and implementation. San Francisco, CA, John Wiley & Sons.

Moineau, G., B. Power, A. M. Pion, T. J. Wood and S. Humphrey-Murto (2011). "Comparison of student examiner to faculty examiner scoring and feedback in an OSCE." Medical Education **45**(2): 183-191.

Morrison, E. W. and R. J. Bies (1991). "Impression management in the feedback-seeking process: A literaturereview and research agenda." Academy of Management Review **16**(3): 522-541.

Moss, S. E., J. I. Sanchez, A. M. Brumbaugh and N. Borkowski (2009). "The mediating role of feedback avoidance behavior in the LMX—performance relationship." Group & Organization Management **34**(6): 645-664.

Moustakas, C. (1994). Phenomenological research methods. Thousand Oaks, California, Sage.

Newble, D. I. and K. Jaeger (1983). "The effect of assessments and examinations on the learning of medical students." Medical Education **17**(3): 165-171.

Niederle, M. and L. Vesterlund (2011). "Gender and Competition." Annual Review of Economics **3**(1): 601-630.

Nind, M. (2017). "The practical wisdom of inclusive research." Qualitative Research **17**(3): 278-288.

Norcini, J. (2010). "The power of feedback." Medical Education **44**(1): 16-17.

Norcini, J., B. Anderson, V. Bollela, V. Burch, M. J. Costa, R. Duvivier, R. Galbraith, R. Hays, A. Kent and V. Perrott (2011). "Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference." Medical Teacher **33**(3): 206-214.

Northcraft, G. B. and S. J. Ashford (1990). "The preservation of self in everyday life: The effects of performance expectations and feedback context on feedback inquiry." Organizational Behavior and Human Decision Processes **47**(1): 42-64.

Obe, R. E. and E. Hogard (2016). "Programmatic assessment: a paradigm shift in medical education." AISHE-J: The All Ireland Journal of Teaching and Learning in Higher Education **8**(3).

Papinczak, T., L. Young and M. Groves (2007). "Peer assessment in problem-based learning: a qualitative study." Advances in Health Sciences Education Theory & Practice **12**(2): 169-186.

Perron, N. J., M. Louis-Simonet, B. Cerutti, E. Pfarrwaller, J. Sommer and M. Nendaz (2016). "Feedback in formative OSCEs: comparison between direct observation and video-based formats." Medical Education Online **21**(1): 32160.

Perron, N. J., M. Louis-Simonet, B. Cerutti, E. Pfarrwaller, J. Sommer and M. Nendaz (2016). "The quality of feedback during formative OSCEs depends on the tutors' profile." BMC Medical Education **16**(1): 293.

Poulos, A. and M. J. Mahony (2008). "Effectiveness of feedback: The students' perspective." Assessment & Evaluation in Higher Education **33**(2): 143-154.

Rauf, A., M. S. Shamim, S. M. Aly, T. Chundrigar and S. N. Alam (2014). "Formative assessment in undergraduate medical education: concept, implementation and hurdles." Journal of Pakistani Medical Association **64**(1): 72-75.

Reiter, H. I., J. Rosenfeld, K. Nandagopal and K. W. Eva (2004). "Do clinical clerks provide candidates with adequate formative assessment during Objective Structured Clinical Examinations?" Advances in Health Sciences Education Theory & Practice **9**(3): 189-199.

Remple, V. P., C. Johnston, D. M. Patrick, M. W. Tyndall and A. M. Jolly (2007). "Conducting HIV/AIDS research with indoor commercial sex workers: Reaching a hidden population." Progress in Community Health Partnerships: Research, Education, and Action **1**(2): 161-168.

Ridder, J. M. M. V. D., K. M. Stokking, W. C. McGaghie and O. T. J. T. Cate (2008). "What is feedback in clinical education?" Medical Education **42**(2): 189-197.

Rushton, A. (2005). "Formative assessment: a key to deep learning?" Medical Teacher **27**(6): 509-513.

Sandars, J. (2015). "The challenge of feedback-insights from non-medical educational research." International Journal of Medical Education **6**: 1-3.

Sargeant, J., K. Mann, D. Sinclair, C. Van der Vleuten and J. Metsemakers (2008). "Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use." Advances in Health Sciences Education Theory & Practice **13**(3): 275-288.

Sargeant, J. M., K. V. Mann, C. P. van der Vleuten and J. F. Metsemakers (2009). "Reflection: a link between receiving and using assessment feedback." Advances in Health Sciences Education Theory & Practice **14**(3): 399-410.

Scarff, C. E., M. Bearman, N. Chiavaroli and S. Trumble (2019). "Trainees' perspectives of assessment messages: a narrative systematic review." Medical Education **53**(3): 221-233.

Schlegel, C., U. Woermann, J.-J. Rethans and C. van der Vleuten (2012). "Validity evidence and reliability of a simulated patient feedback instrument." BMC Medical Education **12**(1): 6.

Schulz, K. F. and D. A. Grimes (2005). "Sample size calculations in randomised trials: mandatory and mystical." The Lancet **365**(9467): 1348-1353.

Schuwirth, L., C. van der Vleuten and S. J. Durning (2017). "What programmatic assessment in medical education can learn from healthcare." Perspectives on Medical Education **6**(4): 211-215.

Schuwirth, L. W. T. and C. P. M. Van der Vleuten (2011). "Programmatic assessment: From assessment of learning to assessment for learning." Medical Teacher **33**(6): 478-485.

Sender-Liberman, A., M. Liberman, Y. Steinert, P. McLeod and S. Meterissian (2005). "Surgery residents and attending surgeons have different perceptions of feedback." Medical Teacher **27**(5): 470-472.

Setyonugroho, W., K. M. Kennedy and T. J. Kropmans (2015). "Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review." Patient Education & Counselling **98**(12): 1482-1491.

Sharma, N., Y. Cui, J. P. Leighton and J. S. White (2012). "Team-based assessment of medical students in a clinical clerkship is feasible and acceptable." Medical Teacher **34**(7): 555-561.

Shepard, L. A. (2000). "The Role of Assessment in a Learning Culture." Educational Researcher **29**(7): 4-14.

Shute, V. J. (2008). "Focus on formative feedback." Review of Educational Research **78**(1): 153-189.

Sinclair, H. K. and J. A. Cleland (2007). "Undergraduate medical students: who seeks formative feedback?" Medical Education **41**(6): 580-582.

Snodgrass, S. J., S. E. Ashby, L. Onyango, T. Russell and D. A. Rivett (2014). "Electronic practical skills assessments in the health professions: a review." Internet Journal of Allied Health Sciences and Practice **12**(1): 8.

Stone, A. (2014). "Online assessment: what influences students to engage with feedback?" The Clinical Teacher **11**(4): 284-289.

Suriowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations, London: Little, Brown**:** 65-71.

Sutton, P. (2012). "Conceptualizing feedback literacy: knowing, being, and acting." Innovations in Education and Teaching International **49**(1): 31-40.

Sweller, J., J. J. Van Merrienboer and F. G. Paas (1998). "Cognitive architecture and instructional design." Educational Psychology Review **10**(3): 251-296.

Tabachnick, B. G. and L. S. Fidell (2013). Using multivariate statistics. Boston, Pearson Education.

Tavakol, M. and R. Dennick (2011). "Making sense of Cronbach's alpha." International Journal of Medical Education **2**: 53-55.

Tavakol, M. and R. Dennick (2012). "Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66." Medical Teacher **34**(3): E161-E175.

Tavakol, S., R. Dennick and M. Tavakol (2011). "Psychometric properties and confirmatory factor analysis of the Jefferson Scale of Physician Empathy." BMC Medical Education **11**(1): 54.

Torre, D. M., D. E. Simpson, D. M. Elnicki, J. L. Sebastian and E. S. Holmboe (2007). "Feasibility, reliability and user satisfaction with a PDA-based mini-CEX to evaluate the clinical skills of third-year medical students." Teaching and Learning in Medicine **19**(3): 271-277.

Treadwell, I. (2006). "The usability of personal digital assistants (PDAs) for assessment of practical performance." Medical Education **40**(9): 855-861.

Turan, S., S. Uner and M. Elcin (2011). "The impact of standardized patient feedback on student motivational levels." Balkan Medical Journal **28**(1): 43-48.

Valle, R. S., M. King and S. Halling (1989). An introduction to existential-phenomenological thought in psychology. Existential-phenomenological perspectives in psychology. Boston, MA, Springer**:** 3-16.

van de Ridder, J. M., C. M. Peters, K. M. Stokking, J. A. de Ru and O. T. J. ten Cate (2015). "Framing of feedback impacts student's satisfaction, self-efficacy and performance." Advances in Health Sciences Education **20**(3): 803-816.

van der Vleuten, C. P. and S. Heeneman (2016). "On the issue of costs in programmatic assessment." Perspectives on Medical Education **5**(5): 303-307.

van der Vleuten, C. P. M., L. W. T. Schuwirth, E. W. Driessen, J. Dijkstra, D. Tigelaar, L. K. J. Baartman and J. van Tartwijk (2012). "A model for programmatic assessment fit for purpose." Medical Teacher **34**(3): 205-214.

Veloski, J., J. R. Boex, M. J. Grasberger, A. Evans and D. B. Wolfson (2006). "Systematic review of the literature on assessment, feedback and physicians' clinical performance: BEME Guide No. 7." Medical Teacher **28**(2): 117-128.

Vogel, J., D. Nel, L. Chapman and V. C. Burch (2012). "'The Game': student 'teaching'objective structured clinical examinations in South Africa." Medical Education **46**(5): 507-508.

Wass, V., C. Van der Vleuten, J. Shatzer and R. Jones (2001). "Assessment of clinical competence." The Lancet **357**(9260): 945-949.

Weaver, M. R. (2006). "Do students value feedback? Student perceptions of tutors' written responses." Assessment & Evaluation in Higher Education **31**(3): 379-394.

Whitehead, C. R., B. D. Hodges and Z. Austin (2013). "Dissecting the doctor: from character to characteristics in North American medical education." Advances in Health Sciences Education **18**(4): 687-699.

Wik, L., H. Myklebust, B. H. Auestad and P. A. Steen (2005). "Twelve-month retention of CPR skills with automatic correcting verbal feedback." Resuscitation **66**(1): 27-30.

Wilkinson, T. J. and M. J. Tweed (2018). "Deconstructing programmatic assessment." Advances in Medical Education and Practice **9**: 191-197.

Winston, K. A., C. P. Van Der Vleuten and A. J. Scherpbier (2010). "At-risk medical students: implications of students' voice for the theory and practice of remediation." Medical Education **44**(10): 1038-1047.

Winston, K. A., C. P. Van der Vleuten and A. J. Scherpbier (2010). "An investigation into the design and effectiveness of a mandatory cognitive skills programme for at-risk medical students." Medical Teacher **32**(3): 236-243.

Winstone, N. E., R. A. Nash, J. Rowntree and M. Parker (2016). "'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience." Studies in Higher Education: 1-16.

Withey, C. (2013). "Feedback engagement: forcing feed-forward amongst law students." The Law Teacher **47**(3): 319-344.

Young, I., K. Montgomery, P. Kearns, S. Hayward and E. Mellanby (2014). "The benefits of a peer-assisted mock OSCE." The Clinical Teacher **11**(3): 214-218.

Yudkowsky, R., S. Downing, D. Klamen, M. Valaski, B. Eulenberg and M. Popa (2004). "Assessing the head-to-toe physical examination skills of medical students." Medical Teacher **26**(5): 415-419.

Yudkowsky, R., J. Otaki, T. Lowenstein, J. Riddle, H. Nishigori and G. Bordage (2009). "A hypothesis-driven physical examination learning and assessment procedure for medical students: initial validity evidence." Medical Education **43**(8): 729-740.

Appendix

Sample feedback emails from students

Email 1

Hello,

We were recommended to contact you regarding feedback for the OSCE feedback we received.

Overall I think the feedback is useful. I've included my results below
So you will understand this point- for station 3 I received a fair amount of free text feedback, which is unusual compared to all the other stations. I believe that one of the points in the paragraph (that I did not ask the patient if I could talk to their relative is not accurate. My recollection is that I did. I hope you will understand that I am not making a complaint about this, or suggesting the examination was not fair but I do feel that feedback in this station would have been more useful and constructive if it had been given in the same way as the other stations (using the A-B-C approach). Whilst the examiner may have had time to write the paragraph immediately after my examination and before the next person it may also be the case that it could potentially lead to error in giving the correct feedback.

I hope you understand that I am not querying this small point- I just think consistency in the giving of feedback is less likely to result in feedback which may conflict with what the student thinks happened in the station.

Otherwise I have found the feedback helpful and reassuring.

Best wishes

Email 2

Hello,

I was just wondering about the table with our domain percentages at the end of the feedback. It seems strange to me that in OSCE 2 I can be above the class average in all but one domain yet score 54% which I presume to be lower than the class average overall?
I'm sure that it must be possible but just doesn't seem very logical?

Thanks,

Ethics Approvals

Figure 22 Ethic Approval 1 Top

The University of Nottingham

Faculty of Medicine and Health Sciences

Research Ethics Committee
School of Medicine Education Centre
B Floor, Medical School
Queen's Medical Centre Campus
Nottingham University Hospitals
Nottingham
NG7 2UH

Direct line/e-mail
+44 (0) 115 8232561
Louise.Sabir@nottingham.ac.uk

18th August 2015

Dr Myo Min Soe
PhD Research Student
c/o Professor Gillian Doody
Professor in Adult Psychiatry and Medical Education
Head of Assessment Unit
School of Medicine Education Centre
Medical School
QMC Campus
Nottingham University Hospitals
NG7 2UH

Dear Dr Myo Min Soe

Ethics Reference No: R14072015 SoMEC AU PhD/BMedSci – please always quote
Study Title: Exploring exam experience. A qualitative study with University of Nottingham Medical Students.
Chief Investigator/Supervisors: Professor Gillian Doody,Professor in General Adult Psychiatry and Medical Education, Head of Assessment Unit, Dr Mohsen Tavakol, Assistant Professor in Psychometrics, School of Medicine Education Centre.
Lead Investigator/Student: Dr Myo Min Soe, First year PhD Student
Interviewer/Transcribers: Dr Robert Cullum, Foundation Year one Doctor, Nottingham University Hospitals NHS Trust, Louisa Hepworth CP1 Medical Student, Thomas Hewson, 3rd Year Medical Student, University of Nottingham.
Duration of Study: 08/2015-11/2015 3mths **No of Subjects:** 30

Thank you for your letter dated 14th August 2015 responding to the comments made by the Committee and the following revised documents were received:

Exploring Exam Experience:

- Study I: Participant Information Sheet Final version 1.0, 14.08.2015

Figure 23 Ethic Approval 1 Bottom

Exploring Exam Experience:

- Study I: Participant Information Sheet Final version 1.0, 14.08.2015
- Study I: Consent Form Final version 1.0, 14.08.2015
- FMHS Research Ethics Committee Application Form
- Short Study Protocol
- Recruitment E-mail
- Information Sheet for participants
- Interview Guide

These have been reviewed and are satisfactory and the study is approved.

Approval is given on the understanding that the Conditions of Approval set out below are followed.

1. You must follow the protocol agreed and inform the Committee of any changes using a notification of amendment form (please request a form).

The University of Nottingham

2. You must notify the Chair of any serious or unexpected event.

3. This study is approved for the period of active recruitment requested. The Committee also provides a further 5 year approval for any necessary work to be performed on the study which may arise in the process of publication and peer review.

4. An End of Project Progress Report is completed and returned when the study has finished (Please request a form).

Yours sincerely

pp Lamagabni

Professor Ravi Mahajan
Chair, Faculty of Medicine & Health Sciences Research Ethics Committee

Figure 24 Ethics Approval 2

University of Nottingham
UK | CHINA | MALAYSIA

Email: FMHS-ResearchEthics@nottingham.ac.uk

Faculty of Medicine &Health Sciences
Research Ethics Committee
c/o Faculty PVC Office
School of Medicine Education Centre
B Floor, Medical School
Queen's Medical Centre Campus
Nottingham University Hospitals
Nottingham, NG7 2UH

4 July 2017

Dr Myo Min Soe
Final Year PhD Student
c/o Professor Gillian Doody
Dean of Medical Education
School of Medicine Education Centre
B Floor, School of Medicine Admin Hub
QMC Campus
Nottingham University Hospitals
NG7 2UH

Dear Dr Myo Min Soe

Ethics Reference No: EV200317 – please always quote	
Study Title: Developing Engagement with Feedback Workshop and Evaluation.	
Chief Investigator/Supervisor: Professor Gillian Doody, Dean of Medical Education, Medical Education Centre.	
Lead Investigators/student: Dr Myo Min Soe, Final Year PhD student, Medical Education Centre.	
Other Key Supervisors/Investigators: Dr Rakesh Patel, Clinical Associate Professor in Medical Education, Dr Mohsen Tavakol, Assistant Professor in Psychometrics, Dr Naomi Winstone, Associate Dean (Learning and Teaching), University of Surrey.	
Type of Study: PhD project, developmental, educational intervention/quantitative evaluation.	
Proposed Start Date: 01/04/17	**Proposed End Date:** 01/07/17 3mths
No of Subjects: 90+	**Age:** 18+years
School: Medicine	

Thank you for notifying the Committee of the following Amendments:

- No 1: 07.04.17 – new recruitment strategy – noted on 18.04.17
- No 2: 27.04.17 – increase inconvenience allowance amount- noted on 28.4.17
- No: 3: 03.07.17 – Addition of new researcher to the team.

and the following documents were received:

- New data collection/recruitment strategy outline
- Poster

These have been reviewed and are satisfactory and the study amendments have been noted and given a favourable opinion.

Yours sincerely

pp

Professor Ravi Mahajan
Chair, Faculty of Medicine & Health Sciences Research Ethics Committee