Assessor Cognition and Inter-Rater Reliability in Nursing Objective Structured Clinical Examinations

Conor Scully

B.A. (Hons), M.Sc. (*cum laude*)

Dissertation submitted to Dublin City University in fulfilment of the requirements for the award Doctor of Philosophy

Supervisors: Prof. Michael O'Leary, Dr. Mary Kelly, Dr. Zita Lysaght

Dublin City University

School of Policy and Practice

January 2023

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID No: 20210001

Date: 04/1/2023

Table of Contents

Chapter One: Introduction

Chapter Two: Literature Review

Chapter Three: Methodology

Chapter Four: Results

Chapter Five: Discussion

List of Tables

List of Figures

List of Acronyms and Abbreviations

AERA	American Educational Research Association
ANOVA	Analysis of Variance
BP	Blood Pressure
DCU	Dublin City University
ICC	Intraclass Correlation Coefficient
IRR	Inter-Rater Reliability
Mini-CEX	Mini-Clinical Evaluation Exercise
NG	Naso-Gastric Tube Insertion
OSCE	Objective Structured Clinical Examination
TA	Thematic Analysis

Acknowledgements

I would like to thank Michael O'Leary for providing me with a frankly inordinate amount of support over the last three years. His advice (both personal and professional) has been invaluable as I have gone on my PhD "journey". I wish him all the best for his retirement. Thanks to Mary Kelly for her continued support and guidance throughout the project, in particular for helping to advance the work at several crucial moments. I am deeply grateful to Zita Lysaght, who went above and beyond on numerous occasions in order to help get this thesis over the line, and whose academic rigour is matched by her good nature.

Undertaking this thesis at CARPE has been one of the most rewarding experiences of my life, and I am lucky to have worked with a series of outstanding colleagues (and friends). Thanks to Anastasios, Vasiliki, Deirdre, Gemma, Eadaoin and Fainche. Thanks in particular to Paula Lehane, who gave me so much support since I strolled into our shared office in 2019 and declared that I "didn't think the PhD was going to be stressful". I was wrong, but I would have been much more wrong without her to help me through various crises. Thanks also to Darina Scully, colleague (and sister-in-law) *par excellence*, who was always available to chat or read over my work even though she was raising two beautiful children, Cian and Lily.

I would like to acknowledge the various people, within and outside DCU, who helped me in ways big and small: Frankie Nesirky, Maeve Power, Elaine McDonald, Cillian Murphy, Alicia Menendez Tarrazo, Marian Traynor, Patrick Doyle, Evelyn Kelleher, Angela Mazzone and Jonny Hobson. Special thanks to Melissa Corbally, whose rigorous scholarship during my progress viva in June 2021 highlighted several issues with the project and helped to make it much better.

I am grateful to have received financial support from Prometric during the last three years. Thanks to all the colleagues I interacted with over there, in particular Dr. Linda Waters and Dr. Li-Ann Kuan. I hope to work with you both again in the future.

All my love to my wonderful friends who enrich my life every day, particularly the "Assistance & Technology" crew: Aisling, Elizabeth, Huda, Ludivine, Matthew, Matthew, Orla and Rachael. Special shout out to Cian Henry, whose friendship has been a constant source of support over the last decade. Love also to Neil Delaney, who was by my side for the majority of time I was doing the PhD: I wish him all the best.

Finally, I would like to thank my family for all they have done for me over the last 28 years. My brothers, Peter and Brian, have never failed to help me in any way they could; while my parents, Aedín and Tony, have provided me with every opportunity I could have wished for (and many I could not). Love to you all.

Conor Scully

Dublin, December 2022

Abstract

Researcher Name: Conor Scully

Thesis Title: Assessor Cognition and Inter-Rater Reliability in Nursing Objective Structured Clinical Examinations

The consistency of judgements made by examiners of performance assessments is an important issue when high stakes are associated with the outcomes of such assessments for examinees. In order to minimize variance between assessors, it is imperative that designers and users of assessments understand and account for variations that may arise when different assessors observe the same performance.

Objective Structured Clinical Examinations (OSCEs) are high-fidelity performance assessments common in the health sciences, which require that students are judged by a range of different assessors. Despite the current prominence of OSCEs within undergraduate nursing programs, two problematic issues are highlighted in the research literature: relatively little is known about the specific cognitive processes that assessors employ when reaching judgements about the students they observe; and inter-rater reliability can be low.

This mixed-methods study sought to address both issues using a combination of semi-structured interviews and a *think-aloud protocol*, in which assessors (*n*=12) shared their thought processes with the researcher as they reviewed four videos of students completing two OSCEs: blood pressure measurement and naso-gastric tube insertion. Participants also completed the associated marking guides for each OSCE, the data from which were used to determine the percent agreement between assessors (inter-rater reliability) of the viewed student performances.

The results of the study indicated idiosyncrasy in the cognitive processes that assessors employed while judging the recorded performances. The data suggested that although each assessor watched the same four videos, they had different methods of determining how well or badly the students performed. Perhaps unsurprisingly, the completed marking guides revealed substantial variance in the scores the assessors awarded, with the harshest assessor awarding 29/52 checklist items across the videos compared to 45/52 for the most lenient assessor. Notably, there were discrepancies at the pass/fail decision for three out of the four performances.

Chapter One

Introduction

1.1 *Introduction*

This thesis examined the inter-rater reliability (IRR) of undergraduate nursing Objective Structured Clinical Examinations (OSCEs), focusing specifically on how assessor cognition could be used to identify and account for threats to IRR. The importance of rigorous, well-designed assessments as part of undergraduate curricula has long been recognised by researchers working in nurse assessment (Rushforth, 2007) as well as the health sciences more broadly (Khan et al., 2013a). Fair and reliable assessments, which provide valid inferences about student performance levels, are crucial if healthcare practitioners are to be well-prepared for the world of clinical practice, with the ultimate aim of improving patient outcomes.

The OSCE is a performance assessment which requires a student to complete a “station” (or series of stations), at which they have a fixed amount of time to perform a specific task, for example the measurement of a patient’s blood pressure. They are judged at each station by a trained examiner, who grades them on the basis of a marking guide that has been designed for that station. A student’s scores at each station are aggregated to form their overall OSCE score (Khan et al., 2013a). Because OSCEs usually require a pool of assessors who examine and allocate marks to students, a key consideration is ensuring that different assessors[1] judge students in ways that are consistent and comparable, such that a student receives the same score regardless of who is assessing them (Gingerich et al., 2014a). However, research into undergraduate nursing OSCEs in particular has indicated that there may be issues regarding this consistency and comparability, and authors writing in this field have urged further investigation (Goh et al., 2019).

In recent years, an emerging body of research known as assessor cognition has provided a novel method for understanding how issues with inter-rater reliability arise. Assessor cognition explores “how assessors/examiners observe and make judgements about students’ and trainees’ clinical and professional skills” (Boursicot et al., 2021, p.59). By de-privatising the cognitive processes that assessors engage in when judging a student’s performance, researchers can identify the sources of divergence between assessors.

[1] In the literature, the terms “assessor” and “rater” are used interchangeably. For the sake of consistency, the term “assessor” is generally used throughout this thesis, unless the term “rater” appears in a direct quotation.

1.2 *Organisation of the chapter*

This chapter outlines the topic explored in this thesis, providing information about nursing OSCEs, concerns regarding the consistency of assessor judgements, and assessor cognition. Following this, the specific aim of the thesis is articulated, namely the use of assessor cognition as a means of understanding how variance in awarded scores arises. The chapter concludes with a discussion of the scope of the thesis and the context in which the study took place.

1.3 *Research Topic and Problem*

1.3.1 *Potential advantages of OSCE format*

There are numerous assessment modalities that can be used in the assessment of undergraduate nursing students. Common assessments include "written exams, assignments and projects", which allow students to demonstrate their theoretical knowledge (Rushforth, 2007, p.483). However, these assessments do not afford students the opportunity to showcase the practical skills which are central to nursing practice. Such skills can only be assessed through a performance assessment, which require students to "construct an answer, produce a product, or perform an activity" (Darling-Hammond & Adamson, 2010, p.7). Performance assessments are distinguished by their fidelity to the real world: a well-designed performance assessment simulates a situation that a student is likely to encounter while working in a clinic or hospital, and as such is considered to be an effective predictor of future success (Rushforth, 2007). As a performance assessment, the OSCE was designed in order to combine the standardisation present in theoretical assessments with complexity and real-world fidelity.

The OSCE was first described in the field of medicine by Harden et al. (1975), and was developed due to concerns with the consistency of awarded scores in other performance assessment methods that were popular at the time (Khan et al., 2013a, p.1439). Before the advent of the OSCE, the dominant performance assessments used to test medical students were short or long case examinations. In a short case examination, a student undertakes a brief clinical examination of several patients, before discussing their findings with a pair of assessors, who award their grade (Khan et al., 2013a). In a long case, a student takes a history and conducts a complete physical examination of one patient, which is followed by unstructured questioning from one or more assessors (Khan et al., 2013a). As Khan et al. outlined, a student's score in either case could be affected by "the patient's performance, examiner bias, a non-standardised marking scheme, and the candidate's actual performance" (p.1439). The OSCE was designed to mitigate the effect of the first three factors on a student's grade, to ensure that a student's performance on the exam is the only determinant of their score.

Accordingly, there are several design features of the OSCE which have the aim of bringing about this consistency. Firstly, the OSCE is structured such that every student who takes the exam has to complete the same series of stations: students enter into the exam hall, begin with the first station, and rotate through the stations until every student has completed every station (Khan et al., 2013a). Secondly, within each station, a student is examined by an assessor who is obliged to award them a grade using that station's marking guide, which reduces the scope for assessors to judge students on the basis of what they personally believe to be important. Thirdly, the use of Standardised Patients (SPs), actors who are trained to present symptoms and interact with exam participants in a routine way, was introduced to ensure that the procedure that students have to perform within a given station is the same every time. Thus, by design, the intention of the OSCE is that every student who completes the exam faces the same problem (or series of problems), and is graded on their performance in the same way.

The original OSCE (Harden et al., 1975) was designed to assess the achievement levels of undergraduate medical students. Subsequent to this, OSCEs proliferated across universities in the Western world, becoming especially popular in the United States and Canada (Rushforth, 2007). OSCEs are now performed in over 50 countries worldwide (Patricio et al., 2013). OSCEs have proven to be a durable assessment format beyond undergraduate level and are now used at postgraduate level, and for certification and accreditation exams for various medical specialties (e.g., Scrimgeour et al., 2019). OSCEs have been popularised beyond medicine, and are now administered in many different medical fields, such as psychiatry, midwifery, and mental health practice (Selim et al., 2012; Smith et al., 2012; Stockmann et al., 2019).

Driven by perceived deficiencies in classic nursing assessments, such as written examinations or portfolio assessments, OSCEs have become especially popular in nurse education, (Rushforth, 2007). In a highly-cited literature review from 2007, Rushforth wrote that the OSCE format had become common in nursing, as it was thought to provide a way of assessing practical skills in a way that maintains the advantages of scoring consistency present in non-practical assessments such as multiple-choice examinations. When administered in nursing, alterations are often made to the OSCE to make it more representative of nursing practice. Common alterations include having fewer stations which take longer to complete - up to 30 minutes - and a less strict delineation of different skills (Rushforth, 2007). Indeed, published work describing the development of nursing OSCEs has often detailed a single-station approach (such as the adapted Objective Structured Clinical Assessment (OSCA) outlined by Najjar et al. (2016)). Such an approach allows for the assessment of multiple, integrated skills in the one station. However, it reduces the number of assessors that each student is graded by, which increases the likelihood that a student's score may be disproportionately impacted by the tendency of certain assessors to grade more harshly or leniently[2] than their counterparts.

1.3.2 *Concerns over scoring consistency*

One of the primary aims of the OSCE is to ensure high levels of score reliability, particularly inter-rater reliability (IRR), which refers to the agreement between assessors on a test in which there is a person (or a series of people) judging a test-taker's performance. In theory, when someone takes a test, their performance level should be the only determinant of their score. As noted by Gwet (2014) in his *Handbook of Inter-Rater Reliability*: "When used by different examiners, a reliable psychometric test is expected to produce the same categorisation of the same human subjects" (p.6). As long as the assessors are adequately prepared, it should not make a meaningful difference who they are – they should complete the marking guide in the same way every time. However, as McHugh noted in a highly-cited piece discussing IRR, "multiple people collecting data may experience and interpret the same phenomena differently" (2012, p.276). When this happens, assessors may award different scores to the same student performance, a phenomenon that would typically be classed as measurement error.

The assumption that a well-designed OSCE will lead to high levels of IRR is not necessarily borne out in evidence. In fact, persistent problems with IRR levels in performance assessments such as the OSCE are well documented. In an instructional piece advising assessment developers how to design and administer OSCEs, Khan et al. (2013b) wrote that "it is essential to perform psychometric analysis on OSCE results and use the outcomes to enhance the quality of the examination" (p.1458). To that end, an evaluation of reliability is often a central component in academic reporting of newly developed OSCEs (e.g., Lee et al., 2020). Recent literature on the subject of OSCE reliability has determined that:

[2] Throughout this thesis, the terms "harsh" and "lenient" are used to describe assessors who systematically award lower or higher grades than their counterparts. The use of these terms is not meant to imply judgement on the part of the researcher as to the defensibility of these awarded grades.

- In medicine OSCEs, awarded scores are "often not very reliable", and are more likely to be classified as "moderate" rather than "strong" (Brannick et al., 2011, p.1181), indicating that there is scope to investigate how reliability could be further improved.
- In nursing OSCEs, data on reliability (and IRR in particular) are often absent from published studies detailing OSCE development (Goh et al., 2019). As such, it is possible that there are systemic issues with IRR that are not being identified.
- When reliability data are published in relation to nursing OSCEs, IRR is at a similar level to what is found in medicine OSCEs; when calculated using common measures such as the Intraclass Correlation Coefficient (ICC) or Spearman's *r*, it is frequently classed as "moderate", and occasionally "poor" (Navas-Ferrer et al., 2017). This suggests that designers of nursing OSCEs have yet to ascertain how to optimise IRR levels.

As such, existing research suggests the need for further investigation into the IRR levels of nursing OSCEs, to determine ways that IRR can be improved. Ultimately, IRR has implications for the validity of the inferences made on the basis of awarded scores (Thompson & Vacha-Haase, 2017). Validity is the core principle underpinning all assessments, and is defined as "the degree to which the evidence and theory support interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p.11). When administering an assessment such as an OSCE, it is imperative that validity evidence is collected so that score inferences are deemed to be valid - this process is known as validation (AERA et al., 2014). One of the key steps in documenting validity evidence is a calculation of score reliability, generally thought to be a necessary condition for valid decisions (Fraenkel & Wallen, 2006). As such, the collection of reliability data such as IRR is an important step in building a validity argument for an assessment. This is especially true when OSCE scores are used to make high-stakes decisions, such as whether a student has demonstrated enough mastery of the curriculum to be allowed to progress to the next year of study. Such decisions cannot be considered defensible unless there is evidence that awarded scores are not affected by idiosyncrasies in assessor judgements (AERA et al., 2014).

There is a substantial body of research which has the aim of determining how score reliability, and IRR specifically, can be improved. Different authors have investigated IRR through various lenses, for example by exploring the design of the marking guide (Setyonugroho et al., 2015), whether the type of skills assessed in an OSCE has implications for score variance (Comert et al., 2016), and the structure and layout of the OSCE overall (Navas-Ferrer et al., 2017). However, the persistence of low levels of IRR has been well-documented in undergraduate nursing OSCEs in spite of these attempts to account for it (e.g., Cazzell & Howe, 2012; Dunbar, 2018). In other words, there is evidence in the literature that inconsistencies in how assessors judge student performances may arise regardless of the design of the OSCE.

Within the field of undergraduate nursing assessments, the majority of published work on IRR has adopted an exclusively quantitative approach. This allows for a determination of whether particular assessors are notably harsh or lenient in terms of their awarded grades compared to their counterparts (e.g., Dunbar, 2018), or whether specific items on the marking guide are causing disagreements between assessors (Cazzell & Howe, 2012). However, while this information is important to ascertain, and can help assessment designers develop solutions to improve IRR, it does not allow for a deep understanding of how specifically variance between assessors arises. For example, it might be the case that the language of the marking guide is not aligned with how assessors view student performances, or that assessors do not share a common understanding of what a "good" performance looks like (Boursicot et al., 2021). Without this deep understanding, efforts to improve IRR are necessarily limited. An alternative potential approach to this issue is researchers engaging with assessors in order to investigate what specifically happens when they judge student performances in assessments such as the OSCE. Such an approach is rare in research on undergraduate nursing assessments, and was the central purpose of this thesis.

1.3.3 *Assessor cognition*

One promising method of investigating IRR, by centring assessors themselves, is an emerging body of work known as assessor cognition. The focus of assessor cognition is understanding "assessors' cognitive processes and their impact on assessment quality" (Gingerich et al., 2014a, p.1056). The underlying assumption of assessor cognition research is that because people are different, it is likely that assessors do not judge performances in exactly the same way (Yeates et al., 2013; Gingerich et al., 2014a). In a piece commenting on this emerging field, Govaerts and van der Vleuten (2013) noted that assessors have increasingly been thought of as "active information processors who interpret and construct their own personal reality of the assessment context" (p.1169). Seen this way, assessors are not merely blank slates who view and grade a student's performance on an OSCE or similar exam in the same way as each other, rather they bring with them individual perspectives which affect their interpretation of a performance (Govaerts & van der Vleuten, 2013). This perspective is one that is central to much research in the social sciences: in contrast to areas such as science, in which complete control and standardisation is possible, the innate complexity of humans' cognitive processing may be inherently incongruous with the goal of complete agreement (Gingerich et al., 2014a).

The potential utility of assessor cognition as a means of exploring IRR has been noted by influential researchers in the area. Indeed, assessor cognition developed in part as a response to persistent reliability issues with performance assessments such as the OSCE. Commenting on this development in 2014, Gingerich et al. noted that "when psychometrics are used to analyse performance assessments, often a greater amount of variance in ratings can be accounted for by the assessors…than the trainees" (2014a, p.1056). In other words, differences in how assessors understand and mark student performances can sometimes affect awarded grades more than actual differences in ability levels between students. This highlights the importance of centring the assessor in IRR analyses, in order to understand how their engagement with the task of judging and grading students leads to divergence in the scores they award.

Researchers such as Kogan et al. (2011) and Gauthier et al. (2016) have devised initial frameworks for mapping out the specifics of what happens when an assessor judges and grades a student completing a performance assessment such as the OSCE. As may be expected due to the aforementioned complexity of human cognition, these frameworks have suggested that there are a range of factors, unique to an assessor, that may affect how a single performance is interpreted. While these factors are often common cognitive processes that help individuals make sense of the world around them, they are notable from an assessment perspective, as they have the potential to threaten the IRR of awarded scores. Initial assessor cognition research, usually conducted in the field of medical assessment, has provided novel insights into assessors' cognitive processes. Thus far, studies have indicated that:

- Assessors' observations (Gauthier et al., 2016) of students are informed by the marking guide. However, different assessors may choose to focus on different aspects of a performance: some assessors might value communication skills and direct their observations in this way, while others might value psychomotor skills (Yeates et al., 2013). Assessors are also prone to making inferences about students that go beyond what is directly visible to them in the performance (Kogan et al., 2011).
- Assessors' interpretations (Gauthier et al., 2016) of students are affected by the fact that assessors may lack a fixed set of criteria against which to judge students, and thus

compare a specific performance to what they judged directly before (Yeates et al., 2015), or to what they would do themselves in a specific situation (Roberts et al., 2020).

- Assessors may also be prone to using a range of subjective criteria, outside of what is contained within the marking guide, to make sense of a particular performance. As such, there is a potential for misalignment between the marking guide and assessors (Hyde et al., 2020), and assessors may choose to openly defy what is in the guide and judge a student based on what they perceive to be important.
- While it has been shown that assessors do have different ways of understanding a student's performance, it is unlikely that there are an infinite number of ways of doing this; rather, there are likely to be identifiable patterns of assessor judgements for any one student performance (Gingerich et al; 2014b). In other words, there may be a notable divide within a specific group of assessors regarding how a performance is judged and graded.
- Where assessor variance exists, it is possible that some of it is due to error on the part of assessors, and some of it is due to assessors having different, but equally defensible, interpretations of student performances (Gingerich et al., 2014a; 2017).

The emergent nature of assessor cognition research means that there are numerous opportunities for further research in the area. Firstly, researchers have emphasised the small scale nature of studies conducted thus far, and have urged others to try to replicate findings across contexts in order to determine generalizability (e.g., Chahine et al., 2016, p.620). Secondly, while assessor cognition research is assumed to be highly relevant to OSCEs, much published work has been conducted using other medical assessment formats, such as the Mini-Clinical Evaluation Exercise (e.g., Gingerich et al., 2014b). Thirdly, little published work on assessor cognition has taken place in the field of nursing specifically (for an exception see East et al. 2014), with the majority of studies situated in the field of medicine. Finally, while there is a substantial body of literature on OSCE design and score reliability, and the relationship between the two (e.g., Setyonugroho et al., 2015); it is uncommon for these issues to be addressed through the lens of assessor cognition. Overall, from the observations made here, which are elaborated further in the following chapter, it is argued that while investigating assessors' cognitive processes may be a useful method of exploring IRR issues, this has yet to be attempted within the field of undergraduate nursing OSCEs.

1.4 *Aim of the research*

The research problem was conceptualised in two parts: the ongoing need to investigate IRR levels within undergraduate nursing OSCEs, and the potential for the field of assessor cognition to provide a novel and fruitful method of understanding how variance between assessors arises.

In light of the issues discussed above, this thesis sought to explore:

- The cognitive processes that nursing assessors go through when grading OSCE performances, and the extent to which assessors of these OSCEs employ subjective or idiosyncratic criteria which have the potential to lead to unwanted score variance.
- The inter-rater reliability of undergraduate nursing OSCEs, focusing in particular on whether certain assessors are systematically harsher or more lenient than their counterparts, and the extent to which specific items within OSCE marking guides are prone to resulting in discrepancies between assessors.
- The links between assessors' cognitive processes and variance in the scores they award. It was expected that collecting and analysing qualitative and quantitative data from assessors would allow for an in-depth understanding of specifically how variance between assessors arises. This information could then be used to determine potential ways that the design of the OSCE could take assessors into account, with the ultimate goal of increasing IRR.

In order to investigate these issues, a mixed-methods, case study approach was employed (Johnson & Onwuegbuzie, 2004; Cope, 2015). The case studied was a School of Nursing (referred to as "the School" for the rest of the thesis) at a university in Ireland, which granted the researcher access to carry out the study. The School offers a range of taught nursing programmes, and uses OSCEs as a method of summative assessment in at least six undergraduate modules. Module coordinators within the School indicated that they had a lack of knowledge as to whether there were issues with assessor consistency. In the absence of a prior investigation, the case represented an opportunity to explore in detail if such issues were present and, if there were, how they arise and how they could be mitigated. The use of a case study approach, therefore, allowed the researcher to carry out a "detailed study" (Cope, 2015, p.681) of assessor consistency within the School, in accordance with the deep understanding that is the aim of assessor cognition research. Two OSCEs, blood pressure (BP) measurement and naso-gastric tube insertion (NG), were used as the object of study, and a total of 12 assessors from within the School participated.

While the focus of this thesis was on assessor cognition and its links with IRR, there were two additional areas of research that informed the study. Firstly, in recent years there has been an emerging discourse in both medical and nursing education that calls into question the narrow focus on reliability that is at the heart of large-scale standardised assessments such as the OSCE. Authors such as Hodges (2013) have argued that concerns over “objectivity” have come to dominate assessments in the last two decades, such that subjectivity has become equated with bias and unfairness. As a result, psychometric evaluations of OSCEs are often exclusively focused on reliability, trying to ensure that assessors all judge performances in a homogenous way. However, this approach fails to recognise that OSCE assessors are often experienced clinical practitioners themselves, and may be able to use their “expert judgement” to appraise students in a way that is not necessarily aligned with the marking guide (or with other assessors) (Eva & Hodges, 2012). Ultimately, such an approach might pose a threat to validity, as talented students may fail to be identified and graded accordingly. This research, while focusing primarily on reliability, was nonetheless informed by debates around the prevalence and utility of subjective judgements within performance assessments. These debates are unpacked and discussed further in section 2.4.4 of the following chapter.

Additionally, the exploration of assessors’ cognitive processes, and the potential effects on IRR levels, necessitated the recording of videos of student OSCE performances. As a result, the thesis has implications regarding the use of video in performance assessments such as the OSCE. In recent years - and particularly since the onset of the COVID-19 pandemic in March 2020 - video has been used to allow educators to assess students’ skills remotely, either by conducting an adapted OSCE through Zoom (e.g., Blythe et al., 2021), or by allowing students to film themselves completing a skill in their own home and uploading the video to an online platform for grading by an assessor (Purpora & Prion, 2018). The potential advantages of the online format in terms of formative feedback, where students can re-watch the video to gain a better understanding of how they could improve, have been noted by researchers (Lewis et al., 2020). However, less is known about how the use of video impacts assessors’ judgement processes; specifically, the potential consequences in terms of IRR. The present study adds to this nascent body of work.

1.5 *Scope of the thesis*

This thesis was positioned at the intersection of the fields of nursing OSCEs, inter-rater reliability, and assessor cognition. Numerous studies (e.g., Gingerich et al., 2014b; Chahine et al., 2016) have investigated assessor cognition and IRR using assessments in medicine as their object of study, while multiple studies exist which address score reliability in nursing OSCEs (e.g., Dunbar, 2018; Goh et al., 2019). Less work has been done on assessor cognition as it relates to nursing OSCEs. A notable example is a significant study by East et al. (2014); however, they did not seek to draw explicit links between assessors' cognitive processes and the scores they awarded. As such, the present study is among the first to investigate these three areas simultaneously.

Researching assessors' cognitive processes also has implications in terms of the fairness of nursing OSCEs. Fairness is a core principle of assessment, along with validity and reliability, and has multiple meanings within the field. One view of fairness is defined in the *Standards* as "the lack or absence of measurement bias" (AERA et al., 2014, p.51). When two candidates take a test, it should not be the case that one receives a better score than the other solely due to bias on the part of assessors, for example on the basis of race or gender. Numerous authors have written about how implicit prejudices such as sexism or racism may impact performance assessment scores (e.g., Lumley & McNamara, 1995; Mortsiefer et al., 2017); however, the evidence available in the literature indicates that such biases, if they do exist within the field of OSCE assessment, account for minimal amounts of score variance (e.g., Denney et al., 2013; Schleicher et al., 2017). For example, a large study of score data from over 2,000 assessors of a postgraduate medicine OSCE found that no assessors displayed evidence of bias on the basis of the sex of the test taker, while only a single assessor displayed potential bias on the basis of test-taker ethnicity (McManus et al., 2013). As such, the investigation into bias on the part of assessors was not an explicit aim of the present study. However, the methodology employed nonetheless had the potential to uncover such biases, which threaten the fairness of OSCEs.

This project provided the opportunity to investigate in detail how assessors of performance assessments in nursing form judgements about students. However, human judgement of performance is a feature of many different types of assessment, both within third-level education, as well as the field of credentialing (see Feldman et al., 2012, for an overview of performance assessments in medical credentialing). As such, in addition to being directly relevant to OSCEs within the School, the results of the present study have relevance to OSCEs administered in other contexts, as well as performance assessments more broadly. These implications are discussed in Chapter Five. This thesis was supported by an organisation within the credentialing industry, and there is a brief reference to the relevance of the study findings to this industry in the final chapter.

As noted by O'Leary et al. (2018), the 21st century has seen a marked increase in the use of technology in assessment, with significant developments in the machine scoring of constructed response items, large-scale assessments of collaborative problem-solving, and virtual reality simulations. The use of technology became a necessity in many areas of education due to the COVID-19 pandemic which closed educational institutions and testing centres across the world (UNESCO, 2020). However, it is important to note that while technology played a role in this thesis (through the use of video), the focus of this work was squarely on human assessors. As such, it was reflective of the ongoing need to keep human judgement at the heart of certain assessments, in order to ensure that decisions made on the basis of these assessment scores are valid.

1.6 *Structure of the thesis*

This thesis consists of five chapters. The following chapter presents the literature review, in which research in the fields of inter-rater reliability, nursing OSCEs, and assessor cognition is critically examined in order to identify gaps in published work and inform the research questions. In particular, empirical research into assessors' cognitive processes is evaluated in detail, in order to determine how these processes have been investigated in the past, and the relevance of existing findings to the context of the present study.

Chapter Three describes the conceptual framework and methodology employed in this study. This chapter contains a detailed description of the various phases of research that were required in order to address the research questions, including the recording of a series of bespoke videos of students completing two OSCEs.

Chapter Four presents the findings of the study, organised into three sections. The first section details the results of the qualitative elements of the study, which allowed the researcher to map how assessors in the sample formed judgements about students, and the likely implications in terms of IRR. The second section details the results from the quantitative aspect of the study, in which descriptive statistics provide an overview of the level of score variance that was observed when the sample of assessors watched and graded the same four videos of student performances. The final section uses outlier sampling, in which both qualitative and quantitative data from the harshest and most lenient assessor in the sample are used to gain a deep understanding of how variance between these two assessors emerged.

Chapter Five discusses the implications of the findings of the study for institutions such as the School who wish to set up, or are currently using, OSCEs as part of their battery of assessments.

The thesis ends with a summary of the findings, the limitations of the study, and recommendations for future research.

Chapter Two

Literature Review

2.1 *Introduction*

As noted in the previous chapter, researchers working in the field of nursing OSCEs have noted a paucity of studies which explicitly address the inter-rater reliability (IRR) of such assessments (Goh et al., 2019). Additionally, of studies which do calculate IRR, there is well-documented variation in the reported levels, leaving open the question of how IRR can be optimised (Navas-Ferrer et al., 2017). Problems with IRR in health sciences assessment more generally have led researchers to focus on assessors themselves, and how their engagement with the task of judging and grading student performances might lead to divergence in awarded scores (Yeates et al., 2013; Gingerich et al., 2014a). Assessor cognition, therefore, represents a relatively novel method of exploring IRR. However, this approach has yet to be implemented in the field of undergraduate nursing OSCEs. In order to unpack these issues in more detail, this chapter critically examines previous research in the fields of IRR and assessor cognition.

2.2 *Organisation of the chapter*

Following a description of the literature search strategy (section 2.3), this chapter is divided into three sections. Section 2.4 discusses existing research into the IRR of nursing OSCEs. The section begins with a summary of studies which have calculated the levels of IRR present in nursing OSCEs, drawing in particular on two recent review pieces by Navas-Ferrer et al. (2017) and Goh et al. (2019). Subsequently, there is a description of how different OSCE designers have tried to maximise IRR levels, and the persistence of issues with IRR in spite of these attempts. The section ends with a discussion of the relationship between reliability data (such as IRR) and validity. If valid decisions are to be made on the basis of awarded scores, there is a need to document evidence of high levels of scoring consistency. However, recent literature has argued that the relationship between validity and reliability is more nuanced than previously imagined, and that efforts to remove all subjectivity from performance assessments are misguided. This literature is examined in order to situate the current study within the broader research landscape of reliability as it pertains to performance assessments.

The need for any empirical study into assessor cognition to have a strong theoretical foundation has been noted by researchers in the field (Gauthier et al., 2016). To that end, section 2.5 outlines the theory of judgement formation that underpins the present study. Using social cognition literature as a lens through which to view the complex ways that individuals make sense of the world around them, assessors' judgement formation is conceptualised as a three-stage process, affected by cognitive load and expertise.

Section 2.6 critically evaluates research into assessor cognition, organised according to the three-stage framework of observation, processing and integration outlined by Gauthier et al. (2016). This section provides a detailed description of studies which have attempted the map the specifics of what happens when an assessor judges and grades an OSCE performance. Thus far, research has indicated that assessors employ a range of subjective judgement processes which have the potential to adversely affect IRR (e.g., Kogan et al., 2011). However, these studies have almost exclusively taken place in the field of medical assessment. As such, the potential for assessor cognition to inform research into the IRR of nursing OSCEs has thus far yet to be realised. The chapter ends with a summary of key findings and an identification of the gaps in the literature that this thesis seeks to address. These gaps form the basis for the research questions and conceptual framework which are outlined in Chapter Three.

2.3 *Scope of the literature review*

An extensive literature review was carried out to uncover and assess research on the bodies of work noted above. Searches were mainly conducted using the PubMed and Cumulative Index to Nursing and Allied Health Literature (CINAHL) databases. Relevant research overwhelmingly came from the disciplines of medicine and nursing. Where possible, nursing research was prioritised in the review, however in some areas (particularly assessor cognition) the nursing literature on the topic was sparse and medical literature was therefore examined. Discussions of the relevant commonalities and differences between the fields of medicine and nursing are present in the chapter when needed. Published work from education was also included, in particular where it pertained to discussions of key assessment issues such as validity and reliability. Important search terms used were 'OSCE', 'validity', 'reliability', 'assessment', 'medicine', 'nursing', 'performance assessment', 'performance evaluation', 'inter-rater reliability', and 'assessor cognition'; with the Boolean operators 'AND' and 'OR' used to identify work that contained various combinations of these terms. While there was no temporal cut-off point for included studies, those which were published more recently (particularly in the last decade) were afforded more prominence in the review, as were older studies considered seminal due to the high number of citations they have received. Reference lists from topical articles were also used to identify other relevant pieces of work. The initial literature review took place between November 2019 and July 2020. The researcher re-appraised the literature in April 2021, to determine whether relevant articles had been published in the interim. Subsequent to this, the literature was checked on an iterative basis to ensure no recent articles were overlooked. Ultimately, over 200 articles were read in-depth, mostly from peer-reviewed academic journals in medicine and nursing.

One point to note at this juncture is that human judgement of performance is a feature of numerous different assessments. As such, the issue of how assessors form judgements about test-taker performance is not one that is exclusive to the healthcare space. However, an explicit decision was made only to focus on empirical research that has taken place within the health sciences specifically. There were two related factors underpinning this decision. Firstly, the research was funded by an organisation within the credentialing assessment industry, who specified that they wanted research conducted about OSCEs, as they administer OSCEs as part of their operations. Secondly, given that the researcher knew that the study would ultimately focus on OSCEs, it was deemed that empirical research from non-healthcare fields would have limited applicability to nursing OSCEs. This is because of the phenomenon of task specificity (Govaerts et al., 2013; Gauthier et al., 2016), which posits that assessors' cognitive processes are affected by the task they are assigned to do. In other words, it cannot be assumed that the cognitive processes employed by assessors who have to test a candidate's driving ability are the same that would be employed when assessing a student's ability to take a blood pressure measurement. As such, the focus on research from within the healthcare field ensured that the researcher would not be making claims of transferability across disparate areas of assessment.

2.4 *Inter-rater reliability of nursing OSCEs*

This section critically examines published research related to the IRR levels present in nursing OSCEs. As noted in the previous chapter, IRR is one method of calculating reliability, and is a particularly important consideration for assessments, such as OSCEs, in which multiple assessors are responsible for judging and grading student performances (Gwet, 2014). The section begins with an overview of the position of OSCEs within the broader array of assessments used in undergraduate nursing programmes, and the purported benefits of the OSCE in terms of reliability. This is followed by a discussion of the state of research into nursing OSCE IRR levels, and an analysis of the relationship between OSCE design and IRR. The latter is particularly important, given that one of the key considerations that informed the design of the OSCE was assessor consistency (Harden, 1975; Khan et al., 2013a). Subsequent to this, there is a focus on the role of the assessor in published work pertaining to IRR in nursing OSCEs, and the attempts (or lack thereof) that researchers have made to bring about higher rates of scoring consistency. The section ends with a discussion of the relationship between reliability and validity. The key takeaways from section 2.4 are as follows:

- It remains important for developers and administrators of OSCEs to conduct reliability analyses (with a specific focus on inter-rater reliability) on OSCE scores, in order to ensure scoring consistency and valid inferences. However, published work on nursing OSCEs has often failed to include a calculation of IRR.
- When IRR is calculated, it is subject to large fluctuations, from near-perfect to poor. As such, the issue of how to optimise IRR is a pertinent one.
- Discussions of IRR in research into nursing OSCEs has often failed to focus on the assessors themselves, and how they engage with the task of assessment. Such an approach might allow for a deeper understanding into how issues with IRR emerge.
- While reliability data are needed in order to make valid judgements about students, recent research has problematized an excessive or exclusive focus on reliability. Some researchers have called for more subjectivity to be brought into the assessment process.

2.4.1 *OSCEs at undergraduate nursing level*

The OSCE is one of numerous assessment formats that are commonly administered in undergraduate nursing programs. As noted by Rushforth (2007, p.483), any discussion of the relative merits of the OSCE should take place “in light of consideration of other modes of assessment” that are used to assess student nurses. Nursing students’ cognitive capabilities are generally assessed using either multiple-choice tests or essays, which require them to demonstrate their theoretical knowledge about a range of different concepts. These tests are generally deemed to have high levels of reliability; however, they merely allow for an assessment of “what students know and understand”, instead of “their actual competence or performance” (p.483, emphasis original).

The most high-fidelity way of assessing actual performance is through assessments that take place within the world of clinical practice (often known as Workplace-Based Assessments (WBAs)). These assessments generally involve an assessor or examiner observing a real, non-simulated encounter between a patient and a student nurse, and awarding them a grade (Rushforth, 2007). However, given that there is no standardisation across different situations that students may encounter, scores derived from these assessments cannot be meaningfully compared, and their potential for informing high-stakes decisions about students is minimal.

As such, the OSCE can be conceptualised as an assessment format that attempts to combine the standardisation (and, in theory, reliability) inherent in assessment modalities such as multiple-choice exams with the practical, high-fidelity advantages of WBAs. As discussed in the previous chapter (section 1.3.1), the OSCE is not the only simulated performance assessment that nursing students may be subject to; however, it is widely perceived as being more reliable and objective than both the short case and long case examination formats. The OSCE, therefore, is often deemed the "gold standard" of performance assessments in nursing, and the health sciences more generally. However, as noted by Rushforth (2007, p.483), the "high level of validity and reliability" that the OSCE is considered to have cannot be assumed to exist a priori, and should be "carefully appraised" where possible.

2.4.2 *Reviews of IRR levels in nursing OSCEs*

A notable body of work has investigated the reliability of nursing OSCEs, as well as other performance assessments used in nursing. This section outlines several large scale systematic reviews that have been carried out on the subject of nursing OSCE reliability, concluding that a calculation of IRR specifically is often missing from published work on nursing OSCEs; and when it does take place, occasionally shows that IRR is below the level required for valid decisions to be made on the basis of awarded scores.

Navas-Ferrar et al. (2017) conducted a systematic review of 19 studies which investigated the validity and reliability of nursing OSCEs. They detailed several factors that are likely to affect the reliability of awarded scores (such as number of stations), but noted a variance in how these factors affected reliability. For example, while increasing the number of stations generally increased score reliability, this was not always the case: some studies reported high reliability in spite of having few stations. Overall, they found that when IRR was calculated, it was generally classed as moderate to high (most commonly measured using the Intraclass Correlation Coefficient), but in at least one case as low. Similarly, Cant et al. (2013) carried out a review of nursing OSCEs in order to determine whether scores demonstrated adequate levels of validity and reliability. In the sixteen papers reviewed, they found that only two contained a calculation of IRR. As such, they urged researchers to provide "adequate statistical justification of instruments' validity and reliability" (p.174).

A recent article by Goh et al. (2019) is illustrative of the persistence in nursing education literature of studies which either do not report on IRR, or else report that it is subject to large fluctuations. They conducted a review of published work pertaining to OSCEs in nursing, with one of the stated aims to determine whether scores derived from OSCEs are generally found to be reliable. 121 studies published since 1982 were examined. They found that IRR was only reported in 16 of the 121 studies, and where it was calculated, there was substantial variation: while one study reported near-perfect agreement of 0.99, another study reported a "poor" score of 0.28 (measured using the Intraclass Correlation Coefficient). Additionally, authors writing about IRR in nursing OSCEs have noted that OSCEs themselves tend to be highly variable, with test administrators likely to tailor the OSCE to suit the specific needs of the institution (such as a university) where it is being administered (Rushforth, 2007). As such, the generalisability of IRR statistics across institutions is limited. In other words, just because some nursing OSCEs have demonstrated high levels of IRR does not mean that this is likely to be replicated in other contexts. As noted by Rushforth (2007, p.484):

> ...it could be argued that IRR data has limited transferability to other situations and thus any group setting up a new OSCE should demonstrate their own IRR for that particular examination and group of examiners, particularly if single examiner stations will be used.

As such, there is an ongoing need to conduct IRR analyses of nursing OSCE scores, particularly those in which a single assessor is responsible for awarding a grade to a student (which is the case in the present study, as will be discussed in the following chapter).

One method of measuring IRR is to get OSCE assessors to watch the same sample of recorded OSCE performances, and complete the relevant marking guides (e.g., Roberts et al., 2020). This allows for a direct comparison of how different assessors graded the same performances. Generally, if agreement between raters is high, it is assumed that there will not be systematic IRR errors across the broader administration of the OSCE. In contrast, if this method reveals significant differences in how the same performances were graded by assessors, it is incumbent upon OSCE administrators to devise a means by which these errors can be corrected (Khan et al., 2013b).

Published work which describes such an approach to IRR is rare in nursing OSCE research. However, when implemented, it allows for a determination of what specific aspects of the OSCE are likely to result in unwanted score variance between assessors. Such studies are often underpinned by the assumption that certain skills, such as communication, might be more difficult to assess in a consistent way, as what constitutes good communication might vary between assessors (more than, for example, what constitutes “good” practical skills) (e.g., Brannick et al., 2011). Cazzell and Howe (2012) is an example of such an approach. They conducted a study in which two assessors watched 207 videos of student nurses completing an OSCE station designed to test their ability to administer medication to a patient. This study will be discussed in detail in section 2.5.1; but it is important to note that their results indicated that certain skills (particularly those in the “affective domain” such as empathy) are more prone to lower levels of IRR. A similar study by Dunbar (2018), in which six nursing assessors watched and graded the same recorded performance, allowed for calculations of IRR for each item on the marking guide, to determine what items were causing assessors to disagree with each other. These studies indicate the utility of an approach to IRR in which recorded performances are used in order to calculate IRR, and pinpoint areas of the OSCE that are causing issues with IRR.

2.4.3 *Relationship of IRR to OSCE design*

An examination of several studies in nursing OSCEs which have conducted thorough IRR analyses reveals a distinct lack of consensus as to how IRR levels can be improved. The review by Navas-Ferrer et al. (2017) discussed this issue in terms of assessment design. As noted by numerous authors, administrators of nursing OSCEs have often opted for a lower number of longer stations, a distinct break from the original OSCE devised by Harden (1975), which contained 18 stations of 4.5 minutes in length (Rushforth, 2007). This approach is thought to be more reflective of nursing practice, but may entail a reduction in score reliability:

> We also noticed a huge number of studies utilizing non-classical OSCEs with only 1 or 2 stations… While the OSCE modifications may be logical attempts to improve its validity to measure the holistic nature of nursing competence, it might come at the cost of reliability.
>
> (Goh et al., 2019, p.14)

As such, authors discussing nursing OSCEs have speculated that reducing the number of stations in an OSCE, and therefore the number of assessors that a student is graded by, may harm reliability levels, as the scope for a single particularly harsh or lenient assessor to impact the score awarded to a student is increased. However, this claim is not necessarily borne out in evidence. As noted by Navas-Ferrer et al. (2017, p.541), “Validity and reliability values in the analysed studies are similar independently of the used model, making it difficult to determine the best model." In other words, reliability coefficients in the reviewed studies were similar regardless of whether the OSCE comprised a larger number of short stations or a small number of long stations. The findings of this review are echoed in the field of medical assessment. In a review of 39 papers which reported on the reliability levels present in medicine OSCEs, Brannick et al. (2011, p.1186) wrote that while "empirical estimates of reliability increase on average with the number of stations, there is surprisingly large variability in reliability at any given number of stations". As such, the literature from both nursing and medicine OSCEs suggests that issues with reliability may arise regardless of the number of stations used.

The lack of a clear causal relationship between OSCE design and score reliability is also present regarding the layout of the marking guide used to assess students. Generally, marking guides can contain either binary checklist items (which ask assessors to classify a certain task or behaviour as done/not done) or global rating scales (which ask assessors to classify students according to a Likert scale, e.g., fail/borderline pass/good/excellent) (Navas-Ferrer et al., 2017). When the OSCE was first devised, it was assumed that binary items would lead to higher levels of IRR between assessors, however research has indicated that high levels of IRR can also be attained through the use of global ratings (Rushforth, 2007).

The results of the review by Navas-Ferrer et al. (2017, p.541) indicate that reliability issues are equally likely to emerge regardless of whether binary checklists or global ratings are used: "In this review, validity and reliability outcomes have not been very different regardless of the option chosen". The evidence, therefore, suggests that if researchers or OSCE designers wish to improve IRR levels, a focus on the design of the OSCE or the layout of the marking guide alone is insufficient. Put simply: "It is hard to determine which characteristics make the OSCE valid and reliable" (Navas-Ferrer et al., 2017, p.540).

2.4.4 *Role of the assessor in published IRR studies*

The logical next step is to examine how researchers documenting the development of OSCEs have attempted to put in place procedures to increase consensus between assessors, and whether these procedures have had the intended effect. Authors discussing the process of designing and implementing an OSCE have long noted the importance of such procedures in order to ensure that the OSCE produces reliable scores that can be used to make valid decisions about students (Khan et al., 2013b).

Three studies which describe the development and administration of nursing OSCEs are illustrative of these processes. In a recent piece describing the development of a five-station OSCE at a university in Singapore, Goh et al. (2022, p.2) documented how, when the OSCE was being devised, there was a process of "standardization based on the examiners' guide, inter-rater meetings, and psychometric testing of the OSCE instrument to ensure the objectivity of the assessment".

Discussing the administration of an eight-station OSCE in mental health nursing, Selim et al. (2012, p.286) went into more detail about the processes in place to bring about high levels of IRR:

> …gathering a team, intense preparation of the blueprint, creating a bank of stations, creating scenarios, training of actors in simulated patient stations, preparation of checklists of rating, training of raters, using two raters in each simulated patient station, creating post simulated patient stations, preparing model answers, reviewing the contents of OSCE by faculty teaching staff and comparing the OSCE content with the intended learning outcomes of psychiatric nursing curriculum.

A similar procedure is described by Corcoran et al. (2013, p.295), who devised a three-station OSCE for palliative care nursing, in which Standardised Patients also acted as the assessors:

> Individuals portraying the standardized patients participated in a 4-hour training session in which they discussed the goals of the learning experience, practiced role-playing the patients, learned how to complete the interpersonal skills and stations checklists, and provided verbal feedback to learners. Faculty certified the accuracy of SPs' portrayal and scoring, as well as their effectiveness in delivering verbal feedback on interpersonal skills.

As such, in published work on IRR in nursing OSCEs, there is usually a description of the means by which the OSCE designers attempted to ensure scoring consistency between assessors. However, what is notable regarding the three studies above is that, in spite of having these seemingly rigorous procedures in place, there were vast disparities when IRR was calculated. The first study resulted in a near-perfect Intraclass Correlation Coefficient (ICC) of 0.948, classified as "excellent" (Goh et al., 2022, p.3). The second reported more mixed results: of the three stations which were piloted and subject to IRR calculations, one reported a "moderate" Spearman's *r* of 0.581, while the other two reported "strong" scores of 0.672 and 0.708. No station recorded a "very strong" value of 0.8 or above (Selim et al., 2012). The third study reported “poor” ICC values of 0.22 and 0.46 for two stations and a “moderate” value of 0.52 for the third station (Corcoran et al., 2013). The extreme disparity in IRR levels across these three studies is puzzling, given that they seemed to have similar safeguards in place to prevent problems with assessor consistency.

What these studies have in common is that, aside from the quoted extracts above, they do not detail the specifics of how assessors engaged with the task of judging and awarding grades to students. This is perhaps to do with resource constraints: as noted by Goh et al. (2019), only 16 of the 121 studies in their review included a calculation of IRR at all. The fact that these studies took the time to document IRR levels in their OSCEs means that they afforded the issue of IRR more attention than most. However, this still leaves open the question of why some OSCE designers managed to bring about almost perfect IRR levels, while others did not. These studies failed to investigate the nuances of how assessors formed judgements about the students they assessed, and whether such an investigation could explain in a deeper sense how inconsistencies between assessors arise, and how they could be addressed.

It is suggested that three conclusions can be drawn from this body of work discussed in section 2.4 thus far. The first is that IRR is often not determined at all in published work on nursing OSCEs. Indeed, Goh et al. noted that there is “insufficient, unclear, or missing data in many OSCE studies” (2019, p.5). This trend has been noted elsewhere: “in the nursing literature related to the evaluation of clinical skills, the establishment of interrater reliability is inconsistently addressed” (Dunbar, 2018, p.137). There is scope for studies which offer a thorough detailing of psychometric measures such as IRR in nursing OSCEs, and what aspects of the OSCE are liable to threaten IRR. Secondly, published work which does report IRR levels indicates that issues with IRR may arise regardless of the design of the OSCE, and in spite of procedures being in place to increase assessor consistency (Navas-Ferrer et al., 2017; Corcoran et al., 2013). Finally, researchers documenting IRR values have thus far failed to explore the link between how assessors form judgements about student OSCE performances and the resulting implications in terms of IRR. Such an approach might allow for a nuanced understanding of how divergence in awarded scores arises, and how it could be accounted for.

2.4.5 *Relationship between IRR and validity*

2.4.5.1 *IRR as an important source of validity evidence*

Ultimately, concerns over IRR have to do with the validity of nursing OSCEs. Validity is a key concept in testing, generally regarded as the most important consideration when designing and administering assessments such as the OSCE. This section examines how validity is conceptualised in medical and nursing assessments. A working definition of validity, based on the *Standards in Educational and Psychological Testing,* is provided, and the process by which inferences from test scores are deemed to be valid is discussed. A calculation of score reliability, such as IRR, is generally seen as a crucial stage when building the validity argument for an assessment, and has traditionally been the dominant concern of researchers writing about OSCEs (Hodges, 2013). However, in recent years an emerging discourse has called into question this singular focus on reliability as the most important piece of validity evidence. Some authors have noted that an excessive focus on reliability might threaten, rather than enhance, the validity of decisions made on the basis of OSCE scores. The section ends with a discussion of this perspective, and its relevance to this thesis.

Validity is defined in the *Standards* as "the degree to which the evidence and theory support interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p.11). Crucially, it is interpretations of test scores that are deemed to be valid or invalid, not the test in and of itself. Just because we can make a certain inference based on a test score does not necessarily mean we can make another inference based on the same score (AERA et al., 2014). For example, a student's score in a test designed to measure how well they have mastered the first-year undergraduate nursing curriculum does not predict how competent they would be at practicing as a nurse in a hospital. In order to determine this, more information about the student (in the form of further assessments) would have to be obtained. As such, validity refers to how appropriate it is to interpret a test score in a certain way. In terms of an OSCE, decisions on the basis of OSCE scores are considered to be valid if we can say with confidence that those scores reflect students' true competency levels at the assessed tasks.

In order to determine whether decisions made on the basis of assessment scores are valid, test designers need to document multiple sources of "validity evidence" which justify the uses of test scores (AERA et al., 2014). This process is known as validation, and developers should have a clear validation plan, and make their sources of validity evidence explicit, so that outside observers can determine the defensibility of inferences made on the basis of test scores (Bandalos, 2018). An important step in the validation process involves determining the extent to which test scores are reliable.

Reliability is broadly defined in the *Standards* as "consistency of the scores across instances of the testing procedure" (AERA et al., 2014, p.33). Generally, if a person takes the same test more than once, it is logical to assume that this person will achieve the same score, assuming that performance on the second taking of the test is not influenced by doing the first test. The numerical value of a test's reliability is usually called the reliability coefficient (AERA et al., 2014). In the same way that it is incorrect to speak of a test on the whole as being "valid" or not, tests cannot be said to be "reliable" – rather, it is the test scores that can be reliable or unreliable (Thompson & Vacha-Haase, 2017). If test scores are unreliable, then we cannot confidently make inferences from them. As such, the importance of calculating reliability levels when building a validity argument is well-recognised. This is especially true when high stakes are associated with the outcome of an assessment; for example, whether a nursing student is permitted to enter into the next year of study. Without evidence that the student would likely have achieved the same score regardless of the assessor (or series of assessors) that assessed them, the decision to allow the student to progress is clearly less defensible. This highlights the importance of conducting thorough IRR analyses of OSCE scores, and using this information as part of a larger validity argument. Indeed, it is the importance of documenting IRR levels that have caused authors writing about nursing OSCEs to encourage research into IRR and how it can be improved (e.g., Dunbar, 2018; Goh et al., 2019).

2.4.5.2 *Concerns with excessive focus on reliability*

In tandem with research which seeks to optimise score reliability in OSCEs, there is an emerging discourse in medical and nursing education that calls into question the perceived narrow focus on reliability that is at the heart of such assessments. Some authors have argued that the proliferation of these tests leads to a situation where students can succeed in examinations by merely completing an assigned series of tasks, rather than having an overall understanding of the "art" inherent in good practice (Khan, 2017). In terms of the assessor, there is a reduced role for expert judgement, leading to a situation where assessors become "box-tickers", a view that ignores the reality that assessors are "active information processors who interpret and construct their own personal reality of the assessment context" (Govaerts & van der Vleuten, 2013, p.1169). This emerging perspective, and its implications for this project, are discussed below. Of particular importance for this thesis is the argument that an excessive or exclusive focus on reliability and objectivity is both an unachievable and misguided goal for assessments.

Hodges (2013) argued that the last two decades have seen an increased focus on psychometric concerns in medical assessment, chief among them the issue of reliability: "It is difficult to overstate the degree to which concerns with reliability dominated assessment during the late twentieth century" (p.564). Because of this, objectivity became one of the primary goals of assessments, and any hint of subjectivity in an exam was deemed an inherently negative thing – subjectivity was equated with bias and unfairness. In a recent article critiquing objectivity in medical assessment, ten Cate and Regehr (2019) wrote that what is often conceived of as "objectivity" is instead best understood as a shared subjectivity. Design and administration of assessments involves decisions from numerous stakeholders, and assessments are generally only finalised after agreement between multiple people. However, just because this agreement has been made does not mean it is "objective" - another group of equally qualified stakeholders might have come to a different conclusion.

Hodges (2013) argued that there has been a shift from assessing character to assessing characteristics - instead of evaluating medical students in a way that allows for judgement of their overall demeanour, they are judged on the basis of being able to perform a series of individualised tasks, such as those listed on the marking tool of an OSCE station (Hodges, 2013). In terms of the assessor, Eva and Hodges (2012) recognised that experienced assessors might be able to identify some gestalt (overall) impression that is demonstrated by medical students, and argued that assessments should be created which allow for assessors to use their "expert judgement" (p.917) to identify otherwise talented students who may not necessarily complete all tasks on the marking guide.

There is an argument in the literature that a strong case must be made for subjectivity to be brought back into the assessment process. ten Cate and Regehr noted that "there is a growing and appropriate challenge of the psychometric premise of objectivity and its attendant construction of variability in assessment" (2019, p.335). Likewise, Hodges wrote that there is currently "an effort to rehabilitate subjective judgement" (2013, p.566), with the aim of allowing assessors greater freedom to identify trainees they believe possess what it takes to be a good practitioner, even if their qualities are difficult to capture on a limited checklist. Van der Vleuten et al. (2010) noted that

> Increased control of the noisy real world by standardising, structuring and objectifying is not the answer. On the contrary, it will only harm and trivialise the assessment. To improve we must 'sharpen' the people rather than the instruments (p.712).

Following on from this, it is reasonable to infer that the different ways assessors might view an OSCE performance is not necessarily just 'error', rather in some cases it can represent different, but equally reasonable, subjective judgements. Such an argument calls into question the idea that when a student participates in an assessment, there exists one "true" score that all assessors should be trained to identify (ten Cate & Regehr, 2019). Efforts to "negotiate a single common perspective among perceivers is problematic" as it precludes assessors from using their idiosyncratic perspectives, and leads to a loss of information about the student (ten Cate & Regehr, 2019, p.335).

Although the body of work discussed above, which examined the possibility of bringing subjectivity back into the assessment process, has taken place primarily in the field of medical education, an argument can be made that such concerns are also applicable to assessments in nursing. Indeed, authors in nurse education have long commented that efforts to completely remove subjectivity from assessments of nursing students are misguided and undesirable (Rushforth, 2007; Walsh et al., 2009; Mitchell et al., 2009). In an early review of 18 studies that described the implementation of OSCEs in the assessment of nursing students, Walsh et al. (2009) noted a consistent worry among nurse educators that the OSCE might be misaligned with nursing practice, as it emphasises the segmentation of skills and competencies in a way that is uncommon in nursing (which focuses on holism). As such, Walsh et al. noted a worry that striving for complete objectivity in nursing OSCEs would ultimately be detrimental, as it would necessitate this splitting up of abilities. Indeed, echoing the concerns of Hodges (2003) described above, Rushforth noted that, regarding nursing OSCEs: "there may be an inverse relationship between validity and reliability, with the strengthening of one potentially weakening the other; achieving good reliability *and* validity is arguably a considerable challenge" (2007, p.488, emphasis original). Likewise, in a highly-cited piece reflecting on best practice for implementing OSCE in nurse education, Mitchell et al. (2009) commented that the OSCE might be "unable to take account of interacting contextual factors that are common in the clinical environment" (p.402). They similarly urged caution about developing nursing OSCEs that were decontextualised and sought to remove all variance between assessors.

It is suggested from these influential papers that concerns around objectivity (specifically the striving for objectivity as the primary goal when developing assessments) are apparent in the nursing literature as well as the medicine literature. However, this does not mean that subjectivity should necessarily be pursued in every case, as an excessive amount of subjectivity may threaten the reliability that is necessary for valid assessment decisions. This tension between objectivity and subjectivity is present elsewhere in the nursing literature. A study by

East et al. (2014), discussed in detail in the following section on assessor cognition, concluded that nursing OSCE assessors frequently employ subjective criteria when evaluating student performances, and that these criteria are often based on their own experiences of clinical practice. However, the authors of that study noted that while assessors invariably bring their own experience to the assessment process, allowing them too much autonomy when judging students poses a threat to the reliability of awarded scores. The question of the extent to which nursing assessors should be encouraged to exercise their subjective judgements remains open.

This discussion of current debates around validity and reliability serves to situate this project within the broader research landscape. This thesis was originally developed due to concerns around the persistence of fluctuating IRR levels in nursing OSCEs, in spite of efforts to identify and account for threats to IRR. The decision to focus on IRR made sense in the context of the well-documented tendency of nursing faculty to use OSCEs as a means of summative assessment (Goh et al., 2019), and the resulting need to ensure that there is validity evidence to support these summative decisions. However, while researching this topic, it became apparent that focusing exclusively on IRR may be considered problematic by some, and that the relationship between reliability and validity is more complex than originally imagined. As such, while the study had the aim of using assessor cognition as a means of understanding how issues with assessor consistency arise, and how they can be mitigated, it took place with full awareness that attempting to bring about complete agreement between assessors may be neither possible nor desirable in terms of validity. As argued by Boursicot et al. (2021, p.2), in some instances, "rater variance is meaningful and such differences should be embraced rather than controlled". However, a detailed investigation into IRR is useful regardless of whether the ultimate goal is to completely remove all instances of assessor variance. Understanding how different assessors approach the task of judging and grading students, and how this leads to divergence in terms of awarded scores, will allow assessment designers to tease apart instances of variance that may be "meaningful" (Gingerich et al., 2014a), and those that can be classed as measurement error. This issue will be revisited in Chapter Five, in light of the study findings.

2.5 *Theoretical foundation*

As noted throughout section 2.4, efforts to understand how issues with IRR arise in performance assessments might be deepened through a focus on assessors themselves. As such, the remainder of this chapter focuses on research which aims to explore what specifically happens when assessors engage with the task of assessing students in performance assessments such as OSCEs, and the potential implications in terms of IRR. Section 2.5 outlines in detail the theoretical foundations of this study. As noted by numerous authors writing about the design of empirical studies (e.g., Fraenkel & Wallen, 2006), and about assessor cognition more specifically (Gingerich et al., 2011), it is important that any researcher explicitly outlines their theoretical perspective. Failing to do so reduces the generalisability of any potential results, and impedes future researchers who wish to replicate the described study. In this case, it was deemed necessary for the researcher to outline a specific framework for how assessors' judgement formation is conceptualised (Gauthier et al., 2016). Influenced by social cognition literature, judgement formation is understood in the present study as a three-stage process influenced by a range of factors, notably cognitive load and expertise. The relevance of these concepts to the issue of assessment is flagged throughout the section. The following section (2.6) contains a detailed discussion of empirical studies into assessor cognition, noting what has been investigated so far, as well as the gaps in the research which inform the present study.

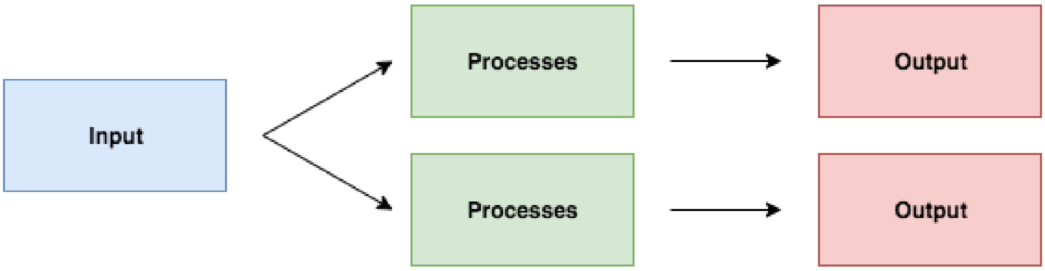
2.5.1 *Social cognition*

Social cognition is an area of research that seeks to determine how individuals understand the world around them. When moving through the world, individuals encounter an infinite number of stimuli, and rely on their own prior knowledge and experiences in order to make sense of these stimuli (Greifeneder et al., 2018).

Because an individual's knowledge and experiences are unique to them, the same stimulus might be understood differently by different people. Sam and Alex meet Pat at a party, and the three of them chat for a few minutes, before Pat leaves to chat to someone else. When discussing their opinions of Pat, it emerges that Sam viewed Pat as friendly and enthusiastic, whereas Alex viewed Pat as superficial and conceited. The reason for this is that Alex had heard from another friend that Pat is rude and self-involved. This affected Alex's opinion of Pat, but not Sam's. Thus, the process by which Alex used a piece of information about Pat in order to make sense of their interaction affected Alex's understanding of the situation differently to Sam's. This resulted in them having different outputs (in this case, opinions) about the same situation.

As such, it can be determined that when individuals make sense of a specific *input* (stimulus), they rely on their own internal *processes* (Greifeneder et al., 2018). This begins to explain why the same situation could be interpreted in completely different ways by two different people. How individuals construct the world around them has as much to do with their own knowledge and personalised cognitive mechanisms as it does the input itself (Greifeneder et al., 2018). This is illustrated in Figure 2.1 below.

Figure 2.1 *Inputs, processes and outputs*



An individual's prior knowledge can be defined as an internal *cognitive structure,* a "mental representation constructed from past experiences, essentially encompassing concepts of all kinds: traits, attitudes, beliefs, values, memories, scripts, stereotypes, self-concepts, and expectancies… These structures contain the person's knowledge, beliefs, and expectations about some stimulus domain" (Hamilton & Carlston, 2013, p.29). A cognitive structure is complex and multifaceted, and individuals have an essentially unlimited number of them.

A *cognitive process* is defined as "an umbrella term that represents the variety of mental activities that are engaged in comprehending and elaborating encountered information" (Hamilton & Carlston, 2013, p.29). Social cognition is concerned with cognitive structures and cognitive processes, and the complex interactions between the two. Examples of cognitive processes that are commonly mentioned in social cognition literature, and that will be explored in detail in the present study, include *making inferences, drawing comparisons* and *synthesising information.*

2.5.2 *Two-systems processing*

There are numerous ideas underpinning social cognition literature. Two interrelated concepts, two-systems processing and cognitive load, are discussed below. There is a consensus in social cognition literature that cognitive processes operate on a two-systems model, comprising System 1 and System 2 processing (Kahneman, 2011); referred to elsewhere as automatic and controlled processes (Hamilton & Carlston, 2013).

System 1 processing "operates automatically and quickly, with little or no effort and no sense of voluntary control (Kahneman, 2011, p.20). System 1 processing happens largely without us realising it, and is responsible for a range of different activities that allow us to make sense of different stimuli. Examples of System 1 activities are: detecting that one object is further away than another, orienting a sound to an object that made that sound, and answering a simple maths problem, such as 2+2=? (Kahneman, 2011, p.21). System 1 processing is "effortless, associational, and intuitive" (van Boven et al., 2013, p.394). System 1 is usually highly efficient at allowing us to process the world around us, however its quick nature means it is susceptible to "systematic errors that it is prone to make in specified circumstances" (Kahneman, 2011, p.25).

In contrast, System 2 "allocates attention to the effortful mental activities that demand it, including complex calculations" (Kahneman, 2011, p.21). System 2 processing is employed when System 1 does not allow an individual to understand something or solve a problem, and is more rational and controlled than System 1. In contrast to the automatic nature of System 1 processing, individuals are almost always aware of System 2 processing, and need to allocate sufficient mental resources to allow System 2 processing to take place. Examples of System 2 activities are: choosing to focus on the sound of one person's voice in a noisy room, filling out a detailed form, and solving a complicated maths problem, such as 43x61=? (p.22).

It is important to note that the cognitive mechanisms ascribed to Systems 1 and 2 are not uniform across all individuals. People who are skilled at or have a significant amount of experience in a specific area are much more likely to have the cognitive processes associated with that area as part of their System 1 (Kahneman, 2011). For example, an amateur chess player may have to think for a significant amount of time to determine the best move in a particular position (System 2), whereas a master player will likely be able to determine it straight away (System 1). Likewise, someone who has a lot of experience doing mental arithmetic, such as an accountant, might be able to determine the answer to the above maths problem (43x61=?) quickly and effortlessly. As such, the cognitive mechanisms of Systems 1 and 2 vary across individuals and are linked to an individual's experiences.

2.5.3 *Cognitive load*

Another concept which is central to the study of social cognition is that of cognitive load or cognitive strain, broadly defined as the amount of mental effort expended by an individual at any given moment. Kahneman (2011) conceives this mental effort as existing on a scale, with cognitive ease at one end and cognitive strain at the other end. When individuals are in a state of cognitive ease, they are likely to be in a good mood, trusting their intuitions, and feeling familiar in a given situation (p.60). Conversely, individuals in a state of cognitive strain are likely to be "vigilant and suspicious", investing more effort in what they are doing, "less comfortable…less intuitive and less creative" (p.60). An individual is more likely to be in a state of cognitive ease when they are experiencing something familiar (or that they have experienced in the past), when they are in a good mood, when they have been primed to encounter a specific situation, and when the object they are processing is easy to make sense of (Kahneman, 2011).

This idea of cognitive ease and cognitive strain is closely linked to System 1/2 processing described above. When an individual is in a state of cognitive ease, they are able to rely almost exclusively on System 1 processing, as the information they encounter is legible enough to them that they can make sense of it easily. However, as cognitive strain increases, an individual is less able to rely on System 1, and is forced to "mobilize System 2, shifting [their] approach to problems from a casual intuitive mode to a more engaged and analytic mode" (Kahneman, 2011, p.65).

Cognitive strain has significant implications for how individuals come to form judgements about others. When relying almost exclusively on System 1 processing, individuals are generally at ease and can make judgements about others quickly; however, they risk relying on mental shortcuts which affect their judgement in a way that is unfair on the target (Kahneman, 2011). When individuals are experiencing cognitive strain so that they have to rely on System 2 processing, they are less likely to make such errors of processing (Kahneman, 2011).

However, if the cognitive strain experienced by individuals is too high, it might overwhelm their mental capacity and they will not be able to attend to all the tasks at hand in the best way (Paravattil & Wilby, 2019). Consider the maths problem 43x61=? mentioned above as an example of a task requiring System 2 processing. An individual might be able to solve this problem if they were able to focus on it for a sustained period of time; however, if they were asked to solve it while simultaneously having to list all the countries of Europe, they most likely would not be able to.

Thus, the cognitive strain of the task(s) at hand has significant implications for an individual's cognitive processing. If an individual is experienced in a given situation, knows what to expect and encounters stimuli that are clear and easy to interpret, they are likely to be in a state of cognitive ease and rely on System 1 processing, however this does leave them open to lapses of judgement. Conversely, if they are under a higher level of cognitive strain, they have to employ System 2 processing, which makes them less likely to make mistakes, but they risk becoming cognitively overwhelmed.

2.5.4 *Expertise*

One of the key concepts which affects judgement formation, and which is particularly relevant to performance assessments, is expertise. Patricia Benner's influential work *From Novice to Expert* (1982) outlined the various stages through which nurses progress in the clinical environment as they gain experience. Benner's work has been instrumental in providing a framework against which the development of nurses can be tracked, and her work is still widely cited in nursing literature today.

Benner determined that, as nurses become more experienced, they progress through five different stages: *novice*, *advanced beginner*, *competent*, *proficient*, and *expert.* She noted that there were two distinct aspects of skilled performance that distinguishes it from unskilled performance. The first is "a movement from reliance on abstract principles to the use of past, concrete experience as paradigms" (1982, p.402). When nurses first begin to practice, the patient issues they encounter are new to them, and they have limited experience against which to compare incidents. As they develop, they are more easily able to interpret incidences in relation to those previously encountered. The second feature of skilled performance is a "change in perception and understanding" such that specific situations are "seen less as a compilation of equally relevant bits and more as a complete whole in which only certain parts are relevant" (1982, p.402). Experienced nurses are able to look at a patient and quickly determine which pieces of information about them are relevant in order to gain a deep understanding of that patient.

Although Benner's work was devised in order to conceptualise the skills acquisition of nurses practicing in the clinical environment, her work is highly relevant to judgement formation more generally. Applying her framework to the case of nursing OSCE assessors allows for an understanding of the skills that experienced assessors have developed. In the case of using past experience as a paradigm against which to interpret new situations, it can be argued that experienced assessors have built their own idea of what constitutes "good" performance, and judge students against this. This argument is a recurring one in assessor cognition research, and will be discussed in more detail in section 2.6.2.2. However, it is noted that an assessors' expertise *as an assessor* is distinct from their expertise as a nurse (or a lecturer). As such, it is possible for an assessor to be an expert nurse, but have low levels of expertise as an assessor.

Secondly, Benner's notion that experienced nurses are able to quickly identify the salient aspects of a given situation has clear relevance to the assessment process. It can be argued that experienced assessors are able to direct their attention to aspects of a student's performance that are most important, and filter out unnecessary information. This idea is similarly common in the field of assessor cognition, and will be examined in more detail in 2.6.1.

Benner's work on expertise allows for a conceptualisation of experienced nursing OSCE assessors as being able to quickly identify the important aspects of a student performance, a process which is affected by their own knowledge and experiences. As such, expert assessors are likely to be able to rely on System 1 processing (Kahneman, 2011), and assess students without putting themselves under a significant amount of cognitive strain. In contrast, less experienced assessors are more likely to be put under cognitive strain when assessing students, and have to rely more on System 2 processing. However, because an individual's knowledge and experiences are unique to them (Greifeneder et al., 2018), it is possible that assessors with a high level of expertise may assess students in ways that are idiosyncratic, which has implications for score reliability. The link between expertise and cognitive processes is explored further in section 2.6, when empirical work on assessor cognition is critically evaluated.

2.5.5 *Three stage process*

There is a general consensus in social cognition literature that, when making sense of a stimulus, individuals go through a series of cognitive processes in a specific order. As noted by Gilbert in *The Handbook of Social Psychology* (1998): "The notion that stimuli are *sequentially* sensed, transduced, encoded, represented, elaborated, integrated, rehearsed, stored, and retrieved is... a ubiquitous aspect of modern theorising" (p.104, emphasis added). These different steps have been classified into three broad stages: identification, attribution, and integration (Gilbert, 1998). These three steps are examined in detail below.

1. Identification: this is the first stage, and is defined by Gilbert (1998) as "the process by which observers identify acts" (p.106). Because the human capacity for processing is necessarily constrained, individuals cannot choose to take in and make sense of every possible stimulus (or aspect of a single stimulus). Instead, we "direct our attention to some aspects of the situation and exclude other aspects from being processed further and thereby taxing resources" (Greifeneder et al., 2018, p.26). Individuals are more likely to direct attention to stimuli that are *distinctive* (in relation to other stimuli) or that constitute a discrepancy vis-à-vis our prior knowledge (for example, if an acquaintance acts in a way that is incongruous with what we know of them) (Greifeneder et al., 2018). Additionally, the identification phase is likely to be

affected by our *goals in a given situation*: if we want to determine how tall someone is in relation to others around them, we would pay particular attention to the heights of everyone in the room (Greifeneder et al., 2018).

2. Attribution: this is "the process by which observers draw dispositional inferences from the acts they have identified" (Gilbert, 1998, p.107). Having observed a stimulus, it is now incumbent on individuals to make sense of that stimulus. The different ways that individuals can give meaning to a particular stimulus is dependent on their own prior knowledge, and generally entails "going beyond the information given" (Greifeneder et al., 2018, p.30) in order to draw general conclusions from a specific stimulus. Because the same stimulus could be understood in a variety of ways (e.g., a person talking at length may be interpreted as *enthusiastic* by some but as *self-involved* by others), it is impossible to say with certainty how a given stimulus will be interpreted by another person.

3. Integration: this is "the process by which observers form impressions from the dispositional inferences they have drawn" (Gilbert, 1998, p.105). Having observed a stimulus, and made sense of that stimulus, the next step is to come to a conclusion about why that stimulus occurred. When the stimulus relates to another person, individuals generally seek to "construct unified explanations" of what that person is like (Gilbert, 1998, p.109). Although it is theoretically possible to think of an individual observing and interpreting one stimulus at a time, in reality individuals process a range of different stimuli, and use a series of these when forming a judgement (Greifeneder et al., 2018).

2.5.6 *Usefulness of social cognition to understand the role of the assessor*

There are four reasons why social cognition is a useful lens through which the view the problem of assessor consistency. Firstly, as documented in section 2.4, in spite of a significant body of work that has sought to remove completely all variance between assessors, such that IRR is at the maximum possible level, it has been noted in both medical (Brannick et al., 2011) and nursing assessment (e.g., Navas-Ferrar et al., 2017) that the reliability of undergraduate OSCEs is prone to pronounced fluctuations, even when assessment designers make attempts to improve it. This has led researchers to question the utility of a purely psychometric approach to the problem of assessor consistency, in which all variance between assessors is treated as error which can be removed through appropriate interventions (e.g., Gingerich et al., 2011; see section 2.4.4).

Secondly, the efforts of social cognition research to investigate individuals' subjective construction of social reality parallels calls in both medical and nursing assessment for a greater level of subjectivity to be brought back into the assessment process. In medicine, authors such as Hodges (2013) have noted that assessment has begun to focus almost exclusively on reliability as the sole indicator of validity, such that efforts to increase reliability have become the central method through which assessments are improved. Although this has brought about transparency and fairness in assessment, it has constrained the role of assessors, leading to a situation where assessors have become "box-tickers" with limited scope to exercise their own informed judgements (see section 2.4.4). Likewise, researchers in nursing assessment have similarly voiced concerns about efforts to remove all subjectivity from assessment, noting that doing so obscures the complexity inherent in nursing practice (e.g., Mitchell et al., 2009).

Thirdly, a significant strand of work within social cognition seeks to understand the process by which individuals come to make a judgement or decision about another person on the basis of a small amount of evidence. Studies such as that by Mohr and Kenny (2006) have investigated how a small amount of information about another person is encoded and interpreted by individuals, and how this process can lead to different people having divergent views about the target person (even when they have access to exactly the same information). This body of work parallels performance assessment situations, such as the OSCE, wherein assessors have a short amount of time to come to a decision about a student's proficiency at performing a range of psychomotor and affective skills.

Finally, initial work in the field of assessor cognition has explicitly emphasised the utility of adopting a social cognitive approach when seeking to understand the issue of assessor consistency. In part influenced by the persistent lack of consistency between assessors noted above, authors such as Gingerich et al. (2011) have noted the parallels between social cognition and the performance assessment space: "social cognitive explorations of [assessor judgements] could be useful in better understanding the cognitive processes used by raters within the social context of rater-based assessments" (p.1). This thesis is an explicit extension of this framework, using social cognition as a lens through which to explore the issue of assessor judgement processes as they manifest in undergraduate nursing OSCEs.

2.5.7 *Social cognition's impact on assessor cognition*

Important early studies in the field of assessor cognition sought to use a social cognition perspective in order to better understand the phenomenon of rater consistency. An influential piece by Gingerich et al. (2011) explored the issue of assessor judgements, noting that the standard approach of trying to remove all variance between assessors had consistently failed to produce the desired outcomes in terms of score reliability. They wrote that "Given the apparent intractability of this problem using our standard frameworks, it might be worth exploring other approaches to understand the manner in which people represent and make determinations about others" (p.1). They used social cognition as a lens through which to deepen understanding of how assessors make decisions in performance assessments, and why variance between assessors arises.

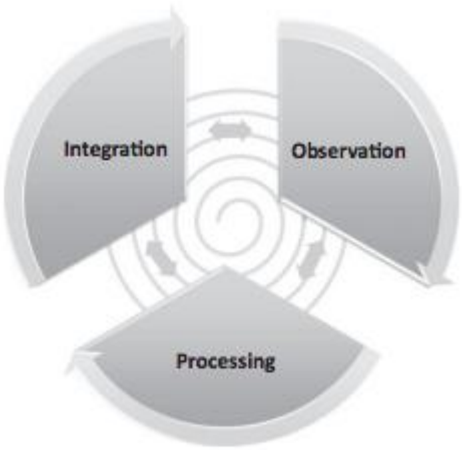
They examined how social cognition research might be applicable to the field of performance assessments. Although they did not draw any firm conclusions about whether a certain approach was likely to yield specific outcomes, their work was important in bringing about a shift in perspective; wherein assessor variance was reconceptualised as an important object of study, rather than a problem to be eradicated in every case. The ultimate goal of this shift is to improve assessment quality: "Through better understanding of how raters make judgements during the assessment process, we may be able to tease apart error attributable to human biases and error unintentionally imposed by assessment systems that are incongruent with innate human cognition" (Gingerich et al., 2011, p.4). In other words, focusing on assessors and how they judge performance assessments may lead to a deeper understanding of how score variance arises.

Following on from this, numerous authors working in the field of medical assessment have incorporated elements of social cognition research into their examinations of assessor consistency in performance assessments. For example, Govaerts et al. (2013) sought to determine whether "theoretical frameworks of social perception can be used to further our understanding of processes underlying judgement and decision-making in performance assessments in the clinical setting" (p.377). Their study will be evaluated in detail in section 2.6, but it is noted that the past decade has seen an increase in published work that has the aim of understanding the mechanisms used by assessors when they make judgements in performance assessments, and how this judgement formation affects score reliability and assessment design.

Subsequent to this body of work which sought to illuminate how social cognition could be useful to the field of medical performance assessments, several authors have devised specific frameworks to map the processes by which assessors form judgements about students. One important piece is by Gauthier et al. (2016), where the authors conducted a review of 34 pieces of peer-reviewed work in assessor cognition, as well as an additional 44 pieces of "grey literature" (such as reviews and commentaries). From this review, they conceptualised judgement formation in performance assessments as a three-stage process. This three stage process is derived from with the general three-stage framework of judgement formation which is prominent in social cognition literature (Gilbert, 1998).

The three stages devised by Gauthier et al. (2016) are illustrated in Figure 2.2 and discussed below.

Figure 2.2 *Three stage framework of assessors' judgement formation (Gauthier et al., 2016, p.516)*



1. <u>Observation</u>: this phase is concerned with "how raters attend to and actively select information about trainees and their performances" (p.513). Within this phase, there are several mechanisms that operate: *generation of automatic impressions* about the student, *formulating high-level inferences* and *focusing on different dimensions of competencies.* Empirical work related to this phase is examined in more detail in section 2.6.1.

2. <u>Processing</u>: this phase is concerned with "how raters retrieve and use contextual information and prior knowledge to inform and compare the performance at hand" (p.513). Having observed aspects of a performance, it is now up to assessors to make sense of that performance. Within this phase, the interpretation of a performance happens in relation to: assessors' *personal concept of competence, comparison with various exemplars,* and is affected by *task and context specificity.* Empirical work related to this phase is examined in more detail in section 2.6.2.

3. Integration: this phase is concerned with "how raters combine different sources of information to form an overall judgement" (p.513). Having observed a student's performance and made sense of what they have witnessed, assessors now have to come to a judgement or decision about that student's level of competence. The mechanisms operating in this phase are: *weighting and synthesising information differently, producing narrative judgements,* and *translating narrative judgements into scales.* Empirical work related to this phase is examined in more detail in section 2.6.3.

The identification of the three stages involved in assessors' judgement formation, and the relevant mechanisms within each stage, highlight the various different ways that assessors might form divergent judgements about the same student's performance. Although the performance itself is fixed, the nine different cognitive mechanisms identified by Gauthier et al. (2016) are "internal" and unique to each assessor; meaning a range of idiosyncratic phenomena affect how a single performance is regarded by an assessor.

This three-stage process provides researchers working within the field of assessor cognition a framework within which to situate various pieces of work. Instead of examining "assessor cognition" as a general phenomenon, researchers can formulate research questions pertaining to a specific stage of the process, or a mechanism within a stage (Gauthier et al., 2016). This three-stage process is used in this literature review to structure the appraisal of previous work in assessor cognition, and identify relevant gaps in the literature from which the research questions will be formulated (see section 2.6).

2.5.8 *Summary of theoretical perspective*

Using social cognition as a lens through which to view the judging of students in performance assessments, the role of the assessor can be conceptualised as a three stage task of *observation, processing* and *integration* (Gauthier et al., 2016). Within each stage, there are a range of different mechanisms, unique to the assessor, which affect how a performance is judged, which begins to explain how a single OSCE performance could be evaluated in numerous different ways, depending on the assessor.

The process of judging student performances is cognitively demanding (Paravattil & Wilby, 2019); however, assessors with a higher level of expertise are less likely to experience cognitive strain during the assessment process, and can rely on System 1 processing when making sense of a performance. As such, the expertise of the assessor, as well as their prior knowledge and experiences, has significant implications for how a performance is evaluated. Understanding assessors' judgement formation as a complex process influenced by a range of personal factors inevitably raises questions about scoring consistency. If assessors are highly idiosyncratic, this might affect their ability to judge the same performance in a homogenous way. In section 2.6, empirical research on assessor cognition and inter-rater reliability is critically analysed, and key conclusions are presented. The structure of the empirical section mirrors the three-stage process outlined above; as such, it follows Gauthier et al.'s (2016) call to situate research on assessor cognition within this temporal framework.

2.6 *Empirical research on assessor cognition*

Assessor cognition refers to research which "focuses on the investigation of assessors' cognitive processes and their impact on assessment quality" (Gingerich et al., 2014a, p.1055). Work in this field has proliferated in the last decade, in recognition of the fact that in spite of extensive research on how to improve test score reliability, numerous authors have noted that it is frequently at a less than ideal level (e.g., Brannick et al., 2011; Navas-Ferrar et al., 2017; Goh et al., 2019). While reliability varies due to several factors, it is noted that variance between assessors often accounts for the highest level of overall score variance, occasionally accounting for more variance than actual differences in students' performance levels (Gingerich et al., 2014a, p.1056). In other words, differences in how assessors interpret a performance can sometimes result in more score variance than actual differences in performance levels. Because of this, the field of assessor cognition shifts the focus of inquiry to the assessors themselves, with the goal of determining how exactly assessors interpret and understand examinee performances in medical assessments.

This section examines the empirical research that has been conducted in assessor cognition. Using the three-stage framework devised by Gauthier et al. (2016) and outlined in section 2.5.7, research is critically analysed under the headings of *observation, processing,* and *integration*. Within each of these three stages, the tentative conclusions that can be drawn from the body of work are discussed, focusing in particular on the potential IRR effects that may stem from a specific cognitive process. Throughout the section, notable gaps in the literature are detailed. These gaps in the literature form the basis for the research questions that are outlined in the following chapter. It is noted that the three-stage framework is used as a way of structuring the discussion of assessor cognition research in this chapter. As such, some articles recur at various points of the discussion, as the different conclusions they make relate to the three different stages. It is also worth noting at this stage that, due to the nascent state of assessor cognition as a research focus - the term was coined a decade ago - there are notable opportunities for further research:

- Assessor cognition research has mostly taken place using Workplace Based Assessments (WBAs) such as the Mini-Clinical Evaluation Exercise, where postgraduate medical students are assessed interacting with real patients in an authentic clinical setting. Research which specifically focuses on the OSCE is less common.
- Almost all work on assessor cognition has been done in the field of medicine, with little published work examining nursing OSCE assessors. While there may be overlap between how assessors in medicine and nursing assess student performances, this will remain unknown unless explicitly investigated.
- Numerous authors working in the field of assessor cognition have noted the potential for assessor cognition research to explain how issues with IRR arise (e.g., Kogan et al., 2011; Yeates et al., 2013). However, the majority of studies have demonstrated assessors' cognitive processes collectively across a sample of assessors, rather than juxtaposing how different assessors vary in their approaches to judgement formation. As a result, the specifics of how score variance arises within a given sample of assessors are under-researched.

2.6.1 *Observation*

In the observation stage, assessors watch students giving performances, and select the information about that performance that they deem to be important (Gauthier et al., 2016). As discussed in the previous chapter, the marking guide designed for a specific OSCE/OSCE station can be conceptualised as an attempt to focus assessors' attention on the salient aspects of a student's performance (Khan et al., 2013a). Rather than giving assessors free rein to determine what they wish to focus on, the marking guide explicitly instructs assessors to look out for specific pieces of information from each student. As such, it is a key mechanism that influences assessors during the observation process. In theory, a well-designed marking guide should unify the observations of different assessors, so that they all focus on the same things. However, research has indicated that:

- In spite of the marking guide, assessors may judge students according to what assessors themselves think is important.
- Assessors may be prone to making inferences about students that go beyond what is directly visible in a particular performance.

The empirical evidence for these claims, and their potential implications for nursing OSCEs, are discussed below.

2.6.1.1 *Judging students according to personalised criteria*

Research in assessor cognition has investigated the phenomenon that, in spite of the marking guide which explicitly instructs assessors how to structure their observations of student performances, assessors may use their own personal criteria when observing students, looking out for what they believe to be important. Kogan et al. (2011) conducted an influential early piece of research into assessor cognition, with the aim of developing a framework for how medicine assessors make judgements about students. In their study, 44 assessors watched video-recorded performances of internal medicine students completing a clinical encounter with a standardised patient. Assessors completed a mini-CEX (an assessment format in which staff members assess a trainee doctor's interaction with a real patient) marking guide while watching the videos, and underwent a semi-structured interview after the video was completed. Using grounded theory, the authors analysed the interview transcripts in order to model the cognitive mechanisms that affect assessors' judgements.

They found that assessors have variable frames of reference when judging performances, which affects the observation stage, as assessors look out for different aspects of a performance. For example, one of their participants noted that "The first thing I always pick is the interpersonal communication portion because it just happens to be 90% of what our job is" (p.1051). This participant is explicit about the fact that they choose to focus their observation largely on the interpersonal component of performance, as they believe this to be the most important aspect of effective clinical practice.

This selective tendency of assessors during the observation phase has been noted elsewhere in assessor cognition research. Yeates et al. (2013) sought to determine the mechanisms by which score variance between raters occurs. They conducted a qualitative study in which 12 assessors watched videos of trainee doctors taking a patient history. Assessors were asked to "think out loud" while watching, and were then interviewed afterwards. One of the themes that emerged from the authors' analysis was *differential salience*: assessors did not weigh every aspect of a performance uniformly. Rather, "what struck one assessor as important about a given performance varied from what struck a different assessor" (p.331). As such, assessors focused their observations on what they perceived to be most pertinent about a performance. The authors noted that this phenomenon was likely to have implications in terms of IRR; however, the purely qualitative approach they adopted did not allow for a calculation of whether awarded scores were actually divergent.

Similar findings were reported in studies by Chahine et al. (2016) and Roberts et al. (2020). These two articles will be discussed in more detail in subsequent sections, but it is noted that in both studies, the authors reported that assessors have idiosyncratic ways of observing students' performances. Chahine et al. (2016) recorded a "hidden pattern" in how assessors observe students, with two competencies (*investigation and management* and *counselling*) afforded more attention, even when they are equally weighted in the marking guide with other competencies. Likewise, Roberts et al. (2020) wrote that "For some assessors, expectations regarding the student's responses were not directly linked to the rubric criteria" (p.9), meaning that assessors used their own ideas of what was important about a performance to guide their observations. The prevalence of studies which have documented that assessors direct their observations based on criteria exogenous to the marking guide raises the question of what they actually believe to be important within an OSCE performance, and whether assessor subjectivity (and problems with IRR) can be reduced if the design of the marking guide is altered.

On this topic, Hyde et al. (2020) conducted a study in which 12 experienced OSCE assessors were interviewed about what they look out for when judging a student performing at an OSCE station. In addition to this, they were asked to develop their own marking guides to reflect what they think should be present. From their data analysis, they found that assessors consistently referred to five "domains" of effective performance: application of knowledge, manner with patients, getting it done, safety, and overall impressions. The authors noted that these five domains are distinct from those generally contained within OSCE marking guides. Remarking on this, they commented that their findings highlight "the complexity of applied judgement… confirming that these experienced and trained assessors do not mechanically apply" marking guides (p.13). This study indicates that when assessors deviate from the marking guide, it is not simply because they are not able to apply it in the correct way; rather they feel that they have seen the student exhibit something important that has been overlooked in the guide. As such, they are explicit about applying their own criteria to the assessment process. These findings suggest that score variance between assessors may arise due to the design of the marking guide failing to take into account what assessors believe to be important.

A recent study by the same authors (Hyde et al., 2021), in which 31 assessors across three countries were interviewed and subjected to a think aloud protocol, further adds to the idea that assessors tend to judge students based on their own unique set of criteria. The authors found that assessors perceived the OSCE to be an "inauthentic" assessment, and that as a result they were willing to "deviate from an organisation's instructions" (p.9) in order to reward students that they believed had performed well. In other words, assessors were honest about their tendency to deviate from the marking guide in order to observe what they believed to be important in an OSCE performance. This has obvious implications in terms of IRR: given the documented disparities in what assessors look out for when judging students (e.g., Chahine et al., 2016), it is likely this deviation from the guide would result in the same performance being awarded different scores. However, the authors did not collect any quantitative data in the form of completed marking guides, so any potential effects in terms of IRR are merely speculative.

As noted previously, the majority of studies conducted in assessor cognition have taken medicine OSCEs (and, therefore, medicine assessors) as their point of study. Assessor cognition research which focuses on nursing is much rarer. The one identified study which sought to determine the judgement processes of nursing OSCE assessors was conducted by East et al. (2014). As part of a larger, mixed methods study, they interviewed 25 OSCE assessors from an Australian undergraduate program in order to ascertain how they make judgements regarding students' clinical competence. They found that "rather than assessing students using the grading assessment criteria…overwhelmingly assessors determined students' competence subjectively" (p.463). Assessors generally chose to focus on whether the students were performing the required tasks safely and confidently, irrespective of what the marking guide stipulated they were supposed to look out for.

The tendency for medicine assessors' observations of student performances to be affected by what they perceive to be important is, therefore, well-documented, and there is some evidence that this phenomenon may be present in nursing assessors as well. However, in spite of frequent comments that this may affect the IRR of awarded scores, almost none of these studies collected and examined quantitative data in order to explore IRR levels. As such, while it is possible to speculate about how these idiosyncrasies in assessors' judgements may have affected score variance, this has yet to be investigated in detail, particularly in relation to undergraduate nursing OSCEs.

2.6.1.2 *Inference formation*

Another key factor which affects the observation stage of assessment is the formation of inferences about a student. In their 2011 study, Kogan and colleagues determined that assessors were prone to extrapolating information about a student that went significantly beyond what was visible to them. Assessors' inferences about students were often "high level", in the sense that there was "significant interpretation based on the behaviour witnessed" (p.1053). However, there was minimal consistency between assessors as to whether inferences were made about a specific student, and if they were made, what these inferences were likely to be. For example, in one of the recorded performances used in their study, one assessor made an inference that the trainee doctor was "stiff... uncomfortable and embarrassed", while another noted that he "wasn't shy" and was "very comfortable" (p.1054). As such, these two assessors observed the same thing (a recorded interaction between a patient and a trainee doctor) and used it to make two opposing inferences about the trainee. The potential effects of this inconsistency in terms of IRR are perhaps obvious: the trainee doctor would be much more likely to receive a higher grade from the assessor who perceived him as being "very comfortable" than the one who described him as "uncomfortable". However, give that Kogan et al. took a purely qualitative approach, this link between the process of inference formation and IRR can merely be suggested.

The tendency of assessors to make inferences about students has been noted elsewhere in literature on assessor cognition. St-Onge et al. conducted a 2016 study in which 11 experienced medical assessors were interviewed about how they come to a decision about trainee doctors' performance levels. One of the key mechanisms they noted at the observation stage was the formation of inferences – assessors were liable to "provide hypothetical explanations" (p.637) of a trainee's behaviour. This was most commonly done in order to explain why a trainee had failed to do something in the correct way: rather than just note that they had got something wrong, some assessors made inferences in order to justify that trainee's behaviour. It is likely that the assessors who made these inferences would award higher marks to that trainee. However, in common with the study by Kogan et al. discussed above, the authors did not collect score data from participants, and so could not test this link between inferences and IRR empirically.

The question of whether assessors making inferences about students is defensible is debated in the literature. As noted by Gauthier et al. (2016, p.516), some authors such as Kogan have described inferences as a “problem to be addressed”. This is perhaps logical considering their findings that assessors made directly contrasting inferences while watching the same recorded performance. The assessment ramifications of this tendency are clear: if different assessors make different inferences about the same student, this could affect the consistency of marks awarded to that student. However, other authors such as Govaerts et al. (2011, 2013) have written about inference formation as justifiable, particularly when assessors are experienced – such assessors are able to leverage their experience and make inferences about a student that are accurate. As such, these contrasting views on inference formation “highlight a need to investigate how this mechanism functions” (Gauthier et al., 2016, p.516) as well as its effects on score reliability.

2.6.1.3 *Conclusion*

Research into the observation stage of assessors’ judgement formation has indicated that while assessors use marking guides to structure their observations of student performances, there is a documented tendency of assessors focusing on what they believe to be the most important aspects of performance. This may be the case even if all aspects are weighted of performance equally in the marking guide. Because assessors may not agree on what these important aspects are, they may direct their observations towards completely different parts of a single performance. Additionally, researchers in assessor cognition have documented the phenomenon of inference formation at the observation stage, in which assessors make (often divergent) inferences that explain student behaviours. Authors writing on this subject have frequently discussed the potential for these divergences in how assessors judge students to affect the scores they award to them. However, this link is under-researched, primarily due to the purely qualitative studies conducted thus far. The relevance of these findings to the current research study are unpacked further in the general conclusion at the end of this chapter (section 2.7).

2.6.2 *Processing*

In the processing phase, assessors give meaning to the acts they have witnessed during a performance, and make sense of these acts in relation to their own prior knowledge or experiences (Gauthier et al., 2016). Research has indicated that assessors:

- Are liable to interpret a student’s performance by comparing it to a range of different exemplars, such as the student who came immediately before, or to what they would do themselves in a specific situation.

- Have their own personal conception of what “good” performance looks like, affected by their own knowledge or experiences.

The empirical evidence for these claims, and the potential implications in terms of OSCE IRR, are discussed below.

2.6.2.1 *Comparison with exemplars*

Yeates et al. (2015) conducted a quantitative study with the aim of investigating whether the grades awarded by assessors are affected by the students they judged immediately before. Their hypothesis was that seeing a particularly good performance might make the subsequent performance seem less good by comparison (and vice versa). They analysed score data from two large scale performance assessments in medicine - the clinical assessment of the 2011 United Kingdom Foundational Programme and a multiple mini-interview at the University of Alberta. Each score was compared with the score given on the same station directly before, as well as with an average of the three previous scores. They found that there was a small, but significant, negative relationship: indicating a “contrast effect” which accounted for up to 11% of score variance. This relationship was stronger when the score on a specific station was compared with the average of the three previous scores on that station, indicating that assessors likely form an amalgamation of the previous performances and judge a candidate against this. They wrote that OSCE assessors lack “any truly fixed sense of competences against which to judge” candidates, and therefore rely on previous candidates to guide their appraisal (p.978).

A similar study by Wood et al. (2018) determined that medicine assessors are overly influenced by their initial impressions of a candidate. If a candidate begins an OSCE station poorly, they are unlikely to receive a good grade, even if they subsequently perform at a high level. This again suggests that assessors tend to anchor their impression of a student in a way that prevents a truly objective judgement of that student. These findings are broadly in line with research from social cognition, which emphasises the relativity of judgement formation (Kahneman, 2011); and suggests that assessors do not apply a “clean slate” to every student they judge.

This phenomenon has also been investigated qualitatively. The study by Kogan et al. (2011), discussed in the previous section, determined that assessors were prone to comparing students to a range of different performances. The assessors who participated in their study reported interpreting student performances based on what they (the assessors) would do if they were in a situation (“I keep thinking…Would I expect this of myself? Hence is it fair to expect this of someone junior to me?” (p.1051)), as well as what other doctors they know would do. This finding was echoed in the study by St-Onge et al. (2016) discussed in the previous section. They found that assessors "continually compared the observed performance" (p.636) to other students and to what they (the assessors) would do themselves in a given situation. Overall, research suggests that assessors draw on a range of different comparisons when interpreting a student's performance, and that this affects the scores awarded to students. This has the potential to threaten IRR levels when an assessment such as an OSCE is delivered, as a student's score may be affected by whoever happened to be taking the OSCE before them.

2.6.2.2 *Personal conceptions of competence*

In addition to drawing comparisons, assessors' impressions of performances are also influenced by personal factors, such as prior knowledge and experiences. The impact of these idiosyncratic factors is that different assessors have unique benchmarks against which students are judged. One of the themes that emerged from the study by Yeates et al. (2013) – discussed in detail in the previous section – is *criterion uncertainty,* namely that assessors lack a fixed sense of what ideal or average performance is supposed to look like, and, consequently, judge students in relation to their own personal conceptions of competence.

This theme recurs in multiple studies of assessor cognition. Roberts et al. (2020) conducted a mixed-methods study in order to determine the extent to which assessor cognition contributed to score variance in an undergraduate occupational therapy OSCE. In their study, 95 occupational therapy students took an OSCE, and were judged on seven domains of performance (on a scale of 1-7) and given an overall score out of ten. Eight assessors took part in a think-aloud protocol in which they viewed six performances (two poor, two medium, and two good), for a total of 48 interviews. From these interviews, the authors found that "assessors' frames of reference or educational perspectives were variable and influenced their expectations during the assessment" (p.8). Assessors were liable to be influenced by their own personal history (e.g., their clinical experience), as well as their knowledge of the curriculum. Because of these factors, assessors' "expectations regarding the student's responses were not directly linked to the marking criteria" (p.9). As such, their study seems to suggest that there was a lack of clarity among assessors as to how a specific performance should be interpreted; and, as a result, assessors used their own idiosyncratic means of determining how well or badly a student had performed. This idiosyncrasy was likely to result in high levels of observed score variance.

A similar study by Govaerts et al. (2013) sought to use a framework of social perception (studying how individuals form judgements about others in a social context) to determine the "processes underlying judgement and decision-making in performance assessments in the clinical setting" (p.377). In their study, 34 assessors (18 experienced and 16 less experienced), watched two videos of sixth-year medical students in a patient encounter. Participants were subject to a think-aloud protocol where they were asked to vocalise their thoughts about the student's performance. In common with other work in this area, the authors found that assessors have idiosyncratic ways of interpreting performances, and form their own "person schemas" about what they think the student is like. They explicitly link this differential formation of schemas to score variance: "Differences in the way raters form person schemas in WBA contexts may therefore be one of the major factors underlying differences in rating outcomes" (p.392). This study further illustrates the tendency of assessors to interpret student performances based on a personalised set of criteria, and the authors were explicit about the fact that this phenomenon is likely to adversely affect IRR.

Likewise, when conceptualising internal and external factors affecting judgement formation, St-Onge et al. (2016) determined that assessors are liable to rely on "external sources of information" when judging students. These factors are unique to each assessor, and in addition are liable to change over time as assessors build up expertise. As such, the findings from these studies further indicate that assessors' judgements of students may be heavily influenced by their own histories (rather than being strictly based on a shared understanding of the marking guide).

While the majority of relevant work in this area has taken place in the field of medicine, the study of nursing assessors, conducted by East et al. (2014) and discussed above, is also illustrative of the variability in how assessors interpret student performances. One of the themes that emerged from their qualitative interviews with 25 nursing assessors is that assessors drew on their own experiences when making sense of student performances. The authors noted that "Some participants revealed that their own clinical practice and experiences had the ability to influence their decision making" (p.465). Assessors reported that it was easier to bring to mind their own clinical experience while assessing students, rather than an abstract (but shared) idea of good performance. This study again suggests that nursing assessors, in common with their counterparts in medicine, draw on personal ideas of competence when interpreting OSCE performances. However, the limited scope of this single study means that the question of whether this takes place in nursing assessors more generally is unanswered. Additionally, and in common with the studies discussed in the previous section regarding assessors' observations, research relating to assessors' processing of performances has generally opted to map cognitive processes across a sample of assessors. This is a necessary first step when conducting research into assessor cognition, and allows for an indication of how subjective judgements emerge. However, they do not drill down into the specifics of how individual assessors judge and grade student performances. Such an approach would allow for a juxtaposition of how specific assessors differ from each other, which is the root of how IRR threats emerge.

2.6.2.3 *Conclusion*

Research into the processing stage of assessors' judgement formation has indicated that assessors interpret student performances by comparing them to a range of different exemplars, such as the students that have come directly before, or what they would do in the same situation. Additionally, several empirical studies in assessor cognition have documented that assessors' own ideas of what "good" performance entails are influenced by a range of idiosyncratic factors. As such, what appears to one assessor to be a competent performance may not appear to be competent by a different assessor. The implications of this idiosyncrasy for score reliability are notable: the more an assessor draws on personal factors when assessing a student, the less likely they are to award the same grades as their colleagues. The relevance of these findings to the current study are unpacked further in section 2.7.

2.6.3 *Integration*

In the integration phase, assessors "combine different sources of information to form an overall judgement" about a student's performance level (Gauthier et al., 2016, p.513). Research that specifically focuses on this stage of the assessment process is less common than the other two phases (Gauthier et al., 2016). Nonetheless, initial research has indicated that assessors:

- May have differing methods of coming to a final decision about a student's performance level, affected by their own opinions about the assessed skill(s).
- Largely think of performance in narrative terms, and may have difficulty deciding on a numerical score when asked to do so.

2.6.3.1 *Overall judgement formation*

Several studies in assessor cognition have noted that, when deciding on an overall score for, or making a final judgement about, a student's performance, assessors are likely to have divergent ideas about what elements of that performance should contribute most to the final judgement.

This may be true even if all elements of performance are equally weighted on a given marking guide (Chahine et al., 2016). For example, St-Onge et al. (2016) detailed that some assessors believed that specific items on the marking guide were so important that if a student failed to complete them properly, they would receive a low grade, even if the rest of the performance was acceptable. Similarly, Kogan et al. (2011, p.1053) noted that a factor affecting assessors in their study was “variable approaches to synthesising judgements”: assessors lacked a unified strategy for determining the extent to which different components of performance should contribute to the final score awarded to a student. Some assessors had more objective strategies, for example averaging the scores across different domains of performance and using this as the final score; while others gave a higher weighting to competencies that they perceived to be the most important. The IRR implications of this are notable: the final score awarded to a student would likely be influenced by a specific assessor’s idea about what elements of performance are most important within a certain skill.

Assessors also may have difficulty translating their narrative judgements about a performance into a numerical score on the marking guide. For example, Yeates et al. (2013) noted that the assessors in their study tended to think about an observed performance qualitatively, and described it as such when discussing it. This phenomenon of assessors forming overall judgements about students and then having to retroactively award scores has been noted elsewhere in the literature (e.g., East et al., 2014). As documented by Gauthier et al. (2016), empirical studies which explicitly investigate cognitive processes at the integration stage are uncommon. As such, any conclusions from this nascent body of work should be treated as tentative. However, studies (e.g., Gauthier et al., 2013; St-Onge et al., 2016) have similarly suggested that when viewing student performances, assessors initially form overall judgements about a student, and then go back and complete the marking guide based on this judgement (rather than completing the marking guide as they go). The issue of what factors influence these overall judgements, and how they may affect the scores awarded to students, is one that is under-researched, particularly in the field of nurse education. As such, there is scope for an investigation into how undergraduate nursing assessors integrate all the information available to them about a student OSCE performance and use it to award a final grade, and whether divergences in how this integration takes place have implications in terms of IRR.

2.6.3.2 *Linking processes with scores*

As noted throughout section 2.6, initial research into assessor cognition has documented a range of factors, unique to an assessor, that can affect their judgement of a student's performance. It is perhaps intuitive that the prevalence of these idiosyncratic factors may have implications for IRR, as different assessors may lack a unified method of judging the same performance, and thus award divergent scores. These potential implications have been flagged throughout this section. However, research which has explicitly attempted to investigate this link is rare. The majority of studies cited thus far have sought to document the cognitive processes that take place across a sample of assessors, rather than exploring differences within these samples. This is notable as it fails to get at the root of how discrepancies between assessors arise: assessors having divergent opinions about, and therefore awarding different grades to, the same performance. As such, without research that seeks to link assessors' cognitive processes with outcomes in terms of awarded scores, our understanding of how IRR issues emerge is limited.

Several high-quality studies in medical assessment have begun to investigate this issue. These studies have focused on dividing assessors in a sample into sub-groups based on commonalities in how they view student performances, and calculating what percentage of score variance is attributable to membership of a specific group. Gingerich et al. (2014b) used a combination of narrative judgements of recorded performances, completed Mini-CEX marking guides, and Latent Partition analysis to determine how many "patterns" of judgement existed within a sample of 48 medicine assessors, as well as the amount of score variance attributable to assessors having specific patterns of judgement. Underlying their study was the assumption that while "as many social judgements could be formed about an individual as there are people providing judgements", it is more likely that there is not an infinite amount of such judgements (p.1510).

Overall, the authors found that while assessors did have divergent judgements for individual performances, there were only between two and five judgement patterns for each one. As such, they concluded that there were multiple "signals" in the "noise" of assessor variability. Additionally, they found that assessors who made the same type of judgement regarding a performance awarded similar scores to that performance, and that the partitioning of judgements accounted for between 9% and 57% of score variance (depending on the video). Thus, their results suggested that assessors within a sample formed distinct and identifiable overall judgements about student performances, and that some amount of score variance was attributable to these different judgements.

Studies by Chahine et al. (2016) and Gingerich et al. (2017) extended research in this direction. Chahine et al. (2016) conducted a mixed methods study in order to determine how OSCE raters valued various competencies - equally weighted in the marking guide - differently, and how this weighting affected a candidate's score. Using Ordinal Logistic Hierarchical Linear Modelling on data from a postgraduate OSCE, combined with cognitive interviews of four assessors, they determined that two distinct groups of assessors could be identified: those who valued the skill *investigation and management* most highly, and those who valued *counselling*. Thus, echoing the above study, they wrote that although assessors make varied assessments, there is "a limited number of patterns of social judgements" (p.619), and that these patterns can be linked to specific outcomes in terms of student scores: the assessors who viewed the performance in the same way were likely to award similar scores. From this study, it is suggested that IRR issues may arise due to assessors within a sample having two divergent opinions about what aspects of a performance are most important, and allocating marks according to these opinions.

Finally, a study by Gingerich et al. (2017b) used Q-sort methodology to identify different patterns of assessor judgements. In their study, 46 participants watched and assigned scores to between one and four videos of different clinical performances. After watching each video, they were asked to sort 44 statements about the video (e.g., *appeared compassionate or showed genuine concern*) into a forced normal distribution ranging from "most consistent with my perspective" of the physician to "most contrary" (p.823). The 44 statements were grouped into those emphasising a rapport-building approach, those emphasising medical expertise, and those signifying social judgements or inferences. Examining how the statements were sorted revealed "two major clusters of consensus" (p.828), meaning that there were two dominant ways that assessors viewed each performance: they either viewed rapport-building skills or medical expertise as the most important. Additionally, the assessors who valued rapport-building skills tended to award similar marks as each other, as did those who valued medical expertise. Accounting for the membership in a specific cluster accounted for between 21% and 53% of the variance in score, depending on the video (p.835). Similar to the study by Chahine et al. (2016), the results of this study suggest that IRR may be threatened due to a split within an assessor pool regarding the most important elements of effective performance.

These three studies were able to uncover a measurable link between the processes that assessors went through when judging a performance, and the scores awarded to that performance. As a result, they went a step further than purely qualitative studies which sought only to map assessors' cognitive processes, or purely quantitative investigations into OSCE reliability levels. By employing mixed methods, they were able to show specific means by which IRR is threatened during performance assessments. However, although the studies discussed in this section are of high-quality and have been widely cited, it is noted that the results of three studies, conducted with medicine assessors, is clearly insufficient to draw any conclusions about the behaviour of nursing assessors. As such, the question of how nursing assessors form final judgements about students, and whether identifiable patterns of judgement can be linked with awarded scores, is unanswered. Additionally, while the focus on identifying "clusters" within a sample of assessors is relatively novel from a research perspective, such an approach might have limited utility in terms of the institutional needs of (for example) a university department which administers OSCEs. As noted by numerous authors (e.g., Bartman et al., 2013; Dunbar, 2018), the goal of IRR analyses is often to identify particularly harsh or lenient assessors operating within a given organisation. The "clustering" approach described in the studies above would not allow for an identification of outlier assessors and an understanding of how they approach the task of assessment.

2.6.3.3 *Conclusion*

Research into the integration stage of assessors' judgement formation has noted that assessors may have different methods of forming an overall judgement about a student's performance, and additionally have different strategies for translating these overall judgements into numerical scores. A burgeoning area of research has attempted to link these judgement formations with awarded scores, by identifying "clusters" (Gingerich et al., 2014b) of assessors. The relevance of these findings to the current study are addressed in the subsequent section.

2.7 *Summary and conclusions*

Since their inception in 1975, OSCEs have proliferated in nursing, and are now a popular assessment format, performed in at least 33 countries worldwide (Goh et al., 2019) at the time of writing. As with all assessments, it is important to continually document evidence that the decisions made on the basis of OSCEs scores are valid (AERA et al., 2014; Khan et al., 2013b). To that end, significant research has been done which seeks to ensure that this is the case.

Recognising that reliability is an important source of validity evidence (Newton, 2017), an extensive body of work has sought to improve the reliability of awarded OSCE scores. However, large-scale systematic reviews of the reliability of nursing OSCEs have indicated that IRR in particular is often overlooked when the reliability of OSCE scores is determined; and, when it is calculated, is frequently at levels classified as "moderate" rather than "strong" (Navas-Ferrar et al., 2017; Goh et al., 2019). As such, it is imperative that researchers make efforts to document the levels of IRR present in undergraduate nursing OSCEs, and make attempts to account for any IRR issues that may arise. Additionally, when IRR has been investigated, published work has frequently failed to focus on the assessors themselves, and how they engage with the process of judging and awarding grades to students. Such an approach might allow for a novel perspective on how IRR issues emerge and how they could be tackled.

However, researchers have also critiqued an excessive or exclusive focus on reliability as potentially undermining the validity of performance assessments. There is an argument in the literature that assessors should be given the space to exercise their "deep expert judgement" (Eva & Hodges, 2012; p.917) when examining students. This conversation has mostly taken place in medical assessment, however the potential of subjective judgements to improve assessment quality has also been discussed in the nursing literature (e.g., Mitchell et al., 2009). The question of the extent to which subjectivity can or should be incorporated into undergraduate nursing OSCEs is a pertinent one.

Assessor cognition emerged as an area of research in part because of the persistent reliability problems with performance assessments in the medical sciences. Assessor cognition is underpinned by the idea that it is unrealistic to expect complete agreement between assessors at all times, and that differences between assessors should be investigated and understood in greater detail. This chapter critically examined empirical research pertaining to the *observation, processing* and *integration* stages of assessors' judgement formation (Gauthier et al., 2016). Within each of these three stages, a range of cognitive processes take place, which affect how student performances are understood by assessors, and the grades they award. However, the majority of studies in assessor cognition have taken place in the field of medical assessment, with only a single study identified in nursing specifically (East et al., 2014). As a consequence, the question of whether assessors in nursing go through the same cognitive processes when judging OSCE performances is one that has not been investigated in detail. Such an investigation would strengthen the validity argument associated with these OSCEs.

Additionally, most published work in assessor cognition has used postgraduate level performance assessments (e.g., Gingerich et al., 2014b), or workplace-based assessments in which trainee doctors are assessed interacting with actual patients, as their point of study. There is an assumption that the findings from these studies may be transferable to the undergraduate context. However, this assumption remains under-investigated, and there is scope for a study that explicitly aims to map the cognitive processes of undergraduate OSCE assessors.

Due to the qualitative nature of most published work on assessor cognition, empirical studies have used small and geographically specific sample sizes (often around 10 assessors working at the same institution). As such, the generalisability of these findings across different contexts is unknown. It is entirely possible that assessors in Ireland approach the task of assessment in a way that is meaningfully different than their counterparts in North America (where most of the cited studies have taken place). Indeed, authors such as Chahine et al. (2016, p.619) have urged researchers to conduct assessor cognition studies in other contexts.

In terms of methodology, the use of a combination of interviews with assessors and "think alouds", in which assessors watch recorded performances and describe their thoughts, have led to rich descriptions of assessors' cognitive processes in several studies (e.g., Yeates et al., 2013, St-Onge et al., 2016). However, as noted in section 2.6, the majority of this work has been purely qualitative, mapping assessors' idiosyncrasies across a sample. There is an assumption in the literature that such idiosyncrasy threatens score reliability, however attempts to test this connection are uncommon. A small but notable body of research has attempted to use mixed methods in order to link assessors' cognitive processes with measurable outcomes in terms of awarded scores (e.g., Gingerich et al., 2014b, 2017), noting that there are identifiable clusters of assessors who award similar marks to each other. Given that one of initial aims of assessor cognition research is the improvement of score reliability, it is important to explore further any potential links between *how* assessors assess students and the *results* in terms of awarded scores.

Chapter Three

Methodology

3.1 *Introduction*

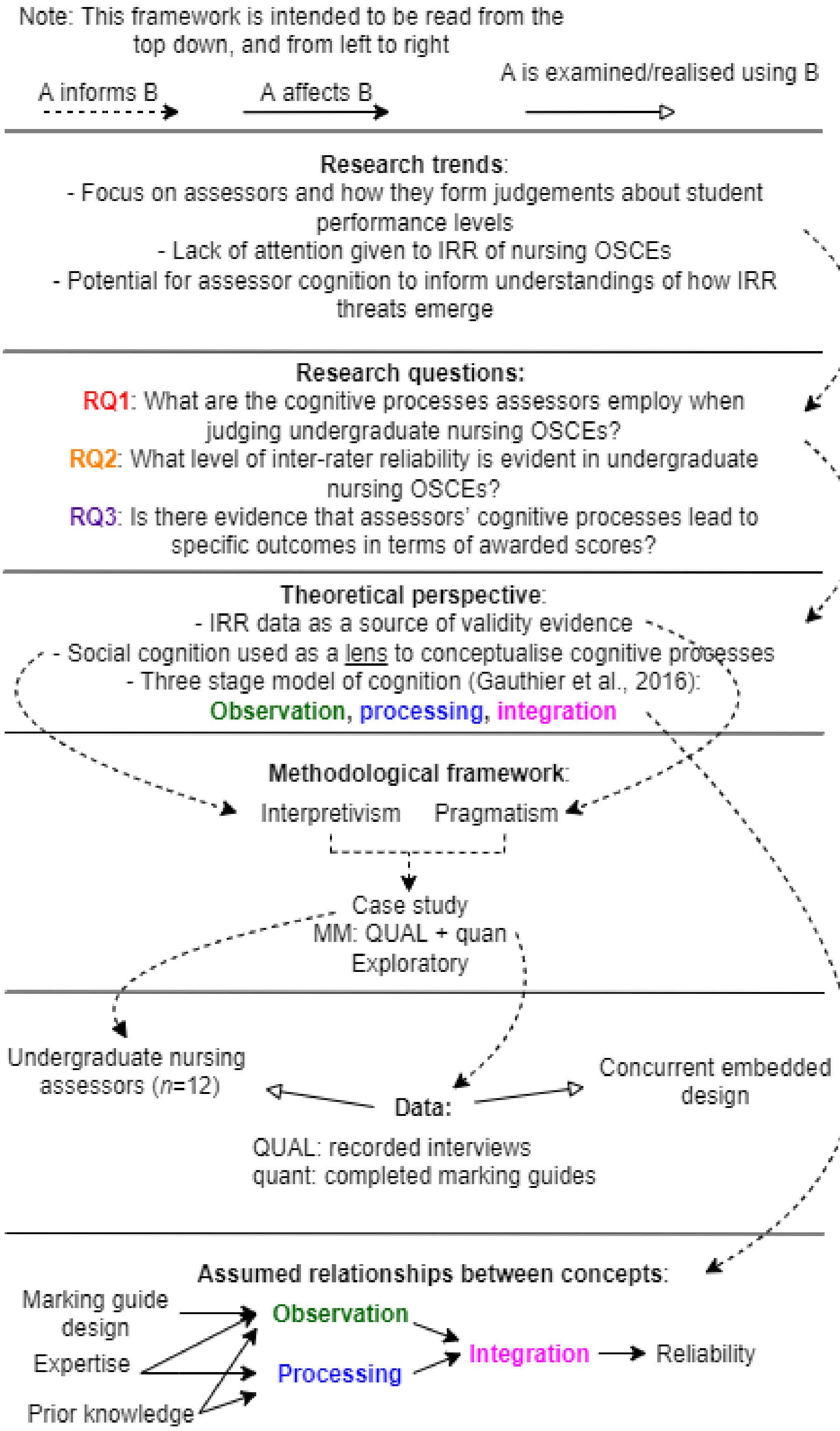
This study explored the cognitive processes that assessors employ when judging undergraduate nursing OSCEs, and how these processes impact the reliability of awarded scores. This study used a mixed methods approach, with data collection taking place from September 2021 – April 2022. This chapter describes in detail the methodology that was employed to carry out this study. Firstly, the conceptual framework and research questions are outlined, followed by an overview of the research design and research methods. Sampling techniques are then described, followed by data collection and analysis procedures. Finally, all relevant ethical considerations are discussed.

3.2 *Conceptual framework*

A conceptual framework is defined by Maxwell (2005) as "the systems of concepts, assumptions, expectations, beliefs and theories that supports and informs your research" (p.222). The conceptual framework is an illustration of the various ideas that influence the research design. A clear conceptual framework indicates to the reader the overall shape of the project, as well as how different elements of the study inform and affect each other. Although the conceptual framework is often derived from the literature review, it is important to note that it is not merely a summary of what was discussed in the literature; rather it goes beyond this, and is a map of how the different concepts that were discussed in the literature review are related, and how they influence the study (Maxwell, 2005). As such, the conceptual framework on the following page is intended to convey how the issues raised in the literature review (and summarised in section 2.7) affected the design of the research study. The key concepts in the framework are briefly mentioned here, before a more detailed discussion throughout sections 3.3 and 3.4.

The framework is intended to be read from the top down. The framework begins with notable *research trends*, identified in the previous chapter. From each of these three trends, a *research question* is derived relating to an unanswered or under-researched topic. As discussed in Chapter Two, the decision to focus on assessors' cognitive processes necessitated the articulation of a *theoretical perspective* within which the research could be situated.

Figure 3.1 *Conceptual framework*



Note: This framework is intended to be read from the top down, and from left to right

Research trends:

- Focus on assessors and how they form judgements about student performance levels
- Lack of attention given to IRR of nursing OSCEs
- Potential for assessor cognition to inform understandings of how IRR threats emerge

Research questions:

RQ1: What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?

RQ2: What level of inter-rater reliability is evident in undergraduate nursing OSCEs?

RQ3: Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?

Theoretical perspective:

- IRR data as a source of validity evidence
- Social cognition used as a lens to conceptualise cognitive processes
- Three stage model of cognition (Gauthier et al., 2016):

Observation, processing, integration

Methodological framework:

Interpretivism
Pragmatism
Case study
MM: QUAL + quan
Exploratory
Undergraduate nursing assessors (n=12)
Data:
Concurrent embedded design

QUAL: recorded interviews
quant: completed marking guides

Assumed relationships between concepts:

This perspective, informed by the work of Gauthier et al. (2016), conceptualises assessors' judgement formation as a three-stage process, influenced by a range of potentially idiosyncratic factors. As such, this decision informed the choice to adopt an interpretivist approach to the *methodological framework*, as interpretivism seeks to understand individuals' subjective perceptions of the world around them (Weaver & Olson, 2006).

In addition to interpretivism, the researcher chose to adopt a pragmatic approach (Onwuegbuzie & Leach, 2005). This decision was informed by the recognition that the research questions would likely require a range of data collection methods, notably the collection of score data, which would be used to calculate IRR. The flexibility afforded by pragmatism, combined with the deep understanding that is the goal of interpretivist research, led to the adoption of an exploratory, mixed methods design. In turn, this affected what *data* were collected: both qualitative and quantitative data were used to explore the research questions, according to the QUAL + quan typology identified by Johnson and Onwuegbuzie (2004). These data were collected using a concurrent, embedded design, with 12 assessors who worked at an undergraduate School of Nursing participating in the study.

At the bottom of the framework, there is an overview of the key concepts informing the study, and (based on findings from the literature review) how they were expected to relate to each other. As discussed in section 2.6 of the previous chapter, assessors' observations of student performances are influenced by the marking guide, but also, in some cases, by prior knowledge and expertise. These latter two factors also affect the processing stage. In turn, these inform assessors' integration of all the information they have about a student, and the subsequent grades they award. The extent of subjective factors impacting assessors' judgements has the potential to threaten the reliability of awarded scores.

3.3 *Research Questions and Study Summary*

Based on the knowledge gaps identified in the literature review, and outlined in the conceptual framework above, the following research questions were devised to guide the study. The questions are outlined below, followed by a table which provides an overview of the study which addressed them.

> RQ1: *What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?*
>
> RQ2: *What level of inter-rater reliability is evident in undergraduate nursing OSCEs?*

RQ3: *Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?*

In order to address these research questions, a range of different methods were employed. Qualitative data, namely semi-structured interviews, a cognitive interview and a think aloud protocol - in which assessors viewed videos of students completing two OSCEs and vocalised their thoughts about how the student was performing – were used to address **RQ1.** Quantitative data, based on the marking guides assessor participants completed while watching the recorded videos, were used to address **RQ2**. Finally, a combination of these qualitative and quantitative data were used to address **RQ3**. Before these methods could be carried out, two pre-studies took place, in order to inform the design of the study (Phase 1) and to film a series of videos which were used for the think aloud protocol (Phase 2). An overview of these methods can be found in the table below.

Table 3.1 *Overview of three-phase research design*

Overview of Phases	**Method(s)**	**Participants**	**Purpose**
Phase 1	Scoping interview	5 coordinators who oversee the development and administration of nursing OSCEs.	To inform the selection of the OSCEs that were used as the basis for filming in Phase 2. To inform the design of Phase 3, in particular: Inclusion of questions in the semi-structured interview guide.
Phase 2	Filming of videos	3 current undergraduate nursing students.	To film videos that were used as the basis for Phase 3.
Phase 3	Semi-structured interview Cognitive interview Think aloud protocol IRR analysis	12 assessors in the undergraduate nursing assessors.	To determine the cognitive processes that assessors use when assessing students (RQ1). To calculate the level of IRR when the 12 assessors viewed the 4 videos (RQ2). To explore the links between assessor cognition and reliability (RQ3).

The following section (3.4) provides an overview of the research design. Information about the research participants and sampling techniques that were employed in the study can be found in section 3.5. Section 3.6 provides a detailed description of the different research methods that were used in the study, as well as how different stages of the study informed each other. Data analysis procedures are detailed in section 3.7, followed by an outline of the pilot study that was utilised (section 3.8). The chapter ends with a discussion of the relevant ethical issues which informed the research choices.

3.4 *Research design*

Section 3.4 outlines the design of the research study. As illustrated in the conceptual framework, this research was informed by a case study design, involving a mixed-methods approach to data collection (Cope, 2015; Johnson & Onwuegbuzie, 2004), informed by the research paradigms of interpretivism and pragmatism (Weaver & Olson, 2006). Section 3.4.1 focuses on why interpretivism and pragmatism were selected as the overarching paradigms, and how they influenced various aspects of the study design. Section 3.4.2 outlines the case study aspect of the design, and includes information about the specific case that was used as part of the study. Finally, section 3.4.3 describes the mixed methods design employed, and classifies the study as QUAL + quan according to the typology described by Johnson and Onwuegbuzie (2004).

3.4.1 *Research paradigm*

One important decision that had to be made when devising the present study was that of the chosen research paradigm. Paradigms are defined by Weaver and Olson as "patterns of beliefs and practices that regulate inquiry within a discipline by providing lenses, frames and processes through which investigation is accomplished" (2006, p.460). The paradigm adopted by the researcher influences all stages of the research process, in particular the chosen research methodologies employed as part of the study.

This study is situated primarily within the *interpretivist* paradigm. Interpretivism focuses on understanding phenomena "through the eyes of people in their lived situations" (Weaver & Olson, 2006, p.461). In contrast to other paradigms, such as positivism and post-positivism, researchers working within the interpretivist paradigm are not seeking to determine objective truths about the world; rather, they are trying to understand how truth is constructed by individuals in subjective ways. The main goal of interpretivist research is *understanding*, as opposed to *control* and *prediction* which are the main goals of positivist and post-positivist research.

Interpretivism is commonly used within nursing research, and has been employed in order to "gain an in-depth and detailed description, understanding and explanation of ordinary occurrences as experienced by those in the field" (Weaver & Olson, 2006, p.464). Interpretivism was chosen as an appropriate paradigm for this thesis as the goal of the research was to produce a rich description of a specific issue – namely, the cognitive processes through which nursing assessors go when coming to judgements about OSCE performances.

Additionally, because initial research that has been conducted in the field of assessor cognition has highlighted the subjectivity inherent in assessors' decision-making (e.g.: Yeates et al., 2013), interpretivism was deemed a useful paradigm to employ, as it emphasises the subjective construction of truth. Research conducted within the interpretivist paradigm commonly employs qualitative methodologies, as they allow for a deep understanding of issues from the point of view of the research participants (Weaver & Olson, 2006).

However, researchers working within the field of nursing (Weaver & Olson, 2006), and in the social sciences (Onwuegbuzie & Leach, 2005), have cautioned against adopting a single research paradigm and allowing it to influence the entire research process. These researchers have called for a *pragmatic* approach to research design. Pragmatism recognises that having a fixed lens through which to view a problem, and relying on one methodological approach to address that problem, limits the flexibility of the researcher (Onwuegbuzie & Leach, 2005). Adopting a pragmatic approach allows the researcher to incorporate elements of both qualitative and quantitative methods, if doing so deepens the understanding of the research problem.

The main focus of this thesis was understanding how nursing assessors form judgements about students, which necessitated the collection of qualitative data. However, because the thesis also explored the link between this judgement formation and score reliability, it was deemed appropriate to collect quantitative data in the form of completed marking guides. As such, the decision to adopt a pragmatic approach allowed the researcher to collect this wide variety of data.

3.4.2 Case study

This research used the OSCEs that are administered in a School of Nursing at an Irish university (the School) as a means of investigating broader questions about assessor cognition and reliability. As such, this project was informed by a case study approach (Cope, 2015). Case study research "focuses on one phenomenon…occurring in a defined or bounded context of time and place to gain an understanding of the whole of the phenomenon under investigation" (Cope, 2015, p.681). While the researcher did not make broad claims about the generalisability of the findings in contexts other than the School, it was nevertheless expected that the results of the study will be of interest to researchers working in nursing assessment more generally.

In case study research, the researcher generally provides information about the case under investigation, and uses a wide range of methodologies in order to gain an understanding of his case (Cope, 2015). One of the key advantages of case study research is that it provides the researcher with flexibility in terms of research methods and paradigms: as the aim to provide a rich description of the case, he is not overly constrained by dogma, and is free to employ both qualitative and quantitative methods (Cope, 2015). In this chapter, information about the OSCEs that take place in the School is provided, as well as detailed information about the array of research methods that were employed: one-to-one interviews, cognitive interviews, a think aloud protocol, and quantitative measures of inter-rater reliability.

The School has been active for several decades. The School offers a series of undergraduate and postgraduate taught programmes across a range of disciplines, and also offers research opportunities for masters and doctoral students. In terms of assessment, single-station OSCEs were used as part of the assessment battery in seven different undergraduate modules at the time of writing. Staff members within the School include academic staff, who are responsible for lecturing, coordinating modules, and conducting research; and clinical skills nurses, who are primarily responsible for the teaching of skills to students, and who also participate in assessment of these skills.

Before the onset of the COVID-19 pandemic in March 2020, the majority of OSCEs were conducted on-site within the School, at a bespoke centre (referred to for the rest of the thesis as the Lab) which simulates a hospital ward. For most OSCEs, students would perform the required skill (either on a real person or a mannequin) within the Lab while an assessor watched and completed the accompanying marking guide. Each marking guide comprised a series of discrete checklist items, which assessors marked as either "completed" or "not completed". Depending on the OSCE, all checklist items were either afforded equal weighting, or certain items would be allocated more marks depending on the perceived importance of the step. The cut score to determine if a student had passed the OSCE was based on a simple calculation of the percentage of marks awarded: students had to receive 40% of the possible marks in order to pass. OSCE scores were used for summative purposes: a student's score on an OSCE accounted for anywhere between 25% and 50% of their total grade for a specific module. As such, the OSCEs can be considered as high-stakes assessment, in that a student's score may affect their ability to progress to the next year of study.

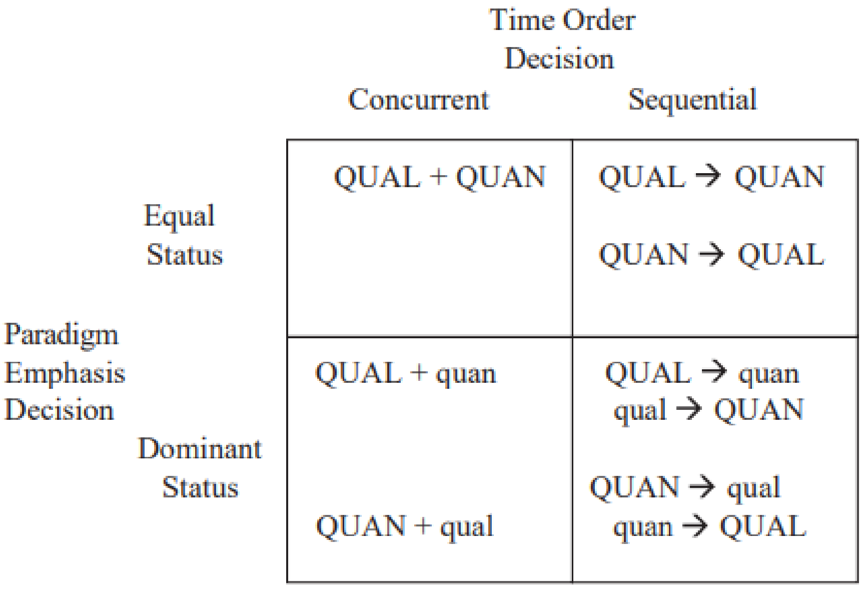
The COVID-19 pandemic dramatically curtailed to ability of module coordinators within the School to deliver in-person assessment. Although the use of video had been incorporated into some modules before 2020, all coordinators were forced to develop some form of remote assessment from March 2020. Some OSCEs, which required less specialised equipment (such as blood pressure measurement) were well-suited to the video format, in which students recorded themselves performing the skill at home and submitted the video to an online learning platform for grading by an examiner. Other skills, which required more equipment, were less well-suited, and coordinators of modules which taught these skills were in some cases forced to abandon OSCEs until the Lab opened again in 2021. The Lab has facilities for recording skills, which allowed students to film themselves completing an OSCE within the Lab and submit this for assessment. As such, coordinators of modules which were poorly suited to home recording were able to run OSCEs in the Lab without having to have an assessor on-site, thus allowing assessment to take place without the risk of COVID contamination. At the time of writing, module coordinators were considering whether to return to "live", in-person OSCEs, or to keep some elements of the video process.

3.4.3 Mixed methods: QUAL + quan

In accordance with the methodological flexibility that case study research allows, this study took a mixed methods approach to answering the research questions. Mixed methods "combines elements of qualitative and quantitative research approaches for the broad purpose of increasing the breadth and depth of understanding" (Doorenbos, 2014, p.207). A central tenet of mixed methods research is that it allows the researcher to gather a wider range of data than would be possible if an exclusively quantitative or qualitative approach was adopted. Mixed methods is deemed to be an appropriate strategy to adopt for some topics in nursing research as it can provide a "middle ground" between the strict objectivity of quantitative research and the subjectivity inherent in qualitative research (Doyle et al., 2009; Doorenbos, 2014).

As noted by Johnson & Onwuegbuzie (2004), it is important for the researcher to be explicit about the specific mixed methods design that is employed. This involves making two primary decisions. The first is whether the researcher wishes to conduct their work with equal weight given to quantitative and qualitative methods, or whether one of these approaches will be dominant and supplemented by the other. The second is whether the quantitative and qualitative data will be collected at the same time (concurrent), or whether data collected through one approach will be used to shape the design of a subsequent study to be conducted through the other approach (sequential). The possible results of these two decisions are illustrated in the figure below (Johnson & Onwuegbuzie, 2004, p.22):

Figure 3.2 *Mixed methods design typologies*



		Time Order Decision	
		Concurrent	Sequential
Paradigm Emphasis Decision	Equal Status	QUAL + QUAN	QUAL → QUAN QUAN → QUAL
	Dominant Status	QUAL + quan QUAN + qual	QUAL → quan qual → QUAN QUAN → qual quan → QUAL

Note. "qual" stands for qualitative, "quan" stands for quantitative, "+" stands for concurrent, "→" stands for sequential, capital letters denote high priority or weight, and lower case letters denote lower priority or weight.[11]

In this study, the dominant method of data collection was qualitative, which was augmented with a smaller amount of quantitative data in the form of completed marking guides. Additionally, because the majority of the qualitative and quantitative data were collected at the same time (assessors completed the marking guides while undergoing the cognitive interviews), the mixing of methods was *concurrent*, with the collection of the quantitative data *embedded* within the qualitative study. Therefore, this research was QUAL + quan as per Johnson & Onwuegbuzie's (2004) typology.

Additionally, because the problem under investigation has not been examined thoroughly in the past, the research was *exploratory* rather than explanatory; any results were treated as tentative, and could later be tested by other researchers on a larger scale or in a true experimental or quasi-experimental design (Fraenkel & Wallen, 2006).

3.5 *Research participants and sampling*

3.5.1 *Sampling approach*

The sample for the study consisted of three different groups: module coordinators (Phase 1), current nursing students (Phase 2), and OSCE assessors (Phase 3). The selection of participants for each of these three phases was based on factors other than random selection; therefore, a non-random sampling approach was employed (Fraenkel & Wallen, 2006). All participants were recruited through convenience and purposive sampling methods. In convenience sampling, the researcher contacts participants who are readily available and accessible (Fraenkel & Wallen, 2006). Although convenience sampling leads to unrepresentative samples and limits the generalizability of results, it is deemed an appropriate methodology for qualitative studies where the researcher is using a specific organisation (or part of an organisation) as a case study, and when he is not claiming that results from the sample will generalize to the population as a whole (Robinson, 2014). Purposive sampling is when the researcher selects participants using personal discretion, who the researcher knows would be suitable for the study (Fraenkel & Wallen, 2006). Purposive sampling is often used in qualitative research when the sample comprises people who are in positions of authority or are considered to be experts in a certain area (Robinson, 2014).

For Phase 1, the population studied was module coordinators who oversee the implementation of undergraduate nursing OSCEs. It was decided that, because the focus of the thesis is on OSCEs that take place within the School, only coordinators from within the School would be contacted. Through consultation with the relevant stakeholders, it was determined that there were seven people who met these criteria. As such, the sampling approach was purposive (Robinson, 2014).

Phase 2 of the study consisted of the filming of a series of videos of current nursing students completing two OSCEs. A selection of these videos were used as the basis for the range of research methods that took place in Phase 3. Details of how the OSCEs were selected can be found in section 3.6.2.1. A key concern when recruiting nursing students to film these videos was to introduce variability in performance levels. As such, it was decided to recruit a convenience sample of one student each from 1st, 2nd and 3rd year nursing programmes. The criteria for inclusion in this Phase of the study were as follows:

- Participants must be currently enrolled in 1st, 2nd or 3rd year in a BSc in General Nursing, Mental Health Nursing, Intellectual Disability Nursing, or Children's/General Integrated Nursing; at a university in Ireland.
- Participants must be at least 18 years of age.
- It was also decided that, if possible, a sample of participants who were demographically similar to each other would be recruited, to remove the possibility that some participants would be judged differently than others on the basis of characteristics such as race or gender.

It is noted at this juncture that, for Phase 2, there were thousands of participants who may have been eligible for inclusion. However, the study took place against the backdrop of the COVID-19 pandemic, during a period in which there were intermittent restrictions on the ability to travel, take public transport, or be in an indoor space with another person. As such, while the pool of potential participants for this phase was in theory large; in reality, it proved difficult to find appropriate people. The process of recruiting participants is discussed in the following section.

For Phase 3, the population being studied was assessors who judge student performances in undergraduate nursing OSCEs. In line with previous work in assessor cognition, a sample of between 10 and 15 assessors was expected to be an appropriate number to recruit (e.g.: Yeates et al., 2013; Hyde et al., 2020). Within the School, OSCE assessment is carried out by academic staff members, as well as clinical skills nurses. Academic staff members are School employees who lecture within the School, and are informally recruited by module coordinators to assist with the administration and grading of OSCEs. There are approximately 60 people employed in this capacity. Clinical skills nurses are employed specifically to teach nursing procedures to students and assist with the grading of these procedures when they are assessed in examinations such as the OSCE. There were five people employed in this role when the study was carried out.

The criteria for inclusion were as follows:

- Participants must be a member of the School; either as an academic staff member or a clinical skills nurse.
- Participants must have experience assessing student OSCE performances.

Module coordinators in Phase 1 noted that the transition to video assessment brought about by COVID had reduced the number of academic staff members who had recent experience assessing OSCEs, as the responsibility for grading had largely fallen on the module coordinators and the clinical skills staff. As such, a challenge for the researcher was identifying potential participants outside of those who acted as either module coordinators of clinical skills staff, as he did not have ethical approval to send a generic recruitment email to all staff members. As a result, the clinical skills nurses were recruited using purposive sampling, while a mixture of convenience and purposive techniques were used to recruit the subsequent group. The procedure for how these people were recruited is discussed in the following section.

3.5.2 *Recruitment procedures*

For Phase 1, the seven module coordinators who were identified as being fit for inclusion were contacted over email by the researcher; a total of five consented to participate in the interviews. These five people coordinated either first- or second-year modules, and had developed OSCEs for use as part of these modules. A summary of these OSCEs, and the ones that were eventually included for selection in the study, can be found in section 3.6.2. Participants were emailed a Plain Language Statement (Appendix A) and Informed Consent Form (Appendix B) in advance of agreeing to take part in the study.

To recruit participants for Phase 2, the researcher emailed class representatives of 1st, 2nd and 3rd year nursing programs at several universities in Ireland. After this process, only a single student each from 2nd and 3rd year expressed interest in participating, these students were included in the study. A total of five students from 1st year nursing programs expressed interest in participating: the researcher selected for inclusion the first student who contacted him. In addition to this, the researcher recruited a further participant to act as the “patient” for one of the OSCEs included in the study. This participant was known to the researcher.

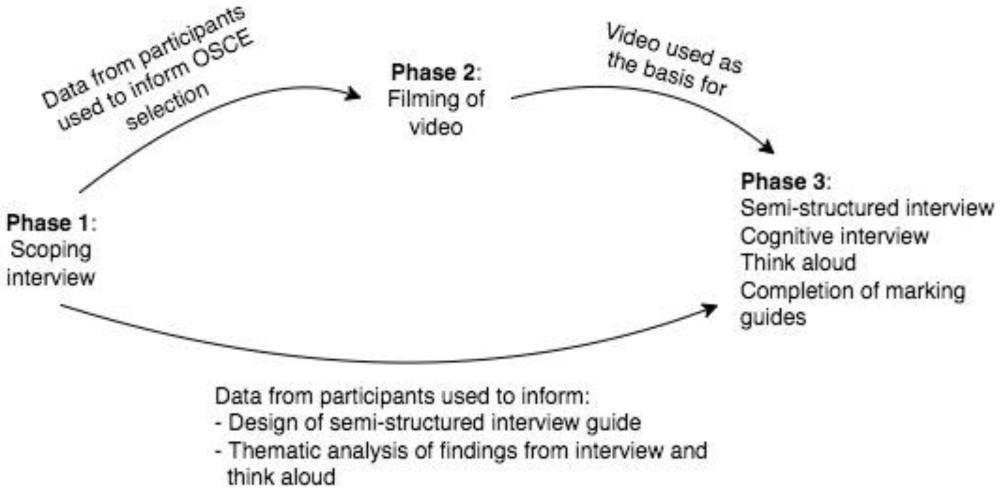
The three recruited participants were all studying General Nursing. All three were white females, and their ages ranged from 19 to 23. Participants were emailed a Plain Language Statement in advance of filming, and also had to fill out a modified informed consent form, which had additional provisions specifying their rights under the General Data Protection Regulation (GDPR), due to video footage being classed as personal data (for ethical considerations, see section 3.9). All participants received €50 for their participation, and were provided with a voucher to buy lunch on campus after filming was completed.

For Phase 3, the researcher began by inviting the five participants who completed the Phase 1 interviews to participate in Phase 3. These five people had specified that they would be happy to be approached to participate in subsequent phases of the study. Of this group, four consented to participate. One of these four assessor participants acted as the pilot participant (see section 3.8). He then used the School's website to identify the five people employed as clinical skills nurses. He approached these five people by email and invited them to participate in the study, all five consented to do so. To recruit further participants, the researcher spoke with the module coordinators, as well as one of his supervisors (who was familiar with some staff members within the School) to determine whether any of the approximately 60 academic staff members had experience grading undergraduate OSCEs, and would therefore be deemed fit for inclusion. As a result, he identified five potential participants. These five people were approached by email, and four consented to participate. Excluding the pilot interview participant, a total of 12 participants took part in this phase of data collection. Demographic information about these participants can be found in section 4.3 in the following chapter. All 12 participants were sent a Plain Language Statement (Appendix A) and online Informed Consent Form (Appendix B), which was completed before participation.

3.6 *Research methods*

This section outlines in detail the research methods that were employed through all stages of the project, as well as how each stage of research informed the next. An overview of how Phases 1 and 2 informed Phase 3 can be found in the figure below. It is important to note at this juncture that Phases 1 and 2 were deemed necessary in order to facilitate Phase 3. As such, the results of the present study (outlined in Chapter 4) are largely based on data collected in Phase 3.

Figure 3.3 *Overview of interactions between study phases*


Data from participants used to inform OSCE selection
Phase 2: Filming of video
Video used as the basis for
Phase 1: Scoping interview
Phase 3: Semi-structured interview
Cognitive interview
Think aloud
Completion of marking guides
Data from participants used to inform:
- Design of semi-structured interview guide
- Thematic analysis of findings from interview and think aloud

3.6.1 *Scoping interview*

For Phase 1, the researcher conducted a group scoping interview with faculty members who design and administer OSCEs in the School. This interview took place over Zoom and the audio from the interview was recorded. A scoping interview is a form of initial interview in which the researcher describes the details of a research study to actors within a specific organisation which plays an important role in the study. The researcher collects information about the organisation from participants, which allows him to build a rich description of the case being studied (Cope, 2015). Participants are also given the opportunity to comment on the design of the overall study (Robertson et al., 2012). In contrast to other forms of interview (such as a structured or semi-structured interview), scoping interview data rarely answers any of the research questions directly; rather, it is used to shape subsequent phases of a research study (Robertson et al., 2012).

The main aims of the scoping interview were to:

- Collect information used to determine which OSCEs would be included in later stages of the study.
- Collect information as to participants' views of how OSCEs are assessed, and whether they perceived any issues with assessor consistency in the OSCEs they coordinate.

The empirical justification for each of these aims is discussed below, followed by a description of how they were instrumentalised, and details of how each aim was be used to shape further stages of the study.

3.6.1.1 *Document analysis*

Prior to the scoping interviews, the researcher asked participants to share documentation they had in relation to the development and administration of OSCEs. These documents were expected to include: instructions for staff members on setting up the OSCE, instructions for assessors detailing the assessment procedure, and marking guides. It was expected that this analysis would be used to deepen the knowledge of the case being studied (Cope, 2015) and inform the design of the scoping interviews. Ultimately, the interviewees indicated that the only documentation used when setting up and running an OSCE was the marking guide.

Document analysis is a common method in qualitative research, and is defined as “a systematic procedure for reviewing or evaluating” relevant pieces of written information (Bowen, 2009, p.27). Document analysis is particularly useful when the researcher wishes to increase their understanding of a specific issue or supplement the knowledge they gain from another research method, such as interviews (Bowen, 2009). Additionally, document analysis is an applicable method when the researcher is using a specific organisation as a case study, and needs to understand as much as possible about that case (Bowen, 2009).

The framework for the document analysis was informed by salient issues in OSCEs and assessor cognition that were outlined in the previous chapter (and which can be found in the conceptual framework). Specifically, the researcher used document analysis to determine the layout of the marking guides, specifically:

- Whether the tasks listed on the guides are judged on a binary basis (done/not done) or on a scale (Setyonugroho et al., 2015).
- The weighting of different tasks (e.g.: how many marks are awarded to each task) (Khan et al., 2013).
- The allocation of individual tasks into a specific domain of competence (e.g.: interpersonal skills, technical skills, critical thinking skills) and the weighting of these different domains (Khan et al., 2013b).

From this analysis, the researcher determined that the OSCE marking guides used in the School generally comprised a series of binary checklist items (marked as done/not done by assessors), with equal marks usually allocated for every step on the checklist. In some OSCEs, the individual items on the checklist were classified according to whether they tested psychomotor or affective skills. The affective items accounted for between 0% and 20% of marks for a given

OSCE. A discussion of how the marking guides for the two chosen OSCEs were adapted for the present study can be found in section 3.6.3.4.

Another important piece of information that was discovered from the document analysis is that all of the OSCEs that are administered in the School are single-station, instead of multi-station as is the case in a traditional OSCE (Harden et al., 1975). This finding is notable from an assessment standpoint, as having only one station (and one assessor) increases the potential effect of a particular assessor's harshness or leniency on a student's score, and as such can be considered a threat to reliability (Rushforth, 2007).

3.6.1.2 *Collect information about which OSCEs to include*

The main aim of the scoping interviews was to collect information which would be used to determine which OSCEs would be included in the study. There are several theoretical and practical considerations that informed this decision. From a theoretical perspective, research has indicated that OSCE items which assess affective skills, such as communication and rapport-building, are more likely to lead to divergence between assessors (Cazzell & Howe, 2012); as are OSCEs which are more difficult for students to complete (e.g., those assessed at the later stages of degree level). Given that the object of study is assessor divergence, it was deemed logical to use OSCEs which contain a significant number of checklist items which assess competency in the affective domain, and which are administered later in degree level.

From a practical perspective, the number of students who complete the OSCE each year was another criterion for inclusion. OSCEs which are taken by a large number of students were prioritised, as it was more likely that there would be a greater number of assessors who have experience grading those specific OSCEs, and therefore more possible participants for Phase 3 of the study.

In order to ascertain this information, interview participants were asked to complete a table (see below), in which they provided details as to the OSCEs they coordinate. Participants were asked to provide the name of the OSCE, the year group to which it is administered, the approximate number of students who complete the OSCE each year, the primary skills assessed by the OSCE, how difficult the students find the OSCE, and whether the OSCE has been difficult for assessors to grade consistently in the past. Participants completed this form in advance of the interviews.

Table 3.2 *OSCE information from module coordinators*

OSCE name	Number of students taking every year	Primary skills assessed	OSCE difficulty for students: Easy, medium, hard	Difficulty in terms of achieving consistency of grading: Easy, medium, hard

During the interview, the researcher asked the following questions, which prompted participants to elaborate further on the theoretical issue of assessor divergence:

> Are there any OSCEs that you have personally found difficult to mark consistently? If so, please describe the OSCE and what makes it difficult.
>
> Please describe OSCEs that have tended to lead to divergence in how they are assessed. In other words, are there OSCEs where assessors are likely to have differing opinions about a student's performance level? What do you think causes this divergence?

The findings from the completed tables and the two qualitative questions were used to select the OSCEs included in subsequent stages of the study. The criteria for selection of the OSCE were as follows, listed in order of importance:

- Grading consistency: the researcher selected OSCEs that have traditionally caused high levels of divergence between assessors. Research has indicated that OSCEs which assess affective skills such as communication may lead to discrepancies between assessors (Rushforth, 2007; Cazzell & Howe, 2012); therefore, it was expected that scoping interview participants would identify such OSCEs as causing assessor divergence.
- Difficulty: closely related to the above criterion, the researcher aimed to select OSCEs in which students have traditionally found it difficult to attain a high grade. Assessor cognition research has consistently used more complex OSCEs/OSCE stations as the basis for research (e.g.: Govaerts et al., 2013), underpinned by the assumption that the

assessment of complex skills places grater cognitive strain on the assessor, making them more likely to rely on individualised judgement mechanisms.

- Number of students taking OSCE: it was deemed prudent to select OSCEs which are taken by a high number of students every year, in order to maximise the number of assessors who have experience grading these OSCEs that could participate in the Phase 3 interviews.

Details of how information from these interviews was used to inform the ultimate selection of the two included OSCEs can be found in section 3.6.2.

3.6.1.3 *Collect information about how OSCEs are assessed*

Research has indicated that there are a range of different cognitive mechanisms, internal to the assessor, which cause idiosyncrasy in how student performances are judged (e.g.: Gauthier et al., 2016). However, when OSCEs are designed and implemented, there are various different procedures which are explicitly designed to bring about agreement between assessors, such as standard setting meetings and reliability analyses (Khan et al., 2013b). Participants in these scoping interviews were deemed likely to have an understanding of possible threats to assessor consistency and how they could be mitigated. In order to determine whether this was the case, the following questions were included in the interview schedule:

> Are there any procedures in place which are designed to bring about consistency between OSCE assessors? If so, please describe them.
>
> Please describe the process(es) by which assessors are trained in how to examine the OSCE.
>
> Have you ever encountered a situation where you found that as assessor was being overly harsh or lenient? If so, why do you think this was the case? Please elaborate on what happened.

The findings from these questions were used to inform future stages of the study, notably the semi-structured interviews which took place with OSCE assessors. In this interview, assessors were asked about the process of judging student performances, and the factors that affect their judgements – these questions incorporated the specific procedures or processes mentioned by scoping interview participants in response to the above questions.

For example, responses to the first question indicated that assessors in the School often chatted informally about OSCE assessment, and how well or badly a particular cohort of students was performing. These informal conversations then moderated their assessment decisions in the future. As such, the researcher included a question about informal moderation practices during the semi-structured interviews. A full copy of the interview schedule for the scoping interview can be found in Appendix C.

3.6.2 *Phase 2: Filming of video*

After the completion of the scoping interview, the researcher recruited three current nursing students (for the recruitment procedure, see section 3.5) and filmed them completing two OSCEs. The OSCEs were selected according to the criteria outlined in the previous section, with additional input from participants in the scoping interview. The decision-making process for selecting the two OSCEs for inclusion in the study is discussed below.

3.6.2.1 *OSCE Selection*

There were several theoretical and practical considerations that informed the choice of OSCEs used as the basis for this study, which were detailed section 3.6.1.2. In addition to these criteria, the interview revealed two additional criteria that participants believed should inform the selection of the OSCEs. Firstly, participants suggested that because the think aloud portion of the Phase 3 interviews would take place with assessors watching videos of OSCEs, it would make sense for the researcher to select OSCEs that had already been adapted for video by the module coordinators. This would ensure that there was a high level of alignment between the OSCEs that were selected for inclusion in the study and those which were administered in the School. One participant noted that “I was about to look at videoing mine when we got the COVID. So you know the [ones which are done through video] are probably a good one to look at”.

Relatedly, participants noted on several occasions that OSCEs where the participant largely stays in the same place should be selected, so as to minimise the chance that the participant’s actions would not be fully visible due to their movements, which may not end up being captured on camera. For example, one participant commented that: “So, I suppose, just having a very practical kind of thing would you be better off picking one that's just maybe they only stay in one place. Like blood pressure or something”.

Between the five participants in the scoping interview, there were six OSCEs that they were responsible for coordinating, and which were therefore eligible for inclusion in the study. Based on the information available, two OSCEs were chosen for filming: blood pressure (BP) measurement and naso-gastric (NG) tube insertion. A detailed description of how these two OSCEs were selected can be found in Appendix D.

3.6.2.2 *OSCE filming*

The OSCEs were filmed in early 2022 in the nursing education laboratory (the Lab) in the School. This facility is designed to facilitate the teaching and learning of essential nursing skills, and has equipment to allow the recording of these skills. In addition to the three recruited nursing students, the researcher also recruited an acquaintance to assist with filming: this person acted as the "patient" in the blood pressure OSCE (the naso-gastric tube insertion was performed on a mannequin). As such, there were three people running the OSCE: the researcher, this acquaintance, and the supervisor, Dr. Mary Kelly. The videos were recorded using UniCam, with a camera and microphone embedded into the ceiling that the researcher was able to operate remotely with his phone.

The arrival time for each participant was staggered – the 1st year student arrived and filmed both OSCEs, followed by the 2nd year student, and finally the 3rd year student. In order to reduce participant anxiety, Dr. Kelly stepped out of the room while the videos were recording, and came back in once they were completed.

The procedure for setting up and recording the OSCEs was adapted from a guide provided by Boursicot and Roberts (2005). It is noted that the goal of these OSCEs was to record videos to be used in the later stages of the study; as such, changes were made to the Boursicot and Roberts' guidelines as needed. The relevant steps for setting up the OSCEs can be found in Appendix E.

Subsequent to filming, the researcher had six usable videos, which ranged in duration from 94 seconds to 345 seconds. Details of the length of each video, and the code names that are assigned to each one, can be found in the table below. The first section of the code (P01/02/03) refers to the student who completed the skill, while the second half (BP/NG) refers to the skill itself. A discussion of which videos were included in the next stage of the study takes place in section 3.6.3.3.1.

Table 3.3 *Details of recorded OSCE videos*

	Blood pressure measurement (BP)	Naso-gastric tube insertion (NG)
First-year	Code: P01BP Length: 161s	Code: P01NG Length: 345s
Second-year	Code: P02BP Length: 132s	Code: P02NG Length: 226s
Third-year	Code: P03BP Length: 94s	Code: P03NG Length: 156s

3.6.3 *Phase 3: Interview series*

The majority of data collected in this study comprised a series of research methods conducted with assessors who have experience judging undergraduate nursing OSCEs in the School (see section 3.5 for recruitment procedures). Each assessor participant completed a one-on-one, face-to-face session with the researcher in which they underwent a series of procedures in a specific order:

1. Semi-structured interview: assessor participants underwent an interview in which the researcher asked them questions about the process of assessing undergraduate OSCEs.
2. Cognitive interview: assessors were provided with the marking guide for the two OSCEs that were filmed in the previous phase and asked to explain how they understand each item contained within the guide.
3. Think aloud: assessors watched a series of videos recorded in the previous phase, and were asked to vocalise their thoughts as they judged the student's performance, as if they were doing so in a real OSCE.
4. Completion of marking guide: while watching the videos, assessors were asked to complete the marking guide used to grade student performance on that OSCE, again as if they were doing so in a real OSCE.

This phase took approximately one hour for each participant to complete. The audio from the interview series was recorded. Details of each of these four stages is provided below. Before the 12 interviews took place, the researcher conducted a pilot interview with one of the participants from Phase 1. More information about this can be found in section 3.8; however, where data from the pilot study were used to inform an assessment decision, it is detailed in this section.

3.6.3.1 *Semi-structured interview*

In the first stage of the interview, assessor participants were subject to a semi-structured interview in which they were asked questions about OSCE assessment. The semi-structured interview guide was informed by the literature review, as well as the research questions. The first interview question was based on previously cited studies in assessor cognition (e.g., Hyde et al., 2020), in which assessors are asked to talk through the process of judging student performances:

> Please talk me through the process of assessing a student in an OSCE. Specifically, how you determine if a student has passed the exam?

The framework used by Gauthier et al. (2016) informed questions relating to assessors' *observation* and *processing* of student performances. In terms of *observation*, the following question was included:

> Is there anything in particular you look out for when judging a student OSCE performance?

The following question was included in order to allow assessors to talk about their *processing* of student performances:

> How do you do determine if a student has completed an item on the checklist properly?

Another key concept identified in the literature was that of *expertise* (Benner, 1982). The following questions were included to prompt assessors to talk about their own levels of expertise and how this affects their judgement processes:

> How do you think you have developed as an assessor since you began assessing? If there anything different about how you approach the task of assessment?
>
> Do you find it easier or more difficult to assess OSCEs than you used to?

Finally, because the aim of this thesis was to uncover instances where variance in assessors' judgements might arise, the following questions were included to prompt assessors to speculate about situations where their judgements might be distinct from those of other assessors. This was particularly important regarding borderline performances:

> Please describe how you would imagine a "minimally competent" or "borderline competent" student. What distinguishes someone who has *just* passed from someone who has *just* failed?
>
> Has there ever been a time when you and another assessor disagreed about how a student performed? How was this resolved?

As noted in section 3.6.1.4, scoping interview participants discussed moderation procedures (whether formal or informal) as a factor assessing assessors' judgements. The final question was informed by this discussion.

A full copy of the interview schedule can be found in Appendix F. In line with the semi-structured nature of the interview, participants were given freedom to bring up topics that emerged naturally as part of the conversation. The researcher also had a series of backup questions which were used as prompts if the main questions did not result in participants giving lengthy answers.

3.6.3.2 *Cognitive interview*

In the second stage, participants underwent a cognitive interview, with the aim of determining how they understand the items contained in the marking guide of the filmed OSCE. Cognitive interviewing is a technique used to gain an understanding of participants' mental processes when they complete an assigned task, such as a survey (Ericsson & Simon, 1980; Willis, 2015). Cognitive interviewing allows researchers to understand better what specifically happens when a research participant is asked to carry out an instruction. Cognitive interviewing originated as an offshoot of quantitative methods, with researchers seeking to determine how survey questions were understood by respondents (Ericsson & Simon, 1980).

In this cognitive interview, assessor participants were presented with the marking guide, and received the following instruction:

> "For each item contained in the marking guide, please explain briefly what the terms mean to you."

Participants then went through each item in the marking guide, one-by-one, until they had explained how they interpret each item in the guide. Analysing data from this stage of the study (see section 3.7 for data analysis procedures) allowed the researcher to determine whether the language used in the marking guide is likely to lead to divergence in how students' performances are judged by assessors.

3.6.3.3 *Think aloud protocol*

In this phase of data collection, assessors watched four of the previously recorded videos and vocalised their thoughts as to how well or badly the students were performing. "Thinking aloud" is a technique that has been applied in numerous previous studies into assessor cognition (e.g., Kogan et al., 2011; Yeates et al., 2013). A well-designed think aloud protocol should allow a researcher to elicit valuable information about how assessors approach the task of assessment, in a simulation that retains a high level of fidelity to the "real world". The decision to adopt a think aloud approach was made after consideration was given to other means of determining assessors' cognitive processes (and the likely effects on score variance). For example, a 2020 study by Roberts et al. used generalisability theory to determine the proportion of variance in awarded scores that was due to assessor idiosyncrasy, and then conducted interviews with assessors in order to better understand how this idiosyncrasy emerged. However, the present study was limited due to the fact that the researcher did not have access to score data for a full cohort of OSCEs (as was the case in the Roberts study). As such, the study lacked the statistical power for more complex analyses. The use of a think aloud, therefore, represented the best possible approach to understanding assessors' judgment processes within the constraints faced by the researcher.

Details of the video selection process are outlined below, followed by an in-depth explanation of the protocol design.

3.6.3.3.1 *Video selection*

In order to determine which of the six recorded videos (see table 3.3) would be included for the think aloud, the researcher developed the following set of criteria:

1. <u>Total length of videos not to exceed 12 minutes</u>: the total time for the Phase 3 methods was one hour (in line with DCU guidelines on conducting empirical research). The researcher decided to allocate no more than 35 minutes to the think aloud protocol, in order to ensure sufficient time was devoted to the other data collection methods. From the pilot (see section 3.8) it was determined that the time required by participants to complete the think aloud would be approximately equal to three times the total length

of the videos. As such, it was concluded that the total length of all included videos would be no more than 12 minutes.

2. Inclusion of videos from all three participants: it was assumed that including videos from all three student participants would maximise the chance for variance between assessors to occur, as there would be a greater chance that one of the participants would display behaviour that was perceived by some of the assessors as ambiguous or debatable.

3. Inclusion of the worst performance overall: a key question in performance assessments is what differentiates assessors' judgements of borderline pass and failing test-takers (Rushforth, 2007), as these decisions are the most high-stakes that assessors commonly make at undergraduate level. As such, it was deemed important to include the video which was likely to be viewed by assessors as the worst performance (while keeping in mind that this was likely to be influenced by subjective factors). Initially, the researcher assumed that the first-year student would likely be the one who gave the worst performance. However, during the pilot study, the participant (who was responsible for co-ordinating and administering the blood pressure OSCE) was confident that the second-year student's blood pressure measurement (P02BP) was the least competent of all the videos. For this reason, it was included in the study.

4. Inclusion of more than one video for each OSCE: research in assessor cognition has consistently indicated that assessors lack a fixed understanding of "good" performance, and are likely to compare student performances against each other (e.g., Kogan et al., 2011). As such, it was decided that (if possible within the time constraint) two videos of each OSCE would be included. This would allow the researcher to determine whether assessors judged the second student they saw completing an OSCE in relation to the first.

5. Inclusion of two videos from the same student: assessor cognition research has indicated that assessors form impressions about students that may affect how that student is viewed subsequently – for example, if a student performs poorly at the beginning of an OSCE, assessors might fail to notice if that student performs well later on (Wood et al., 2018). Research also suggests that assessors' judgements might be

influenced if they know a student, or feel like they do (Schleicher et al., 2017). As such, it was decided to include two videos from the same student, to allow for a determination of whether assessors' impressions of the student completing the first OSCE would affect their subsequent judgement of that student completing the second OSCE.

Taking these criteria into consideration, the following four videos were included in the think aloud portion of the study: P01NG, P02BP, P03BP, P03NG. The shortest videos for each OSCE were those completed by student P03 – these videos were both included for this reason. As discussed in step 3 above, P02BP was included at the behest of the pilot interview participant, leaving P01NG as the final performance to be included in order to make sure there were two videos of each OSCE in the study. Stills from two of the videos (published with permission from the student participants) can be found in Appendix G.

It should be noted at this juncture that the decisions made by the researcher throughout the design of this phase of the study had the aim of trying to maximise variance between assessors. However, while such variance was discovered (as documented in the following chapter) it often emerged from unexpected places. For example, it was assumed that the most inexperienced student (P01) would be the worst performing student overall, and this poor performance might lead to variance between assessors. Ultimately, participants perceived the P01NG to be (broadly speaking) the video demonstrating the highest level of performance, and this decision had a broad degree of unanimity. As such, this demonstrates the unpredictability of empirical research, and that researchers can never control or predict perfectly what is likely to happen in a given study.

3.6.3.3.2 *Think aloud protocol design*

There were numerous procedural decisions that had to be made when setting up the think aloud protocol. These decisions, as well as the process by which previous studies in assessor cognition addressed them, are outlined in the table below. A detailed explanation of how each of these decisions was made follows the table.

Table 3.4 *Think aloud protocol strategies*

Study	**Probing strategy**	**Pausing allowed?**	**How many times can video be watched?**	**What order to show the videos?**
Kogan et al. (2011)	Retrospective	No	Once	Random order
Yeates et al. (2013)	Retrospective	No	Once	"Balanced order within a Latin square design" (p.329)
Govaerts et al. (2013)	Concurrent, but only if participant was silent for more than a few seconds	Pausing allowed once (when participant felt ready to make an initial judgement)	Once	Two videos - half watched each first
Gingerich et al. (2014)	n/a	No	Once	Seven videos: Half watched 1-7 in order, half watched 7-1 in order
St-Onge et al. (2016)	Retrospective	No	Once	Only one video
Wood et al. (2018)	n/a	Video stopped automatically after first minute	Once	Same order for each participant
Roberts et al. (2020)	n/s	No	Once	n/s

Note. n/a = not applicable, n/s = not specified

The first decision concerned the probing strategy adopted by the researcher. Probing can be either concurrent or retrospective. In concurrent probing, the researcher intervenes while the participant is thinking aloud to ask them to elaborate further on something they have just said; while in retrospective probing, the researcher waits until the task has been completed and asks a series of questions, again with the aim of elaborating on an aspect of what they said during the task (Willis, 2015). For these think alouds, the decision was made not to intervene in the process of watching the video, so as not to disrupt the process of assessment in a way that would be unnatural for assessors. This approach is in line with the majority of studies listed in the table. However, the researcher also decided (in line with the strategy taken by Govaerts et al. (2013)), that if the participant was silent for more than ten seconds in a row, he would prompt them to continue their vocalisation, by asking “Can you please explain what you are thinking at the moment?”

The second decision was whether the assessors were allowed to pause the video while it was playing. The majority of studies listed in the table did not allow pausing to take place, as it would not replicate the experience of assessing a “live”, in-person OSCE. The exceptions to this were the studies by Govaerts et al. (2013) and Wood et al. (2018). In both cases, the pausing strategy was closely linked to the aim of the research. In the former, the researchers were trying to determine how quickly experienced assessors felt confident making a judgement about a student, and so instructed assessors to pause the video, once, when they felt they could to this. In the latter, the researchers wanted to ascertain whether an assessor’s first impression of a student could be changed by a subsequent improvement in that student’s performance, so they stopped the video automatically after exactly one minute. In neither instance was the assessor allowed to pause the video as many times as they liked. Initially, a similar strategy was to be adopted for this study. However, during the pilot study (see section 3.8), the participant found it too cognitively demanding to vocalise what they were thinking while still watching and listening to the video. Additionally, participants in Phase 1 of the study frequently mentioned the ability to pause the video as something they do when assessing real OSCEs. For these reasons, it was decided that participants could pause the video as often as they wanted.

The third decision was how many times the participants would be allowed to watch each video. In every study listed in the table, participants were only allowed to watch the video a single time. A similar approach was adopted in this study. However, as noted above, assessor participants were allowed to go back to a specific aspect of the video they found confusing or difficult to judge in order to watch it again. When this happened, the researcher asked why they felt this was necessary.

The final decision the researcher had to make was the order in which to show the videos. Out of the studies which used multiple videos, only one (Wood et al., 2013) presented the videos to the participants in the same order every time. For the present study, a balanced Latin square design was used, in line with the approach taken in Yeates et al. (2013). The researcher labelled the videos as follows: P01NG = A, P02BP = B, P03BP = C, P03NG = D. He then entered A, B, C, D into an online Latin square generator (Masson, n.d.). The following square was generated:

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

He then used this as the order for the videos presented to participants, e.g. the first participant watched P01NG, followed by P02BP, P03NG and then P03BP. As there were 12 participants, this square was repeated three times. A key advantage of the Latin square design is that it equalises all order effects. For example, A is followed by B the same number of times that it is followed by C and D; B is followed by A the same number of times that it is followed by C and D, and so on (Edwards, 1951).

3.6.3.4 *Completion of marking guides*

While watching each video, assessors were asked to complete the relevant marking guides in order to award each student a grade. Data from the completed marking guides were then used as the basis for the quantitative analysis, to determine whether there were measurable differences in how the same four performances were graded by the 12 assessor participants. As determined from the document analysis in Phase 1 of the study (section 3.6.1.1), both the blood pressure OSCE and the naso-gastric tube insertion OSCE are assessed solely through a series of binary checklist items, each of which assessors tick if they deemed the student to have completed the step successfully. For the BP OSCE, there are 12 of these items, for the NG OSCE, there are 14. These items were included in the marking guides devised for the study.

In addition to these binary checklist items, the researcher also chose to include three “global” items at the end of each marking guide. These items were labelled as “global” because they pertained to the student’s overall performance, rather than their completion of a specific step (Rushforth, 2007). The inclusion of global items was deemed essential as part of this study, as research has indicated that assessors have different strategies for “translating” their checklist scores on a given marking guide into an overall score (e.g., Yeates et al., 2013). Additionally, empirical studies have shown that when checklist items are included, and allocated equal number of marks, assessors have differing ideas about which are the most important (Chahine et al., 2016). The inclusion of global items, therefore, allowed the researcher to investigate the relative importance of specific checklist items when assessors decide on an overall OSCE result. Indeed, as discussed in the following chapter, these global items revealed considerable variation between assessors, even if they had ticked a similar number of checklist items. As such, the inclusion of these items reduced the fidelity of the OSCEs to their real-world administration (as these items are not included when the OSCEs are administered in the School), but allowed for a deeper insight into assessors' cognitive processes.

The first global item related to a student's communication skills:

> How would you rate the participants' **communication** skills? Fail / Borderline pass / Good / Excellent

Research has suggested that "good" communication might be more subjective and open to interpretation on the part of assessors (Cazzell & Howe, 2012). As such, it was determined that including this question would allow for both a discussion about, and quantitative measurement of, whether there were differences in how participants interpreted a student's communication skill.

The second item was as follows:

> How confident would you be allowing this student to perform this procedure **on you**? Not at all confident / Just about confident / Confident / Very Confident

The aim of this question was to determine whether, broadly speaking, participants found the student in the video to be "entrustable", e.g.: whether they thought the student demonstrated enough competency in the video to perform the skill in real life. This is in line with research that suggests that a key factor guiding assessors' judgements is if they think a student would succeed in the "real world" of clinical practice (Hyde et al., 2021).

The final question was as follows:

> How would you rate this performance **overall**? Fail / Borderline pass / Good / Excellent

This item is one which is commonly included in OSCEs in both nursing and medicine (Rushforth, 2007; Khan et al., 2013b). Its inclusion allowed for a determination of each participant's final judgement about the recorded performance (particularly at the crucial pass/fail decision) and acted as a means by which participants could discuss any aspect of the performance they wished. It is important to note that it was this global item which was used to classify a student's performance in this study as either a pass or fail. This is distinct from when the OSCEs are administered in the School, when the pass/fail decision is made using a cut score of the number of items completed correctly (usually 50%). This has implications for the results of the study in the context of the School, as the frequency of "fail" decisions based on the above global item cannot be assumed to correlate to the actual frequency of “fail” decisions when they OSCEs are administered in the School.

As a final piece of data collection, participants were asked at the end of each video to complete the following statement: “I think this is the type of student who…”. Adapted from Gingerich (2014b), this allowed for a further qualitative insight into their opinion about each student’s performance.

As such, for each of the two OSCEs, the marking guide contained a series of *binary checklist items* (12 for the BP OSCE, 14 for the NG OSCE) and three *global items*. A full copy of the marking guide can be found in Appendix H.

3.6.3.5 *Procedure*

The 12 interviews took place from February-April 2022. Each interview was conducted in-person, in the assessor’s office. After introducing himself to the assessor, the researcher began by collecting demographic information about the assessor, such as their age and years of experience. This was followed by the semi-structured interview, which lasted about 20 minutes. After this, the researcher conducted the marking guide cognitive interview, which took approximately five minutes. Finally, the researcher presented the first video to the assessor on his laptop, who underwent the think aloud protocol while concurrently completing the marking guide. Once this was finished, the same procedure was repeated for the second video, and so on until all four videos had been watched, discussed and graded. This took approximately 35 minutes. Following this, the researcher asked participants if they had any final questions or comments, and instructed them to email him in the future for any reason they wished.

Audio data from the interviews were recorded on the researcher’s laptop, which is encrypted and password-protected. The marking guides were completed with pen and paper, and were subsequently stored in a locked drawer in the researcher’s office. As such, for each interview, the data collected were as follows: one audio file of approximately an hour in length, and four completed marking guides (one for each of the four videos: P01NG, P02BP, P03BP and P03NG). At the end of the interviews, the researcher had 12 audio files and 48 marking guides.

3.7 *Data Analysis*

This section outlines the data analysis procedures that took place in order to allow the researcher to address each of the three research questions:

RQ1: *What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?* This question was addressed using qualitative data from the 12 interviews, analysed using thematic analysis. The analysis procedure is outlined in section 3.7.1.

RQ2: *What level of inter-rater reliability is evident in undergraduate nursing OSCEs*? This question was addressed using quantitative data from the 48 completed marking guides. The analysis procedure is outlined in section 3.7.2.

RQ3: *Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?* This question was addressed using a mixture of qualitative and quantitative data. The approach taken at this phase of analysis is outlier sampling, outlined in detail in section 3.7.3.

3.7.1 *RQ1: Thematic analysis*

In order to address the first research question, interview data were transcribed and analysed according to the principals of thematic analysis (TA). TA is a "method for identifying, analysing and reporting patterns (themes) within data" (Braun & Clarke, 2006, p.79). TA is concerned with "searching across a data set… to find repeated patterns of meaning" (p.86), rather than trying to understand the meaning of one piece of data (such as a single interview). When applied correctly, TA goes beyond merely counting how many times certain concepts or ideas are mentioned by participants, and instead allows the researcher to develop a deep understanding of an issue from participants' point of view (Castleberry & Nolen, 2018).

As noted by Castleberry and Nolen (2018), thematic analysis is the most commonly cited method of qualitative data analysis, however it is often mentioned without a thorough description of the specific procedures employed. Explicitly detailing how the analysis took place ensures that the conclusions drawn from the data are robust and defensible (Castleberry & Nolen, 2018). The qualitative data analysis in the present study followed the six-step process outlined by Braun and Clarke (2006). An outline of each stage of the process is detailed below.

1. *Familiarisation*: the data (726 minutes) were initially transcribed using an online transcription service, Happy Scribe. Subsequently, the researcher read each transcript through once, then read them again while listening to the audio, to amend any errors. Each transcript was then read a third and final time. The process of amending errors to each transcript took

approximately three times the length of the audio file, i.e., 180 minutes for a 60-minute recording.

2. *Coding*: the data were then read systematically on a line-by-line basis, with each relevant "chunk" of data assigned a particular code. The coding strategy was *deductive* (Braun & Clarke, 2006) in that it was informed by previous studies in assessor cognition that the researcher had encountered in the literature review. However, in line with flexibility inherent in TA, the researcher also identified new codes that had not been anticipated.

One distinction drawn in thematic analysis is whether the approach to coding is *semantic* or *latent*. In semantic coding, the researcher is largely concerned with the "explicit meaning" of the data, while latent coding is concerned with the "implicit meaning" behind the data (Terry et al., 2017, p.11). Researchers have noted that "one approach is not inherently superior to the other", and that those employing TA need to ensure their coding approach aligns with the research question (Terry et al., 2017, p.12). The approach taken for this TA was largely semantic, as the researcher aimed to determine the explicit mechanisms by which assessors made judgements about students, rather than the implicit philosophies and assumptions that may have guided these judgements. However, while being immersed in the data, a series of latent codes were identified in relation to how assessors interacted with the video technology used in the study. These codes were grouped together in the sub-theme *differences between video and face-to-face assessment*, discussed in the following chapter. Additionally, in addressing RQ3, the researcher analysed two transcripts in-depth, in order to link the processes these two assessors went through when judging the recorded performances with the scores they awarded. At this stage of analysis, the researcher re-coded some of the data in order to identify the latent meanings behind what participants said. More details of this can be found in section 3.7.3.

Data saturation was determined to have been reached after 11 interviews, as no new codes were identified in the 11th transcript. This was confirmed by the 12th transcript, in which no new codes were identified. A coded except from interview 4 can be found below:

Researcher: When you're grading someone how do you know if they're doing well?

Interviewee 4: How do I know that they're doing well? I would know the literature, the research behind it. Right. And then I would have my own experience as well from doing OSCEs. And I'll have myself prepared as well. And we have videos of the... Because the technique can slightly change over the years. So we have up to date videos. So I'd have my OSCE template.

So I have myself prepared and from my own experience as well. Safety would be a big one with safety, that they don't do any harm, any damage to a patient. And I suppose the communications, the verbal or nonverbal, the proficiency.

Now, how would I know that they're doing well? I suppose that their demeanour. and you know that they're confident, their steps, speaking to the patient and talking to the patient and, you know, they're talking to the patient that it's not talking at it. Or you know, “next, and this next”, they're checking on the patient that they're okay, they are caring, compassionate. If there is a human being that they're actually dealing with instead of being just focused on the next step, this step, and you can tell if they've practiced more that they're able to integrate all that.

Codes: theory / experience / safety / communication / not just step by step / confidence

3. *Theme search*: the initial coding process resulted in the identification of 58 codes over the 12 transcripts. These codes were then organised into broader themes in an iterative process. At this stage, seven themes were identified. An example of an initial theme is found below.

Theme label: **performance comparisons**

Example codes:

- *previous student*: “The other girl, I think, was very skills confident, not so much social confident, whereas this girl is both” (Interview 3)
- *self as patient*: “So if I was a patient, I would have felt out of place. I would have just felt like I'm a mannequin. Like someone is just experimenting on me and just leaving me out of my own show” (Interview 2)
- *real world*: “And that's really what I'm looking for is that the standard here is the standard there in clinical practice” (Interview 8)

Theme description: this theme described how students’ performances were interpreted by assessors by comparing them to a range of different people, approaches or scenarios; for example, in comparison to the previous student in the study who performed the same skill, by imagining what it would be like if they (the assessors) were the patient, or by comparing a student’s performance against what would be expected of them in clinical practice.

4. *Theme review*: At this stage, the researcher reviewed the identified themes to assess how well they reflected the complexity of the data. After this review, the researcher chose to group the identified themes and sub-themes according to the three-stage framework of decision-

making outlined by Gauthier et al. (2016) and discussed at length in Chapter Two. As such, three main themes were identified, with each theme corresponding to a stage in Gauthier et al.'s (2016) conceptualisation of assessors' cognitive processes: *observation, processing* and *integration*. A fourth theme, *assessor development,* was also identified, relating to medium- and long-term factors which participants discussed as affecting their decision-making, but which did not correspond to any of the three-stages identified in the 2016 study.

5. *Theme definition*: The fifth stage involved the researcher defining each of the four themes. Given that three of the themes were derived from an existing paper, the definition of these themes followed those described by Gauthier et al. (2016). The researcher also thought about the "essence" of the fourth theme, *assessor development*, and decided on a definition of it. A detailed description of the four themes, and relevant sub-themes, can be found in the following chapter.

6. *Write up*: For the final stage, the researcher wrote up his findings. These are presented in section 4.4 in the following chapter. As noted by Braun and Clarke (2022), researchers undertaking TA have to strike a balance between presenting data verbatim from the transcripts, and analysing and contextualising that data. In line with their recommended approach, the researcher aimed for an approximate 50/50 split between original data and interpretation of that data.

3.7.2 *RQ2: Quantitative analyses*

The aim of the second research question was to determine whether there were notable differences in how the four recorded videos were graded by the 12 assessor participants. In order to address this question, a series of descriptive statistics was provided. These statistics were first detailed for the *checklist items* for each of the four videos, in order to determine whether there was identifiable variance in how many checklist items were ticked by assessors. A similar approach was taken for the three *global items* at the end of each of the four marking guides. Given that the pass/fail decision is usually the most important one associated with undergraduate OSCEs (Khan et al., 2013b), a brief summary of whether the four recorded videos were deemed to be at the passing level by the 12 assessors was provided.

Subsequent to this, descriptive statistics were provided for the *summated* checklist and global items for each assessor, to allow for an identification of whether some participants were consistently harsher or more lenient than their counterparts. Determining "hawks" and "doves" is an important step when conducting reliability analyses (Bartman et al., 2013), and additionally was used to inform the outlier analysis which is detailed in section 3.7.3.

Finally, IRR was calculated for each item on the binary checklists, using the *percent agreement* statistic. This simple calculation allowed for an identification of which items on the marking guide caused divergence between assessors, and follows the same procedure outlined by Dunbar (2018). This calculation was made by determining what percentage of assessors ticked (or did not tick) a specific item on the marking guide. For example, if all 12 participants ticked item 1 for video P01NG, then the percent agreement statistic is 1 (perfect agreement). Likewise, if all 12 participants failed to tick the item, the percent agreement is also 1. Possible figures for percent agreement range from 0.5 (lowest possible agreement) to 1. As noted by Gwet (2014), a figure of less than 0.75 is deemed unacceptable; all items which resulted in such a figure were noted. Results from this phase of analysis can be found in section 4.5 of the following chapter.

3.7.3 *RQ3: Outlier analysis*

The aim of the final research question was to explore whether any links could be drawn between the processes that assessors went through while judging the recorded performances, and the scores they awarded. This would allow for a deeper understanding of how divergence in awarded scores arises, and thus how the reliability of scores is threatened. The approach to data analysis taken for this question was *outlier analysis* (Teddie & Yu, 2007, p.81), which involves "selecting cases near the 'ends' of the distribution of cases of interest" for further examination, with the expectation that these cases "yield especially valuable information about the topic of interest".

For this analysis, the researcher used the quantitative data outlined in the previous section to identify both the harshest and most generous participant. He then went back to the qualitative data to determine whether any distinct differences could be identified between how each of these two assessors approached the task of assessment, and whether these approaches could be linked with divergent outcomes in terms of awarded scores.

This phase of analysis involved a partial re-coding of some of the data (Braun & Clarke, 2022), to uncover more nuance in the potential connection between the process of judging students and the result of that process in terms of awarded scores. For example, one of the initial codes identified after the thematic analysis was *limitation of guide,* which was labelled to data extracts in which assessors expressed frustration at the guide "forcing" them to assess the students in ways they did not want to. At this stage of analysis, this code was re-classified into two codes:

limitation of guide (harsh) and *limitation of guide (lenient).* The former described instances where assessors felt that the marking guide led them to awarding lower marks to students than they felt the students deserved, while the latter described instances where it led to higher scores. This re-coding resulted in the finding that the harsh assessor was more likely to describe the guide as resulting in marks that were too high, while the opposite was true for the generous assessor. This phase of analysis also involved the identification of latent codes, which were used to identify the assumptions underlying assessors' decisions. For example, a latent code that was identified for each assessor was *confidence*. Identifying these codes allowed for an explanation as to why these assessors might have awarded different scores to the same videos. In this case, the harsh assessor used confidence in her decision-making in order to justify awarding low scores to students, while the lenient assessor invoked a lack of confidence which resulted in her giving the benefit of the doubt in some instances, and thus awarding marks to students even if she was not convinced they had completed an item properly. Results from this phase of analysis can be found in section 4.6 of the following chapter.

3.8 *Pilot Studies*

In addition to the three-phase study outlined in detail above, the researcher conducted a pilot study in January 2022, with one of the participants from Phase 1 of the data collection. The aims of the pilot study were as follows:

- Discuss the interview guide for the semi-structured interview, to assess whether any questions needed to be altered in order to ensure participants provided lengthy answers.
- Determine any issues that may have arisen during the initial cognitive interview.
- Discuss how the think aloud protocol functioned and identify changes to the prompts that had to be made.
- Determine whether completion of the think aloud while simultaneously filling out the marking guide was likely to be too cognitively demanding for participants.
- Identify any technological problems that may have arisen during the think aloud.
- Estimate how long each stage of the interview was likely to last, and in particular how long the think aloud would take in relation to the length of the videos.
- Discuss whether the participant had strong opinions about which videos should be included in order to maximise the chance for variance between assessors to emerge.

The results of this pilot study affected several decisions made when setting up the research design, as discussed throughout section 3.6. Once these decisions had been made, data from the pilot interview were excluded from the rest of the study.

3.9 *Ethical considerations*

Ethical approval for the study was granted on April 7th, 2021 (Appendix I), and an amendment was approved in December 2021 (Appendix J). A number of ethical issues required attention, most notably the capturing and storage of video data of student participants. The use of personal data such as video footage is controlled as part of DCU's Ethics Guidelines (2019), and the use of video data automatically classified the study as requiring an expedited review.

A number of procedures were put in place by the researcher to ensure that video data were stored securely and only used for the intended purpose of the study. Once the OSCE videos were recorded, the researcher immediately stored them in a password-protected folder on his laptop, with backup copies on DCU's secure Google Drive. Only the researcher and his three supervisors had access to these files. The video files were shown to interview participants on the researcher's laptop, thus the files did not have to be moved from this laptop in order for the interviews to take place.

The three student participants who took part in the video recordings were informed they could request a copy of their videos if they wished, however none did so. The videos – along with other data collected as part of the project – will remain on the researcher's laptop for the duration of the doctoral thesis and for five years afterwards. Subsequent to this, the researcher will delete these files permanently, both from his own laptop and from Google Drive. It is likely at this point that the hard drive on the researcher's laptop will have been wiped, as he will no longer be a student at DCU.

A second ethical concern related to the issue of test anxiety (Howard, 2020) – because student participants had to complete two OSCEs that approximated a high-stakes examination they undertook as part of their degree, it was possible that they might become anxious or stressed. In order to mitigate this possibility, the researcher made it clear to the students that it was of no consequence how well they performed, that no one would be in the room grading them, and that they could take a break or cease participation at any point they liked. The researcher's supervisor, Dr. Mary Kelly, who assisted in the recording of the OSCEs, stepped out of the room while students were completing the OSCEs, so that her presence would not cause them anxiety.

There were other less pertinent but still notable ethical considerations. Audio data collected were subject to the same protection procedures as video data, with the caveat that participants were unlikely to reveal any personal or compromising information about themselves. Physical data, such as completed marking guides, were kept in a locked drawer in the researcher's office, which was itself kept locked when unoccupied. Physical data will be shredded and disposed of upon completion of the project. Participants at all stages of the study were informed about the study with a Plain Language Statement (Appendix A) and had to complete an Informed Consent Form (Appendix B).

Chapter Four

Results

4.1 *Introduction*

This chapter presents the findings of the current study in three sections. As documented in Chapters Two and Three, the present mixed-methods study aimed to address pertinent research questions about assessor cognition and score reliability as they relate to undergraduate nursing OSCEs. Developed from the gaps in the literature identified in Chapter Two, the three research questions were as follows:

> RQ1: *What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?*
>
> RQ2: *What level of inter-rater reliability is evident in undergraduate nursing OSCEs?*
>
> RQ3: *Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?*

These questions are addressed one-by-one in the present chapter. The first question was explored using purely qualitative data, analysed through thematic analysis. The second question was addressed using purely quantitative data. The final question was investigated using a mix of qualitative and quantitative data.

Figure 4.1 summarises the data that were collected for the study, and how they were used to answer the research questions. The green section at the top refers to the four videos that were used for the think aloud protocols. The number after the P relates to the year of study that the student participant was in when she was filmed (e.g. P01 refers to a first-year student). The letters after this number denote the OSCE that was completed (NG = naso-gastric tube insertion, BP = blood pressure). For example, P01NG denotes the video in which the first-year student participant completed the naso-gastric tube insertion OSCE. All four videos were watched by 12 assessor participants, who additionally completed a semi-structured interview and a cognitive interview pertaining to the marking guides for the two OSCEs. All assessor participants also completed the relevant marking guide for each of the four videos, which comprised a series of binary checklist items, as well as three global items. A copy of the marking guide that was used in the study can be found in Appendix H. The orange boxes in the table indicate that a specific piece of data was used to address one of the research questions (e.g., data from the semi-structured interviews were used to address RQ1 and RQ3).

Figure 4.1 *Summary of data collected*

Students (*n* = 3)	**P01**	**P02**	**P03**	
Videos	P01NG	P02BP	P03BP	P03NG
Assessor participants (watched all videos)	*n* = 12			

	RQ1	**RQ2**	**RQ3**
Data source		**Qualitative data**	
Semi-structured interview	Y		Y
Marking guide cognitive interview	Y		
Think-aloud protocol	Y		Y
		Quantitative data	
Binary checklist items*		Y	Y
Global questions**		Y	Y

** The blood pressure measurement OSCE contained 12 items, while the NG tube OSCE contained 14.*

*** Participants were asked three global questions at the end of each video, pertaining to the student's communication, entrustability, and overall performance.*

4.2 *Organisation of the chapter*

The volume of information collected during this study resulted in significant planning on the part of the researcher regarding how best to tell the "story" of the data. The structure of the present chapter went through numerous iterations before the current layout was finalised. This reflects the difficulty, inherent in mixed-methods research, of trying to build a coherent narrative to describe phenomena that are often complex and multifaceted. The layout of this chapter thus represents the optimal way of presenting the data such that the research questions are addressed in an engaging and logical manner.

Section 4.3 details demographic information about the 12 assessor participants. Section 4.4 outlines the findings of the thematic analysis, which allows for a description of *how* the assessor participants approached the task of assessment (RQ1). As documented in Chapter Two, this issue has been investigated regarding assessors in medicine, but is largely unaddressed in nursing assessment. The potential implications of these findings regarding the core assessment principles of validity and reliability are flagged throughout this section, and are unpacked and discussed further in Chapter Five. It is noted at this juncture that idiosyncrasies in assessors' cognitive processes as they judged the recorded performances were likely to impact the inter-rater reliability (IRR) of the awarded scores. However, this information alone does not constitute *evidence of* low IRR, which can only be ascertained through quantitative examination of these scores. This quantitative examination takes place in section 4.5, which provides details of the scores awarded by the 12 assessor participants to the four videos (RQ2), and responds to calls in nursing research for continued investigation into the reliability of scores awarded in undergraduate nursing OSCEs (Navas-Ferrar et al., 2017; Goh et al., 2019). As such, sections 4.4 and 4.5 together represent the entire breadth of the collected data, detailing the results of the qualitative and quantitative data across all 12 assessor participants.

Section 4.6 uses outlier sampling (Teddie & Yu, 2007) to focus on two assessor participants, examining the data they provided in depth to explore whether any links can be made between the cognitive processes they went through when judging the recorded videos, and the grades they awarded to the student participants. As noted in the literature review, assessor cognition research has been largely qualitative in nature, while investigations of IRR have been quantitative. The section builds on work by Gingerich et al. (2014b, 2017) to combine these approaches, in order to begin to draw causal links between processes and outcomes. When the quantitative data were analysed, it was evident that specific assessor participants were particularly harsh or lenient compared to their colleagues: this section describes how these participants approached the task of assessing students, and the resultant implications in terms of the reliability of the awarded scores. Section 4.6, therefore, represents an exploration of the collected data in as much depth as possible.

4.3 *Demographic information*

In total, 12 assessors participated in the study. One of the challenges in writing the current chapter was to give the reader enough information about these 12 people in order to inform the subsequent pieces of data analysis, but without compromising their anonymity by making them identifiable. As such, the approach taken here was to group participants into distinct "bands", particularly in terms of their years of experience. For example, instead of saying that someone has exactly one year of experience, they were grouped into the category "< 5 years". Additionally, a decision was made not to classify specific assessors based on whether they had a clinical role or an academic role. Their age is also not detailed. All participants had experience in the teaching and/or assessment of both blood pressure measurement and naso-gastric tube insertion. Assessors were asked to classify themselves in terms of their ability to assess OSCEs, as either Advanced Beginner, Competent, Proficient or Expert (derived from Benner (1982)). The results of this information can be found in Table 4.1 below.

Table 4.1 *Demographic information of assessor participants*

	Gender	**Years in Role**	**Ability**
Assessor Participant 1	Female	< 5 years	Advanced Beginner
Assessor Participant 2	Female	< 5 years	Competent
Assessor Participant 3	Female	< 5 years	Competent
Assessor Participant 4	Female	> 5 years	Proficient
Assessor Participant 5	Female	> 5 years	Competent
Assessor Participant 6	Female	> 5 years	Proficient
Assessor Participant 7	Female	> 5 years	Proficient
Assessor Participant 8	Male	> 5 years	Proficient
Assessor Participant 9	Female	> 5 years	Proficient
Assessor Participant 10	Female	> 5 years	Expert
Assessor Participant 11	Female	> 5 years	Expert
Assessor Participant 12	Male	> 5 years	Proficient

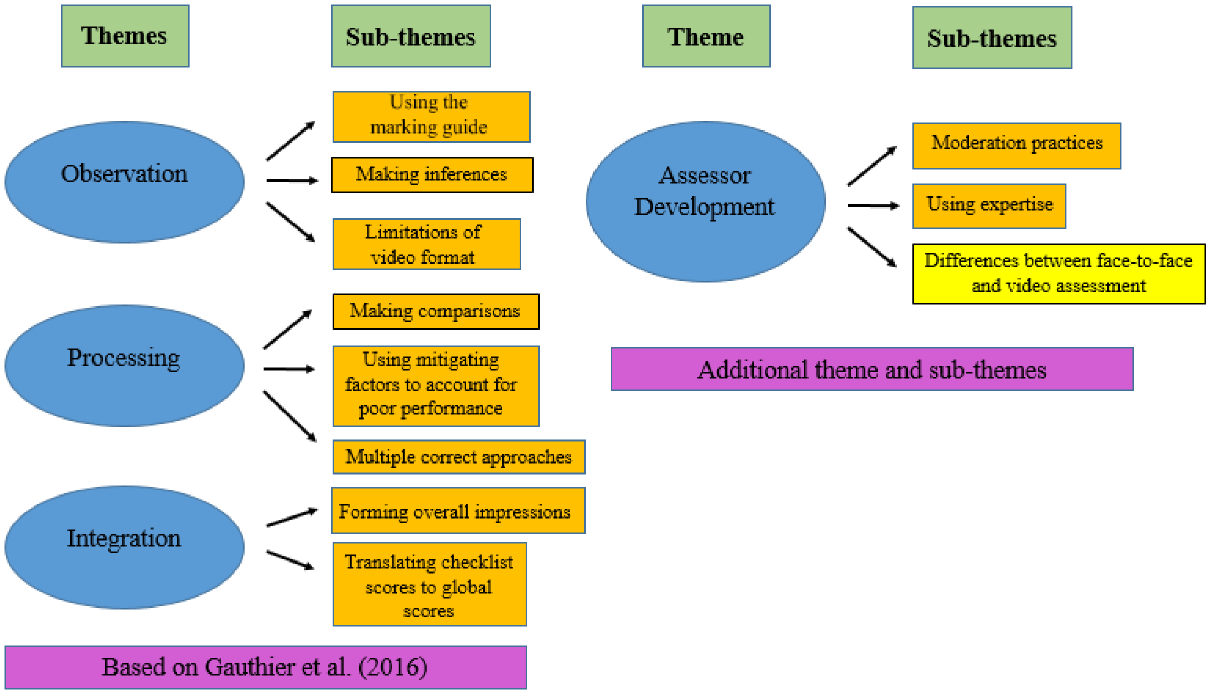
The interviews ranged in length from 50 minutes to 67 minutes. In total, 726 minutes of audio was recorded, for an average of 60.5 minutes per interview. Quantitative statistics were calculated using Microsoft Excel, while NVivo was used for qualitative analysis.

4.4 *RQ1: What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?*

4.4.1 *Introduction*

Regarding RQ1, data from the thematic analysis is presented, organised into four themes (following Braun and Clarke's (2006) guidelines for thematic analysis). The first three themes relate to the cognitive processes that assessors employed at the *observation*, *processing* and *integration* stages of decision-making, in line with the framework described by Gauthier et al. (2016) and discussed in Chapter Two. The final theme, *assessor development*, relates to longer-term factors that affect assessors' judgements. In line with the approach suggested by Braun and Clarke (2022), approximately half of this section contains data, and the other half is the researcher's interpretation and contextualisation of that data. After the explanation of each sub-theme, there is a brief discussion of how it relates to the existing literature, and any potential issues in terms of score validity and reliability are flagged. An overview of these themes can be found in Figure 4.2 below. The overall theme is shaded in blue, with sub-themes in orange if they are semantic (sometimes referred to in the literature as *manifest*) and yellow if they are latent. As noted in section 3.7.1, this stage of analysis focused largely on the semantic meaning of what assessors said; it is for this reason that the majority of identified sub-themes are semantic.

Figure 4.2 *Overview of identified themes*

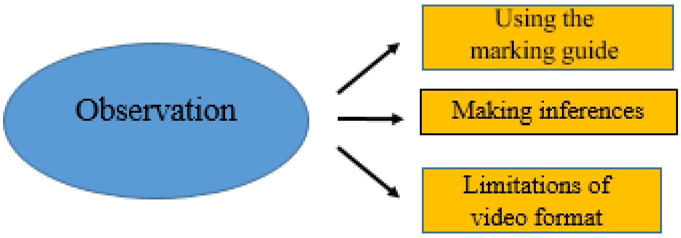

Themes
Sub-themes
Theme
Sub-themes
Observation
Using the marking guide
Making inferences
Limitations of video format
Processing
Making comparisons
Using mitigating factors to account for poor performance
Multiple correct approaches
Integration
Forming overall impressions
Translating checklist scores to global scores
Based on Gauthier et al. (2016)
Assessor Development
Moderation practices
Using expertise
Differences between face-to-face and video assessment
Additional theme and sub-themes

Taken together, these four themes, and related sub-themes, provide an explanation for how the assessor participants formed judgements about students who complete undergraduate nursing OSCEs. It is noted at this point that analysis of the qualitative data was carried out before, and independently of, the quantitative data; it is for this reason it is presented first. Additionally, in line with similar research in assessor cognition where multiple qualitative methods are used in the same study (e.g., Yeates et al., 2013), all qualitative data were analysed together and are presented together in this chapter. As such, it is noted that participants spoke about assessment in general terms, as well as in relation to the specific videos that were used in this study. For the sake of clarity, every time a quote from an assessor describes something happening in a specific video, it is mentioned what video they are referring to. Otherwise, their comments relate to their assessment practices more generally.

4.4.2 *Observation*

This theme describes factors that guided participants' observations of student performances. As outlined in Figure 4.3, there were three sub-themes identified relating to the larger theme of observation: *using the marking guide*, *making inferences* and *limitations of video format*. These three sub-themes are discussed in turn below.

Figure 4.3 *Observation*



4.4.2.1 *Using the marking guide*

All 12 assessor participants mentioned the marking guide as being the primary means by which their observations of the video recordings were directed. Participants stressed that the marking guide was an essential tool for ensuring that different assessors can judge student performances in a way that is broadly consistent with each other:

> I think if they're not doing it right, it's very obvious because you've got your steps and your steps are quite tight and you know what you have to follow and you know the mark kind of associated with it. So I think you've got less of the grey area.
>
> Assessor 9

In this excerpt, the assessor described the process of observing a student's performance as one of following the steps contained in the guide. In her opinion, the steps in the guide are "quite tight", allowing minimal room for a "grey area", as all assessors have to use the same set of criteria to guide their observations.

Five assessor participants also discussed the importance of standardised procedures which describe how to perform a specific skill in line with current best practice. These standards inform the development of the marking guides used in the present study, and additionally allow assessors to watch the videos and know whether the student is following the correct steps:

> I suppose myself, knowing the standards, we use the Royal Marsden clinical procedures... So everything is aligned the standard way. A lot of our skills are just standard. So I suppose I know by that whether or not [students are performing well]. I kind of know them off by heart, I suppose being an experienced nurse as well, I just kind of keep myself up to date with my standards. That's probably what guides me with doing it right around.
>
> Assessor 3

The fact that all 12 assessor participants mentioned the marking guide as being the primary means by which they direct their observations of OSCE performances is notable from an assessment perspective. A well-designed marking guide has the explicit purpose of creating uniformity in what assessors look for as they judge student performances (Khan et al., 2013b). Given that all assessors in the present study were explicit about using the marking guide in this way, it would be logical to assume that this would increase the inter-rater reliability of the scores they awarded. This reliance on the marking guide is not one that is consistently found in similar studies of assessor cognition. For example, in interviews with nursing assessors, East et al. (2014) found that "overwhelmingly assessors determined students' competence subjectively" (p.463), openly admitting to ignoring the provisions of the marking guide if they felt it appropriate. This was echoed in a recent study by Hyde et al. (2022), who found that medicine assessors were also explicit about their tendency to deviate from the marking guide, even if it meant that their judgement would be different from those of their colleagues (which would adversely affect the reliability of the assessment scores). As such, the present study indicates that nursing assessors used the marking guide to shape their observations of students, in a way that should bring about a high level of score reliability.

However, assessor participants also acknowledged that specific items on the marking guide were especially important for them, and that they were likely to direct their observations towards these items in particular. Some participants mentioned psychomotor criteria as being particularly important for them, while others mentioned communication or affective criteria:

> …the “length of tube to be passed, the distance from the nose to earlobe, then earlobe to the lowermost section of the breastbone. Must be clearly visible on video.” [item 5 on NG OSCE] Yeah. That's very important. That's something I would look out for and kind of, if I needed to stop and start the video, I would do it for that part. It has to be right. And with that one, there'll be no, "I don't know, maybe I'll give it to them." It's either you got it or you don't.
>
> Assessor 11
>
> Then “explain procedure to the patient and gain consent” [item 3 on BP OSCE]. I'm looking out, definitely looking out for this because I can't over-emphasize this enough. In every one of my classes, I make sure to say it and drill it and keep drilling it.
>
> Assessor 2

These two quotes suggest that specific assessors are likely to have their own ideas about items in the marking guide that are especially important, even if these items are afforded the same amount of marks on the guide as the other items (as was the case in this study). These ideas about what is important are then used to direct their observations of the videos, such that they are likely to focus more on the items they perceive to be important. This phenomenon has been identified in numerous other studies discussed in Chapter Two (e.g., Yeates et al., 2013; East et al., 2014), which found that assessors are likely to direct their attention towards different dimensions of the assessed skill, based on what they believe to be the most important aspects of that skill. There are notable implications of this phenomenon in terms of score reliability: a student who displays good psychomotor skills but poor communication could receive a much higher score from an assessor who values psychomotor skills than they would from an assessor who values communication skills. Their score would then be overly affected by who happened to be assessing them, rather than by their ability level. Indeed, this issue has been discussed in the literature in a study by Chahine et al. (2016), who found that assessors could be “grouped” into clusters based on what aspects of OSCE performance they believed to be most important, and that the scores awarded by the assessors in each cluster were measurably different.

4.4.2.2 *Making Inferences*

During the observation stage, ten assessor participants made inferences about students that went beyond what was visible in the videos. These inferences usually had to do with the student’s experience (e.g., whether they had worked in the clinical setting before). In some cases, making an inference about one of the students in the video was used to explain why they did not complete a specific task in the correct way:

> The patient and herself seemed to have discussed this so consent was obtained before this procedure, and she's using his first name. So I make an assumption that he has consented to that and that they have that rapport.
>
> Assessor 6

In this extract, relating to the video P03NG, the assessor noted that the student participant obtained consent from the patient too quickly; however, she believed that the student and the patient had conversed some time before the video began and discussed the procedure. As such, even though she believed that consent was not obtained in the best way, the assessor was happy to award the student the mark for this step on the guide, due to an inference made about events that may have taken place before the beginning of the video.

In other cases, assessors made an inference based on their observation of a student, but were explicit that this inference did not affect their decision to award marks to the student:

> She's probably a nursing assistant somewhere. I might be first to ask her, "have you worked somewhere before?" Yeah, but of course, part of what I want to score in the OSCE is demonstrating that skill of getting the estimated pressure. So asking her is just by the way, that's not going to make me give her the marks for doing that. Of course, my curiosity will just make me ask, "have you worked somewhere before? Have you done this before?"
>
> Assessor 2

In this extract, relating to the video P03BP, the assessor participant made an inference that she believed the student has worked (or is working) as a nursing assistant, and that it is this experience that has caused her to skip taking an estimated systolic blood pressure reading (item 9 on the marking guide for the BP OSCE). However, in spite of this explanation, the assessor was clear that this perceived experience did not justify her skipping the step, and so did not award her the mark.

The issue of inference formation is one which has implications for validity and reliability. As noted by Gauthier et al. (2016), there is a divide in the literature between researchers who problematise inferences as having negative effects on score reliability (e.g., Kogan et al., 2011), and those who believe that experienced assessors can make "high level" inferences about students that accurately capture their abilities (e.g., Govaerts et al., 2013). The potential effects of inference formation in the present study in terms of reliability are twofold. Firstly, not all assessor participants made inferences, with some basing their observations solely on what was visible to them in the recorded videos. Secondly, when inferences were made about students, they were sometimes used to justify awarding marks that otherwise may not have been awarded if an inference had not been made. However, in other cases, assessors made an inference about a student but did not use this inference to change the awarded scores. As such, inferences were likely to cause inconsistencies in how the recorded videos were graded by the 12 assessors. Additionally, there was no discernible profile in terms of which assessors were likely to make inferences, with both experienced and inexperienced assessors equally likely to make them. From the present study it is therefore impossible to conclude that it was only experienced assessors who made inferences, and that these inferences were uniformly of a "high level" (Govaerts et al., 2013).

4.4.2.3 *Limitations of video format*

Ten of the assessor participants mentioned the limitations of the video format, specifically in terms of being unable to see clearly certain aspects of a student's performance. These discussions took place regarding the use of video in general, as well as in relation to the specific four videos used in this study. Assessors mentioned that, when assessing OSCEs in the past, they have had to grade video performances that were filmed in such a way so that some aspects of the assessed procedure were not visible to them. This would occasionally result in assessors having to guess whether a student had completed a specific step on the guide:

> Sometimes it can be due to a camera angle as well, and the quality of the video sometimes doesn't be that good. So you are kind of guessing, "did they or didn't they?"
>
> Assessor 11

However, some assessor participants noted that in spite of the potential limitations of the video format in terms of allowing them to observe all aspects of a performance, the students could make up for this by narrating what they were doing. In this way, students could still communicate to assessors what they were doing, even if the video could not show it in detail:

> With the patients [in the NG OSCE], sometimes you can get curling of the tubes at the back of the throat and so forth. You might not see that all the time with the mannequin,

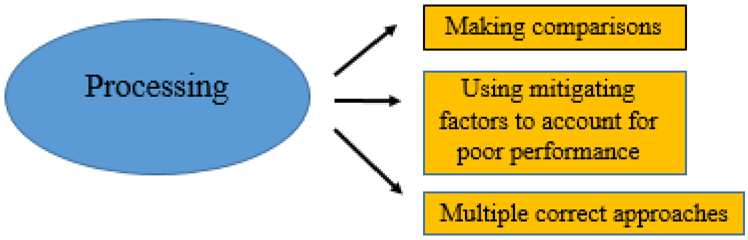
> it might just have one way down and it just goes down. So that can be a little bit difficult to see. But if they're vocalising that to you, they may say, well, when I'm sliding it in, I want it to go upwards and backwards, inwards and backwards, and then they know the basis of it.
>
> Assessor 9

The issue of assessors' observations of recorded OSCE performances being limited by the video format is not one that has received significant attention in the literature. However, a series of articles published since the onset of the COVID-19 pandemic in March 2020 have begun to highlight this as a potential negative effect of moving the assessment online (e.g., Major et al., 2020). The potential implications of this issue in terms of score reliability are notable, as assessors may have different strategies for choosing to award (or not to award) a mark for an item on the marking guide that they could not see the student performing clearly. This issue is discussed in more detail in section 4.6, when the contrasting views of two of the assessors who took part in the present study, and the resultant effects on score reliability, are described.

4.4.3 *Processing*

This theme relates to how assessor participants were able to determine whether a student had performed well in the recorded videos. As outlined in Figure 4.4, there were three sub-themes identified relating to the larger theme of processing: *making comparisons*, *using mitigating factors to account for poor performance*, and *multiple correct approaches.*

Figure 4.4 *Processing*


Processing
Making comparisons
Using mitigating factors to account for poor performance
Multiple correct approaches

4.4.3.1 *Making comparisons*

All 12 assessor participants used comparisons to inform their judgements of how well or badly a student was performing in the recorded videos: both in relation to their overall performance, and their ability to carry out a specific item on the marking guide. Most frequently, assessors compared a specific student's performance with that of the previous student who had completed the same skill as part of the study:

> She would have been a little bit better than the last girl because she expanded a little bit on what she was doing: “Are you okay?” And she explained the actual pressure which is important, that your blood pressure is normal. So she'd be a little bit better than the other. But the communication with the two of them are just average... So there's just a basic concise information. So what would I give? She's a little bit better than the other one. It's reasonably good, but it's bare minimum.
>
> Assessor 4

In this extract, related to the video P03BP, the assessor participant drew a comparison between the performance of the student in the video, and the other student who completed the blood pressure measurement OSCE (P02BP). Here, the assessor clearly used the P02BP video as a benchmark in her interpretation of how the student in the P03BP video was performing, noting twice that the latter was doing “a little bit better” than the former in terms of her communication skills. She then used this to justify awarding a higher mark for communication to the student in the P03BP video.

Assessors also used themselves as a benchmark when interpreting a student’s performance, by drawing a comparison between what the student did, and what they (the assessor) would do if they had to carry out the same skill:

> I don't think she's finding a brachial artery there for me. I wouldn't look for a brachial artery there.
>
> Assessor 9

Here, the assessor noted that the student’s (P02BP) location of the patient’s brachial artery is incorrect, by contrasting the student’s effort to locate the artery with what she (the assessor) would do herself. The five assessors (not identified in the extracts below for reasons of anonymity) who were involved in the teaching of clinical skills to students also compared the performance of the students in the videos to what they would have taught in class:

> And of course, as a clinical skills instructor, you want to demonstrate it first and you get to know they're doing the right thing if they replicate what you've taught them.
>
> Assessor X

> I'd give her a fail. I think it's like a student who didn't know the steps at all. Yeah, didn't know the steps. Didn't know the correct procedure. Didn't demonstrate safe practise. And it’s very evident there because that's not what she was shown in class.
>
> Assessor Y

In the first extract, the assessor participant was explicit about the fact that she knows whether a student is performing well or not based on the similarity of that student’s performance to what the assessor would have taught in class. Likewise, in the second extract, the assessor justifies her decision to award a failing mark to the student (P02BP) on the basis of the disparity between what the student did and what she would have been taught in her clinical skills classes.

The phenomenon of assessors drawing comparisons in order to interpret a student’s performance has been widely noted in the literature on medical assessment (e.g., Kogan et al., 2011; Yeates et al., 2013). Yeates et al. (2015) found that assessors’ OSCE scores are influenced by the standard of performances seen directly before: if an assessor judges a series of highly competent students in a row, a subsequent student may receive lower scores than they otherwise would have, due to being compared unfavourably to these high-performing students. The present study is novel in that it is among the first to document these comparisons as they manifest in undergraduate nursing OSCEs. The potential implications of this in terms of score reliability are twofold. Firstly, a student’s score may be affected by who happened to be assessed directly before them. In other words, the same performance might receive two different scores due to the skill level of the previous student(s), rather than the skill level of the assessed student. Secondly, the tendency of some assessors to compare a student’s performance to what they would do themselves means that a student’s score might be affected inordinately by the personal opinions of whatever assessor happened to be judging them. The frequency of performance comparisons therefore threatens the validity of decisions made on the basis of OSCE scores, as it would be impossible to say with certainty that these scores were not influenced by the order in which students were seen by assessors, and the assessors’ own ideas of how they would complete a skill.

4.4.3.2 *Using mitigating factor to account for poor performance*

The second sub-theme relates to the tendency of participants to be lenient in their interpretation of a student’s performance, either because of reflexivity on the part of students, the artificiality of the OSCE environment, or the perceived inexperience of the students. Five assessor participants mentioned the possibility for students to compensate for a poor performance by expressing awareness of the mistakes they made:

> Because they might say, “oh, yeah, I know. I stopped in the middle of that where I shouldn't have or I was just nervous when I dropped that onto the sterile field or whatever.” And they acknowledged where they've gone wrong. At least that might push them back into the pass range where you would have to kind of give them that nerves thing. But that only happens in face to face.
>
> Assessor 9

In this extract, the assessor discussed how a student demonstrating knowledge of the mistakes they had made while completing an OSCE might be the difference between them receiving a pass or fail grade. Four assessor participants noted that, when assessing in-person OSCEs, they would ask the student at the end how they thought they had performed, and those that had awareness of their mistakes would be graded less harshly than those that did not. However, as noted by the above assessor, this can only take place when the OSCE is delivered in-person (as opposed to via video), as the video format does not allow for assessors to question students at the end of the exam. The differences between face-to-face and video assessment are discussed further in section 4.4.5.3.

Ten assessor participants also mentioned the artificiality of the OSCE environment as influencing their decisions regarding a student’s performance:

> And you have to bear in mind that if a student is carrying out a clinical procedure like that on a mannequin, to understand that that can be a difficult thing for them, to achieve a kind of rapport. And not every one of us are born actors. You know what I mean?
>
> Assessor 6

Here, the assessor discussed the difficulty for students who are performing an NG tube insertion on a mannequin to demonstrate appropriate communication skills. She noted the fact that not all students will be able to achieve the necessary “rapport” with the dummy, and that she would not penalise a student too harshly for this. A similar idea is found in the following excerpt:

> Communication is more about a commentary to us rather than to the patient. She's telling us what she's doing rather than communicating hugely with the patient, [but that doesn’t bother me] because it's a dummy.
>
> Assessor 5

The assessor in this case explicitly referred to the fact that the student (P01NG) is performing the required skill on a mannequin, and therefore that she (the assessor) is not judging the student’s communication skills too harshly. In both cases, the artificiality of the exam setting, specifically the fact that students have to interact with a mannequin, served as a mitigating factor when students’ communication skills were assessed.

Finally, three assessor participants also reported being lenient in their interpretations of student performances due to the perceived inexperience of the students at performing the assessed skill. These assessors mentioned that students could not be expected to be perfect at the skill due to being at an early stage in their education. As such, they would not judge a student too harshly if they felt the student had not done something in the correct way. This was particularly relevant when students’ psychomotor skills were remarked on:

> I think something that's very psychomotor related that you can make an assumption about someone and their clinical skills… their psychomotor might not be particularly good but their cognitive and their communication, their affective domain is very good quality. And you know that this student would not have huge practice, and you have to factor that in. But what you're looking for is do they have the capacity.
>
> Assessor 6

In this excerpt, the assessor commented on the fact that a student’s poor psychomotor skill level could be attributed to the fact that they may not have had a great deal of opportunity to practice the assessed skill. This assessor would be willing to overlook poor psychomotor skills as long as the student was strong in the affective domain, which for her indicates the “capacity” for improvement.

The presence of various mitigating factors affecting assessors’ decisions is likely to impact score reliability, especially given the inconsistency with which they were used across the assessor pool. For example, only three of the 12 assessors mentioned that they would be lenient in their judgement of a student based on the perceived inexperience of that student. This is suggestive of a potential discrepancy in how these three assessors would grade a student compared to their counterparts. Roberts et al. (2020) noted that the decisions made by assessors in their study were influenced by an assessor's "educational perspectives" (p.9). The results of the present study indicate a similar phenomenon is present among the 12 assessors who participated: their assessment decisions were likely to be affected by the extent to which they believe reflexiveness, artificiality, and inexperience should influence their interpretations of OSCE performances.

4.4.3.3 *Multiple correct approaches*

The final sub-theme in this section relates to the idea that assessor participants' interpretations of student performances were affected by their knowledge that there may be more than one correct approach to completing an assessed skill. Six assessors noted that, for some items on the marking guide, there was no "one-size fits-all" method for completing it. This idea was particularly notable in discussions of communication skills:

> So I suppose if somebody who might just want to get the job done, they might just say, "okay, I'm going to take your blood pressure." Now you might get somebody else who will say "hello, how are you?" going into more detail really to kind of get their consent. And maybe different personalities will bring a different type of communication. And those are both OK.
>
> Assessor 1

In this extract, the assessor participant discussed the process of obtaining consent from a patient before beginning a procedure, a step on the marking guide in both OSCEs used in this study. She stated that a student could be brusque and upfront with the patient, or chattier and "going into more detail", and that both approaches would result in her awarding the mark for that item on the marking guide. While this idea of different approaches was most frequently mentioned regarding communication skills, it was also discussed in relation to psychomotor skills:

> Or if they said "the next thing I would do is take an aspirate". Take an aspirate, yeah, I would give it [the mark] to them, but again, if it wasn't shown but they said it, I would still give the marks. If they didn't say it and they showed me how to do it, then I'll give the marks as well. "Mark the Ng tube with indelible marker." Yet again, they can either show me that or they can tell me they would do it.
>
> Assessor 11

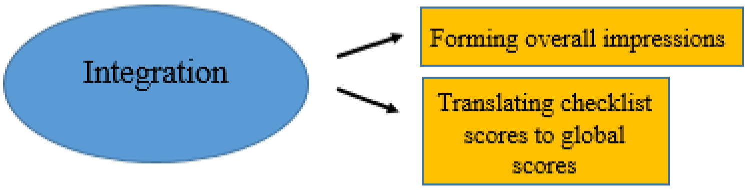
Here, the assessor discussed the idea that, for some psychomotor steps on the NG tube OSCE, she would not even need to see the student completing the step in order to award the mark. As long as the student mentioned that they would do a certain step (such as taking an aspirate) that would be sufficient for this assessor. As such, for this participant there is equivalence (for some items) between saying that a task would be done and showing how to do it. This has the potential to cause a discrepancy regarding how this assessor would grade an OSCE performance compared to other assessors, who may decide only to award a mark if the student actually performs the necessary skill, rather than just stating that they would perform it.

All in all, the findings within the *processing* theme suggest significant levels of idiosyncrasy in how assessors interpret a student's performance. Yeates et al. (2013) described the idea of criterion uncertainty: assessors in their study reported lacking a fixed internal sense of what "good" OSCE performance looks like, and so had to rely on a range of different factors, such as comparisons, in order to determine how well or badly a student was performing. The present study indicates that this idea is relevant to undergraduate nursing OSCE assessors as well. As noted previously, this may adversely affect score reliability, as a student's grade will be unduly influenced by the nature of the assessor who happens to be grading them. Recommendations for how this problem can be addressed are found in section 5.3 of the subsequent chapter.

4.4.4 *Integration*

This theme relates to how assessor participants used all the information available to them to come to a final decision about students' ability levels. As outlined in Figure 4.5, there were two sub-themes identified relating to the larger theme of integration: *forming overall impressions* and *translating checklist scores to global scores*. These two sub-themes are discussed below.

Figure 4.5 *Integration*


Integration
Forming overall impressions
Translating checklist scores to global scores

4.4.4.1 *Forming overall impressions*

The first sub-theme relates to how assessors formed a final impression or judgement about a student's level of competence, in particular by focusing on the student's communication skills, as well as whether they were perceived to have demonstrated safe practice. Ten of the assessor participants reported that they were looking out for a student who was able to integrate effective communication into their OSCE performance. This did not just relate to the checklist items for each OSCE which assessed communication skills, but to a student's entire performance. Indicative quotes relating to this sub-theme often emphasised the importance of being able to carry out all the required steps in an OSCE while also maintaining an appropriate demeanour and engaging with the patient:

> I suppose that their demeanour and you know that they're confident, their steps, speaking to the patient and talking to the patient… not talking at it. Or you know, "next, and this next", they're checking on the patient that they're okay, they are caring, compassionate. If there is a human being that they're actually dealing with instead of being just focused on the next step, this step, and you can tell if they've practiced more that they're able to integrate all that.
>
> Assessor 4

In this quote, the assessor discussed how she would be unhappy with a student who was overly focused on completing each step in the OSCE, without taking the patient into consideration. Later on in the interview, the same assessor criticised a student in one of the videos (P03BP) for not engaging with the patient:

> Her communication then it's quite... There's no rapport building with them. But they come in and just do it. They don't say, "how are you? Are you okay? Can I take this?" And it's not reassuring to the patient. Her body demeanour shows that she's comfortable, capable of being able to do the skill and follow them in relation to the criteria. So she can do the skill okay.
>
> Assessor 4

Here, the assessor was critical about a student's performance in the blood pressure OSCE, even though she acknowledged that the student was capable of completing the required steps in order to take a blood pressure measurement. It is notable that this assessor participant still awarded the mark for the item *explains the procedure to the patient and gains consent*, even though she was critical about how the student accomplished this. As such, the student's ability to maintain communication with the patient while completing the required steps affected the assessor's overall impression of the student’s performance, rather than any specific item on the marking guide.

Assessor participants’ impressions about students were also affected by whether they perceived the student to have demonstrated safe practice, even though “safety” was not an explicit criterion within the marking guide for either OSCE. Six of the assessors invoked the idea of safety as being a key indicator for determining how well a student had performed:

> The main thing is that they're safe. They're demonstrating safe practice. If they're not demonstrating safe practice, I just won't pass them… And that's just bottom line, like, you might have a student that comes in that's all chatty and great with the talking to the mannequin, but they might not demonstrate safe practice, where you might have the student that comes in and doesn't open his or her mouth. But they have demonstrated safe practice.
>
> Assessor 11

This assessor was emphatic about the fact that the concept of safety is a crucial means by which she makes a pass/fail decision about a student. For her, if the student is unsafe, they will not receive a passing mark, even if they demonstrate other skills (for example, effective communication with the patient). Comments relating to the idea of safety generally invoked it as being the ultimate arbiter of a student’s competence – if they were not safe, they were not competent, regardless of anything else.

The importance of the concepts of communication and safety in affecting assessors’ judgements has been noted in the literature on OSCE assessment (e.g., Rushforth, 2007). In a study of nursing assessors, East et al. (2014) found that safety was the single biggest criterion affecting assessors’ judgements of students. This is, in a sense, unsurprising, given that safety is a core principle of effective nursing practice. However, from an assessment perspective, the invocation of safety as a judgement criterion is problematic, given that neither of the OSCEs used in the study contained safety as one of the assessment criteria. This means that, in spite of assessors’ earlier comments that they used the marking guide to assess student performances, their judgements are additionally guided by their own conceptions of safety. This suggests that, in practice, student OSCE scores are not only determined by the criteria contained within the marking guides. Additionally, assessors’ own ideas of what constitutes “safe practice” may be inconsistent: what strikes one assessor as unsafe may not be the same to another. As noted by East et al. (2014, p.466): “participants’ perceptions of safety differed substantially according to their subjective experiences". This has implications for score reliability, as a student's score may be affected by whether their performance in an OSCE aligns with what a specific assessor happens to consider safe.

East et al. (2014) also noted the relevance of communication in informing assessors' judgements, discussing how assessors frequently mentioned a student's communication skill as being a core aspect of their performance (even when it was not explicitly mentioned in the marking guide). All in all, the findings from the present study show clear parallels with those of East et al., in terms of assessors using the potentially subjective criteria of safety and communication to inform their judgements, in a way that goes beyond what is contained within the marking guides. These two important criteria are re-visited in section 4.6.

4.4.4.2 *Translating checklist scores to global scores*

The second sub-theme relates to the relationship between the checklist items on the marking guide (of which there were 12 on the blood pressure OSCE and 14 on the NG tube) and the global questions asked about the student at the end (e.g., "How would you rate this student's performance overall?"). In neither of the two OSCEs used in the present study (which can be found in Appendix H) were any of the items afforded a "weighting" based on their perceived importance. In seven of the 12 interviews, assessors reported that a specific item on the marking guide was so important that failure to complete it would have a significant and negative impact in their overall assessment of a student's performance. These "red line issues" were specific to different procedures:

> Adhering to basic infection control principles, that's one of the things… that could compromise the patient's safety, they need to understand that this is just not... This is something that cannot be done. And the rationale behind it. So they're asking, for example, you're taking out sutures, basic wound care, and you have a basic principle of how you maintain the sterile field. There is a kind of cut-off point. If they contaminate the field, that's it…Whereas something like inserting the gastric tube, I think if you didn't communicate to the patient that there's a stop sign. That's a critical part… That communication - that the patient has control and is able to indicate. So that's something that's really important training for. So that would be something to say, okay, they're borderline, but it's critical they achieve certain points there.
>
> Assessor 6

Four of the assessor participants noted that when OSCEs were administered in the past, it was common for some items to be given this “red line” status, such that if a student did not complete them, they would fail the OSCE automatically, even if they had done every other step correctly. These participants believed that this process should be brought back into the assessment.

Finally, assessor participants also discussed the difficulty of awarding a final mark to a student who had completed many of the checklist items, but had not demonstrated an overall level of competence. Five of the assessors expressed reservations about the potential incongruence between a high score on the checklist and a low mark for overall performance:

> The thing is, she has done some of the steps, but my overall impression would be to fail her. But what she has done here, I don't know what the marks are, but she has 1, 2, 3, 4, 5, 6 ticks. So she has done some parts correctly. But I know that she hasn't got the right pressure.
>
> Assessor 11

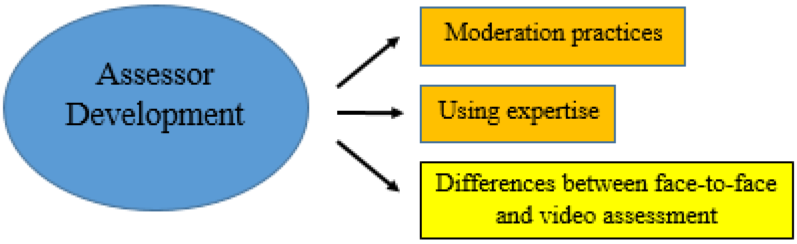
In this extract (related to video P03BP), the assessor discussed how the student had completed six of the 12 steps in the checklist, but that she did not want to award the student a passing grade because she did not believe the student had taken an accurate blood pressure reading. As such, she expressed confusion as to how to reconcile her overall impression about the student’s performance with the fact that the student had completed many of the steps on the guide. This suggests a difficulty with the way that she processed the global question relating to overall performance, being unable to confidently award a failing grade even though that is what she believed was warranted.

Assessors’ comments about the marking guides used in the present study are notable from an assessment perspective. Over half the assessors in the present study discussed the idea that specific “red line” items were so important that a student should not be able to pass the OSCE if they fail to complete them (which was not the case in this study). This suggests a misalignment between the design of the marking guides and what assessors actually believe to be important, a finding that has been widely noted in the literature on performance assessments in medicine (e.g., Tavares & Eva, 2013; Hyde et al., 2020). Given the documented tendency of assessors to explicitly act outside the provisions of the marking guide if they don’t believe it allows them to capture their opinions of a student (as documented here and also in a recent study by Hyde et al. (2022)), it is imperative that marking guides are designed in consultation with assessors. Doing so should ensure that assessors stick to the marking guide when judging students, which will increase the reliability of awarded scores. These findings also correlate with those in Yeates et al.'s (2013) study, which documented how assessors in medicine have difficulty weighing all the information available to them about a student in order to come to a final decision about that student. Assessors in the present study expressed difficulty deciding on an overall global score for a student (e.g., Fail/Borderline pass/Good/Excellent), particularly in relation to the number of checklist items they believed a student to have completed successfully. This may affect the reliability of awarded scores, as two assessors may have differing opinions about whether they could fail a student who completed half of the checklist items successfully.

4.4.5 *Assessor development*

The final theme relates to factors which affected assessors' decision-making, but which did not pertain to specific elements within Gauthier et al.'s (2016) framework. These factors influenced assessors' judgements in the long-term (or over the course of their careers as assessors), and as such do not pertain to the act of watching a specific video or judging a single performance. None of the indicative quotes within this theme relate to the four videos used in the study, instead coming from assessors reflecting about their general assessment practices. Within this theme, the sub-themes of *moderation practices* and *using expertise* were identified. Additionally, a latent sub-theme that emerged from the data was that of *differences between video and face-to-face assessment*, which also affected how assessors formed judgements about students. These three sub-themes are discussed in turn below.

Figure 4.6 *Assessor development*



4.4.5.1 *Moderation practices*

This sub-theme relates to the means by which assessors discuss OSCEs (and related performance assessments) with their colleagues, and use these conversations to inform their judgements of subsequent student performances. These processes can be formal (e.g., double-marking of a certain percentage of performances) or informal (e.g., chatting to colleagues); and the findings indicate inconsistencies in the extent and nature of moderation practices. In terms of formal practices, four assessor participants discussed referring specific videos back to the module coordinator if they were unsure as to how well or badly the student had performed. This was particularly important when the student may have been on the verge of failing the OSCE:

> I think being familiar with the grading system and I suppose those little pieces that are maybe vague where you're not sure if the student should pass or should fail, as I say. At the moment, I'm referring them back to the module coordinator. So, yeah, I suppose the details. I might be able to make my own decisions more as I go on.
>
> Assessor 1

In this extract, the assessor (who had only assessed a single administration of OSCEs at the time of the interview) discussed the fact that she does not have the confidence to make a unilateral decision to fail a student based on their performance, and so generally refers such borderline students to the module coordinator in order to make a final decision. The aim of these referrals is to allow her to develop a better sense of what distinguishes a passing student from a failing student, so that she can make such decisions herself in the future.

Less formally, two assessor participants mentioned having discussions with their colleagues about their assessment practices, which affected how they assessed students going forward:

> I suppose one of the examples would have been mentioned earlier is like, as I became more confident, I talked to others, they'd say maybe "I gave her a few minutes and told her, come back and try again". And then I thought, oh, yeah, I can do that.
>
> Assessor 10

Here, the assessor noted that, in the past, she had spoken to colleagues who were assessing the same OSCE, who mentioned that they were happy to pause the OSCE if the student was performing poorly, and give the student a chance to begin again. After this conversation, the assessor began incorporating this into her own assessments.

Assessor participants’ accounts of institutional moderation processes were inconsistent. One assessor noted that in the OSCE he assesses, there is a strict procedure of internal moderation, to ensure that assessor idiosyncrasy is minimised:

> Those ones are usually, we do internal moderation here anyway, so there's always two or three people looking at stuff. So it's just to make sure that we are all thinking the one thing here. And then you have a look at their feedback sheets to see were the same points being highlighted by three different people. And to be honest, it's usually unanimous.
>
> Assessor 8

However, another assessor was explicit about the fact that, in the OSCEs she judges, no such procedure is in place:

> At the moment... we don't moderate OSCEs or those kind of assessments. We have to try to think about what's feasible to do, but also what's the right thing to do.
>
> Assessor 10

These divergent responses indicate a lack of strict moderation policies at the School level. As such, some module coordinators may choose to implement formal processes of internal moderation, while some do not.

The importance of moderation procedures in improving score reliability has long been noted in the literature on OSCEs, as well as assessments more generally. Ideally, moderation should happen both before and after an OSCE is administered: before the OSCE, assessors should meet and try to come to a consensus on the standards required in order to pass each item on the guide, as well as what is expected from students at different levels in terms of their global score. After the OSCE, some percentage of the performances should be checked by more than one assessor, to ensure adequate levels of IRR (Khan et al., 2013b). The findings of the present study are concerning in this regard, given the discrepancies in moderation practices reported in the School. It is likely that these discrepancies have reduced agreement between assessors as to what constitutes good performance, as they seem not have had an opportunity to discuss this in a formal setting. Indeed, the admission from Assessor 10 that she began allowing nervous students to start the OSCE again after a colleague informed her that she did this as well has obvious implications for score reliability: a nervous student may receive a significantly higher score from this assessor than they would from a different assessor who does not allow such a restart.

4.4.5.2 *Using expertise*

The second sub-theme relates to expertise on the part of assessors, and how their process of judging students evolved as they gained more experience. All seven of the assessor participants who had been assessing OSCEs for more than ten years at the time of interview reported developing a keen internal sense of whether a student was performing well or not. These assessors mentioned that, when assessing a student, they usually “know straight away” (Assessor 11) how well the student is doing, and that they rarely have to revise their initial impression. Indicative quotes relating to this sub-theme invoked the idea of a “gut instinct” when judging students:

> And that's what I say, that sometimes it's not tangible. And that's what I mean by going with the gut. That as a lecturer, you have to go with your gut with this: "no, I'm not happy with this lady".
>
> Assessor 7

These experienced assessors also reported feeling confident in the grades they award to students, rarely second-guessing themselves. One also reported that other assessors would ask her to assess videos that they were having trouble with, and that her word would always be taken as final:

> And I'm sure there's a lot more justification for my decisions on my grades… Rarely someone will say to me, “no, I don't think that's right.” They always kind of send them into me [if they are unsure] and say, “well, you examine them now and see.” And they will take my word, if I say yeah they're competent or not. There's never any question.
>
> Assessor 11

As such, these assessors reported that they have developed an innate ability to determine quickly if a student is performing well or not, and that their decisions are respected by their colleagues. This above quote also has to do with moderation, as it indicates that this assessor seems to be frequently used as means by which to check that another assessor’s opinion about a student is the correct one.

Findings within this sub-theme suggest that experienced assessors discuss their own assessment practices in a way notably similar to Benner’s (1982) conception of expertise. Benner noted that expert nurses can focus on the salient aspects of a given situation and come to a quick decision about what to do. A similar idea was noted here: some assessors discussed how they quickly form an instinctive decision about a student, from which they rarely deviate. However, the issue of assessor expertise is one that is subject to debate within the assessor cognition literature. Some authors (e.g., Hodges, 2013; Govaerts & van der Vleuten, 2013) have argued that experienced assessors should be allowed some level of subjectivity in their judgement practices, as they possess specialised knowledge that less experienced assessors do not, and can use this in order to make informed judgements. Seen this way, subjectivity on the part of assessors increases the validity of assessment decisions, as it results in competent students being identified and rewarded. In contrast, others (e.g., Gingerich et al., 2014a) have written that this subjectivity threatens the reliability of awarded scores, as it reduces the consistency between assessors in terms of how they judge students. This tension between subjectivity and reliability was discussed in section 2.4 of the literature review and will be unpacked further in the following chapter. For now, it is noted that, on the basis of these findings, it is apparent that experienced nursing assessors reported leveraging their expertise in order to make quick judgements about students, that they rarely perceived as being incorrect.

4.4.5.3 *Differences between video and face-to-face assessment*

This sub-theme describes assessor participants' opinions about the perceived differences in assessing students through video, rather than in-person. These discussions were focused on how the use of video affected their ability to make objective judgements about students. All assessors noted that video technology had been incorporated in the School since the COVID-19 pandemic in 2020. Ten of the 12 assessors were entirely positive about the benefits that video affords. These assessors noted that the use of video allowed them to make decisions with more confidence, and therefore rely less on instinct or guessing, which they admitted had happened occasionally when OSCEs were administered and assessed live:

> This just reminded me of when I'm doing it, that when you're doing so many of them in real life, sometimes you're going "Did she do that?" and that's where the video is helpful because you can actually stop it and look back and you can't do that in real life.
>
> Assessor 10

> I would probably watch the whole thing through and then I'd go back. I might highlight something that might pop up, that I need to look a little bit more carefully. And then I

> would go back and I might stop it a few times if I'm not sure, especially with the first few until I become familiar with exactly what the student is doing.
>
> Assessor 1

For these assessors, the use of video technology within the OSCE allows them to view the same performance multiple times, to ensure they have not missed anything and can confidently award a grade to the student. As such, the use of technology affected their process of coming to a judgement about a student. Seen this way, using video technology should improve score reliability, as assessors would be less reliant on guesswork or short-term memory when judging a student.

However, it is also notable that two of the assessors spoke about the use of video in more negative terms. Both assessors emphasised that when OSCEs are administered through video, students can record their attempt at the skill multiple times; as such, the disconnect between the "real world" of clinical practice, where a nurse may only have one attempt at performing a procedure, and the testing environment, is increased. For these assessor participants, the use of video may decrease the objectivity of the OSCE:

> But of course, if it's physical assessment, you will probably have gotten more information… And I would have been able to assess more objectively because it's real time. And whatever mistakes they're making, they're making it in real time... Whatever thing they're doing, it's real time. And it gives you more objectivity for sure.
>
> Assessor 2

Given that one of the perceived advantages of the OSCE is its close approximation to the real world of clinical practice, the use of video may threaten the validity of the assessment, as it decreases this proximity. All in all, the use of video technology as a mediating factor on participants' decision-making was discussed in mostly, but not entirely, positive terms. As such, the results of the present study add to the nascent literature concerning the use of video in performance assessments such as the OSCE. As noted by Major et al. (2020) and others, the COVID-19 pandemic resulted in the rapid incorporation of video into performance assessments, as it was the only way that these assessments could take place. Ten of the participants in this study were positive about this development, noting that it had the potential to increase score reliability, due to the reduction in guesswork or errors on the part of assessors as to whether a student had completed a task successfully. Put simply: the ability to pause and rewind the video allowed assessors to see aspects of a student's performance more clearly, and make a more confident decision as a result. However, two assessors spoke about video in more negative terms. Their comments have to do with validity, specifically the fact that using video increases the disconnect between the OSCE and the real world of practice. This potential disconnect has long been the subject of debate in OSCE literature (Hodges, 2003; Khan, 2017). The present study suggests that OSCE designers and administrators should think about what is "lost" in terms of validity when the assessment is moved online, in addition to what may be gained in terms of reliability.

4.4.6 *Summary of thematic analysis*

This section addressed the first research question, *What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?* The thematic analysis resulted in four identified themes: *observation, processing, integration* and *assessor development*. The first three of these themes are derived from Gauthier et al.'s (2016) framework for assessors' judgement formation. In the *observation* stage, assessors used the marking guide to direct their attention to the relevant aspects of a student's performance, however they were also explicit about their tendency to focus on certain aspects of a performance that they believed to be important. Assessors were also prone to making inferences about the students that went beyond what was directly visible in the videos. Finally, assessors' observations were in some instances hindered by the limitation of the video format in allowing them to see all the salient features of the recorded performances.

In the *processing* stage, assessors "made sense of" a student's performance by comparing it to a range of other performances, for example that of the student who performed the skill before them. Several assessors also discussed mitigating factors that may cause them to be more generous in their judgements of students. Finally, three assessor participants discussed the idea that their interpretations of the recorded performances were affected by the knowledge that there was more than one correct way to complete a specific task or procedure.

In the *integration* stage, assessors were affected by how they perceived a student to have performed overall, usually informed by those aspects related to communication and safety. Some assessors also expressed confusion regarding the marking guide, particularly in terms of translating their checklist scores to global scores, and the importance of some "red line" items within each marking guide.

The final theme, *assessor development*, described how assessors' decision-making processes are influenced by a range of long-term factors; namely, moderation procedures (whether formal or informal) and expertise. The findings also indicate divergent opinions regarding how the use of video technology affects assessors' abilities to make judgements about students.

The relevance of these findings in relation to published work on assessor cognition were discussed throughout the section, as were the potential implications in terms of score validity and reliability. The question of what assessment designers and administrators can do to mitigate these potential implications is discussed in the Chapter Five. The following section addresses the second research question, detailing the results from the quantitative element of the study.

4.5 *RQ2: What level of inter-rater reliability is evident in undergraduate nursing OSCEs?*

4.5.1 *Introduction*

This section addresses the second research question by examining the data from the quantitative element of the study; namely, the results of the completed marking guides. Given the extent to which assessor participants in the present study employed a range of potentially subjective judgement criteria when grading the four videos (as documented in the previous section), it is logical to assume that the IRR of these videos would be negatively affected. This section documents the extent to which this was the case. Firstly, there is a reminder of the quantitative data that were collected as part of the study. Secondly, quantitative results are presented for each of the four videos, to demonstrate how each performance was graded by the 12 assessors. Thirdly, results from this phase of the study are presented on a per-assessor basis, allowing for an identification of which assessors were consistently harsher or more lenient than average. This analysis is used as the basis for the third and final phase of analysis which follows in section 4.6. Subsequently, there is a calculation of inter-rater reliability (IRR) for each item across the two OSCEs, to allow for a determination of which items were more likely to lead to divergence between assessors. The section ends with a discussion of the research question and the key issues that arise as a result of these data.

4.5.2 *Reminder of quantitative data collected*

As part of this study, 12 assessor participants each graded four videos of students completing OSCEs: two videos of a blood pressure measurement (BP) OSCE, labelled as P02BP and P03BP; and two videos of a naso-gastric tube insertion (NG) OSCE, labelled as P01NG and P03NG. The number after the P refers to the student who completed the OSCE: P01 is a first-year student, P02 is a second-year student, and P03 is a third-year student. For both OSCEs, assessors were given a series of *binary checklist items*, relating to different steps required to complete the skill. For each video, assessors were asked to tick which items they perceived the student in the video to have completed satisfactorily. The BP OSCE contained 12 such items, while the NG OSCE contained 14. A copy of the marking guides, containing these items, can be found in Appendix H. Additionally, for all four videos, assessors were asked to answer the same three *global questions*, relating to the student's overall performance:

1. How would you rate the participants' **communication** skills? Fail / Borderline pass / Good / Excellent
2. How confident would you be allowing this student to perform this procedure **on you**? Not at all confident / Just about confident / Confident / Very Confident
3. How would you rate this performance **overall**? Fail / Borderline pass / Good / Excellent

For the first and third questions, answers were coded as follows: fail = 0, borderline pass = 1, good = 2, excellent = 3. For the second question, responses were coded as follows: not at all confident = 0, just about confident = 1, confident = 2, very confident = 3.

As such, for each of the four videos, assessors completed either 12 or 14 binary items, and three global items. In total, each of the 12 assessors answered a total of 52 binary items and 12 global items.

4.5.3 *Quantitative data per video*

This section details the results of each video in turn, to demonstrate how each recorded performance was graded by the 12 assessor participants, and the extent to which there was divergence in how the same performance was graded by the assessors. For each video, there are two tables presented: the first details the results of the checklist items, while the second details the results of the global items. As described above, the checklist items were binary in nature: during the interviews, assessors ticked each item they perceived the student in the video to have completed successfully. Table 4.2 below demonstrates the checklist responses for the video P01NG. Each "tick" given by assessor participants is represented by the number 1, and the total number of ticks awarded is given in the far-right column. For example, the first assessor who watched the video P01NG ticked all the items on the checklist except items 4, 11 and 14, for a total of 11.

The checklist item tables also include a calculation of IRR for each item on the marking guide, calculated using the percent agreement statistic, found in the bottom column. This was calculated by determining what proportion of participants ticked that item on the guide while watching a specific video, and subtracting that number from 1 if it was below 0.5. As such, the possible results for this calculation ranged from 0.5 (least possible agreement) to 1 (all participants agreed). For example, for video P01NG, eight out of 12 participants awarded the mark for item 1 (*wears correct full uniform – hair tied up/off shoulders*), giving a percent agreement calculation of 0.67; while for item 4 (*pre-arrange a 'stop' signal patient can use*), two out of 12 participants awarded the mark, for a percent agreement of 0.83. As noted by Gwet (2014), percent agreement figures of less than 0.75 are classified as unacceptable.

The second table details the results for the *global* items. As described above, these items related to a student's overall performance, rather than specific steps on the marking guide. Possible results for each of these items ranged from 0 to 3, with 0 being a fail. Table 4.3 details the responses to these three global items for the video P01NG. Each table is colour-coded such that "Excellents" are green, "Goods" are yellow, "Borderline passes" are orange and "Fails" are red. For example, Assessor 1 awarded an "Excellent" for questions 1 and 3, and a "Very confident" for question 2; while Assessor 2 awarded a "Borderline pass" for question 1, a "Confident" for question 2, and an "Excellent" for question 3. The mean and standard deviation for each of these questions is found at the bottom of the table.

4.5.3.1 *Results for P01NG*

Table 4.2 *Results from checklist items for video P01NG*

Item:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Sum
Assessor 1	1	1	1		1	1	1	1	1	1		1	1		11/14
Assessor 2	1	1	1		1	1	1	1	1	1	1	1	1	1	13/14
Assessor 3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14/14
Assessor 4		1	1	1	1		1	1	1	1	1	1	1		11/14
Assessor 5		1	1		1	1	1	1	1	1	1	1	1	1	12/14
Assessor 6		1	1		1	1	1	1	1	1	1	1	1		11/14
Assessor 7	1	1	1			1	1		1	1	1	1	1	1	11/14
Assessor 8	1	1	1		1	1	1	1	1	1	1	1	1	1	13/14
Assessor 9	1	1	1			1	1	1	1	1	1	1	1	1	12/14
Assessor 10	1	1	1		1		1	1	1	1	1	1	1		11/14
Assessor 11		1	1		1		1		1		1	1	1		8/14
Assessor 12	1	1			1	1	1	1	1	1		1	1	1	11/14
Percent Agreement	**0.67**	**1**	**0.92**	**0.83**	**0.83**	**0.75**	**1**	**0.83**	**1**	**0.92**	**0.83**	**1**	**1**	**0.58**	

For video P01NG, the number of checklist items awarded ranged from 8 (Assessor 11) to 14 (Assessor 3). The mean number of marks awarded was 11.5 with a standard deviation of 1.51. In terms of IRR, five items (2, 7, 9, 12 and 13) reported complete agreement between the 12 assessors, while three items (1, 6 and 14) reported a percent agreement of 0.75 or below.

Table 4.3 *Results from global items for video P01NG*

Item:	Communication	Self as Patient	Overall
Assessor 1	3	3	3
Assessor 2	1	2	3
Assessor 3	3	2	3
Assessor 4	*0*	1	1
Assessor 5	2	2	2
Assessor 6	2	1	2
Assessor 7	2	3	1
Assessor 8	3	3	3
Assessor 9	3	1	2
Assessor 10	1	2	2
Assessor 11	2	1	2
Assessor 12	2	2	2
Mean	**2**	**1.92**	**2.17**
SD	**0.95**	**0.79**	**0.71**

Note: 3 = excellent, 2 = good, 1 = borderline, 0 = fail

For video P01NG, there was variation between assessors for all three global questions. This was most pronounced for the communication question, for which assessor participant responses ranged from Fail (Assessor 4) to Excellent (Assessors 1, 3, 8 and 9). This question also reported the highest standard deviation of 0.95.

4.5.3.2 *Results for P02BP*

Table 4.4 *Results from checklist items for video P02BP*

Item:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Sum**
Assessor 1	1	1	1	1	1	1	1	1	1	1	1	1	12/12
Assessor 2	1	1	1	1	1	1	1	1	1	1	1	1	12/12
Assessor 3	1	1	1	1	1	1	1	1	1	1	1	1	12/12
Assessor 4	1		1	1	1	1	1		1	1		1	9/12
Assessor 5		1	1	1	1	1	1	1	1	1	1	1	11/12
Assessor 6	1	1		1	1	1	1	1	1	1	1	1	11/12
Assessor 7	1	1	1	1	1	1	1		1	1	1	1	11/12
Assessor 8	1	1		1	1	1	1						6/12
Assessor 9	1	1	1	1	1	1	1			1	1	1	10/12
Assessor 10		1		1	1	1	1	1	1	1	1	1	10/12
Assessor 11			1	1	1	1			1	1		1	7/12
Assessor 12	1	1	1	1	1	1	1	1	1	1	1	1	12/12
Percent Agreement	**0.75**	**0.83**	**0.75**	**1**	**1**	**1**	**0.92**	**0.58**	**0.83**	**0.92**	**0.75**	**0.91**	

For video P02BP, the number of checklist items awarded by assessors ranged from 6 (Assessor 8) to 12 (Assessors 1, 2, 3 and 12). The mean number of marks awarded was 10.25 with a standard deviation of 2. In terms of IRR at the item level, three items (4, 5 and 6) resulted in complete agreement between assessors, while four (1, 3, 8 and 11) resulted in percent agreement of 0.75 or below.

Table 4.5 *Results from global items for video P02BP*

Item:	**Communication**	**Self as Patient**	**Overall**
Assessor 1	2	3	3
Assessor 2	3	3	3
Assessor 3	3	3	3
Assessor 4	1	1	2
Assessor 5	3	2	2
Assessor 6	2	2	2
Assessor 7	2	3	1
Assessor 8	2	*0*	*0*
Assessor 9	2	*0*	1
Assessor 10	2	2	2
Assessor 11	2	*0*	*0*
Assessor 12	2	2	2
Mean	**2.17**	**1.75**	**1.75**
SD	**0.58**	**1.22**	**1.06**

Note: 3 = excellent, 2 = good, 1 = borderline, 0 = fail

For video P02BP, there was notable variation between assessors regarding the global items. This was most pronounced for questions 2 and 3, each of which resulted in the full range of grades being awarded, from Fail/Not at all confident to Excellent/Very confident. These two questions also reported standard deviations of 1.22 and 1.06, which were the highest and fourth-highest found across the four videos. Notably, there was a discrepancy regarding the assessors' decisions about the student's overall performance at the pass/fail decision: two of the 12 assessors deemed that she had failed the OSCE, while the other 10 determined that she passed.

4.5.3.3 *Results for P03BP*

Table 4.6 *Results from checklist items for video P03BP*

Item:	1	2	3	4	5	6	7	8	9	10	11	12	Sum
Assessor 1			1	1	1	1	1	1				1	7/12
Assessor 2	1		1	1	1	1	1	1				1	8/12
Assessor 3	1		1	1	1	1		1			1	1	8/12
Assessor 4	1		1	1	1		1	1	1	1	1	1	10/12
Assessor 5			1	1	1	1	1	1	1	1	1	1	10/12
Assessor 6	1	1	1	1	1	1	1	1			1	1	10/12
Assessor 7	1		1	1	1	1	1	1	1	1	1	1	11/12
Assessor 8	1	1	1	1	1	1	1	1	1	1	1	1	12/12
Assessor 9	1		1	1	1	1	1	1			1	1	9/12
Assessor 10	1		1	1	1	1	1	1			1	1	9/12
Assessor 11	1		1	1		1		1				1	6/12
Assessor 12			1	1	1	1	1	1				1	7/12
Percent Agreement	**0.83**	**0.75**	**1**	**1**	**0.92**	**0.92**	**0.83**	**1**	**0.58**	**0.58**	**0.75**	**1**	

For video P03BP, the number of checklist items ticked by assessors ranged from 6 (Assessor 11) to 12 (Assessor 8). The mean number of items awarded was 9.33 with a standard deviation of 1.87. In terms of IRR, four items (3, 4, 8 and 12) resulted in perfect agreement between assessors, while four (2, 9, 10, and 11) resulted in a percent agreement statistic of 0.75 or less.

Table 4.7 *Results from global items for video P03BP*

Item:	Communication	Self as Patient	Overall
Assessor 1	1	1	2
Assessor 2	2	1	1
Assessor 3	3	2	2
Assessor 4	2	3	2
Assessor 5	2	2	2
Assessor 6	2	3	2
Assessor 7	2	2	2
Assessor 8	2	2	2
Assessor 9	2	1	2
Assessor 10	3	3	2
Assessor 11	2	*0*	*0*
Assessor 12	1	2	2
Mean	**2**	**1.83**	**1.75**
SD	**0.6**	**0.94**	**0.62**

Note: 3 = excellent, 2 = good, 1 = borderline, 0 = fail

For video P03BP, there was variation between assessors for the global questions, although this was less pronounced than in the other videos, with relatively low standard deviations reported, particularly for questions 1 and 3. However, in spite of this, there was still a discrepancy regarding the overall pass/fail decision, with one assessor (Assessor 11) awarding the student a Fail. As such, the student would have passed the OSCE with 11 assessors, but failed with one.

4.5.3.4 *Results for P03NG*

Table 4.8 *Results from checklist items for video P03NG*

Item:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**Sum**
Assessor 1	1	1	1		1	1	1	1	1	1		1		1	11/14
Assessor 2	1	1	1		1	1	1	1	1	1		1			10/14
Assessor 3	1	1	1		1	1	1	1	1	1	1	1			11/14
Assessor 4		1	1	1	1							1			5/14
Assessor 5	1	1	1		1	1	1	1	1	1	1	1			11/14
Assessor 6	1	1	1		1	1			1	1		1			8/14
Assessor 7	1	1			1	1		1	1	1	1	1		1	10/14
Assessor 8	1	1													2/14
Assessor 9	1	1	1		1	1	1	1	1			1		1	10/14
Assessor 10	1	1	1		1	1	1	1	1	1	1	1			11/14
Assessor 11	1	1	1		1		1	1	1			1			8/14
Assessor 12	1	1	1		1	1		1		1	1	1			9/14
Percent Agreement	**0. 92**	**1**	**0.83**	**0.83**	**0.92**	**0.75**	**0.58**	**0.75**	**0.75**	**0.67**	**0.58**	**0.92**	**1**	**0.75**	

Finally, for video P03NG, the number of items ticked by participants ranged from 2 (Assessor 8) to 11 (Assessors 1, 3, 5 and 10). The mean number of items ticked was 8.83 with a standard deviation of 2.79. In terms of IRR, two items (2 and 13) reported perfect agreement between assessors, while seven (6, 7, 8, 9, 10, 11, 14) reported a percent agreement of 0.75 or below.

Table 4.9 *Results from global items for video P03NG*

Item:	Communication	Self as Patient	Overall
Assessor 1	3	3	3
Assessor 2	1	1	1
Assessor 3	3	3	3
Assessor 4	*0*	*0*	*0*
Assessor 5	3	2	2
Assessor 6	1	2	2
Assessor 7	2	2	*0*
Assessor 8	1	*0*	*0*
Assessor 9	3	2	2
Assessor 10	1	2	2
Assessor 11	3	2	2
Assessor 12	1	2	2
Mean	**1.83**	**1.75**	**1.58**
SD	**1.11**	**0.97**	**1.08**

Note: 3 = excellent, 2 = good, 1 = borderline, 0 = fail

For video P03NG, there was considerable variation between assessors, with all four grades being awarded for each of the three questions. This resulted in relatively high standard deviations, with questions 1 and 3 reporting the second- and third-highest figures across all four videos. Notably, in terms of the pass/fail decision, three assessors awarded the student a "Fail" in terms of her overall performance.

Taken together, the results of the checklist items across the four videos indicate notable variation in how each recorded performance was graded by the participants. This variation was most pronounced for video P03NG, which had the highest reported standard deviation in terms of items ticked (2.79), as well as the highest number of items resulting in a percent agreement of less than 0.75. Even the video with the lowest standard deviation (P01NG, 1.51) resulted in a situation where the most lenient assessor perceived the student to have completed six more steps in a satisfactory way than the harshest assessor. This has implications in terms of validity, as making a decision on the basis of an assessment result in which there was such range between assessors may not be considered defensible (AERA et al., 2014).

Similarly, the results of the three global items revealed considerable variation in how assessors graded the same four recorded videos. As with the results from the checklist items, these results were occasionally stark: for video P03NG, two assessors awarded the student an "Excellent" in terms of her overall performance, while three awarded her a “Fail”. Indeed, in terms of the overall pass/fail decision:

- The student in P01NG passed regardless of the assessor.
- The student in P02BP passed with ten of the assessors.
- The student in P03BP passed with 11 of the assessors.
- The student in P03NG passed with nine of the assessors.

As such, there were notable discrepancies in how the 12 assessors graded the four recorded student performances, both in terms of the checklist items and the global items at the end of each OSCE. These discrepancies have implications for reliability and validity: if the same student performance (P03NG) receives anywhere between two and 11 marks (out of 14) for the checklist items and “overall” grades ranging from Fail to Excellent, it is clearly less defensible to make a decision on the basis of a single assessment score. The implications of these results are unpacked further in section 4.5.6.

4.5.4 *Quantitative data per participant*

This section details the results of the checklist items and global questions on a per-assessor basis across all four videos. This allows for an identification of which assessors were consistently harsher or more lenient in comparison to the rest of the sample. Identifying outlier assessors is an important step when conducting reliability analyses (Khan et al., 2013b), and will further be used as the basis for the final phase of analysis conducted in section 4.6.

Table 4.10 details the *checklist* item responses for each assessor across all four videos. For example, Assessor 1 ticked 11 items (out of 14) for video P01NG, 12 (out of 12) for P02BP, 7 (out of 12) for P03BP and 11 (out of 14) for P03NG. As such, Assessor 1 ticked 41 out of a possible 52 steps, or 79% of all possible steps. By this measure, Assessor 3 was the assessor who awarded the greatest number of checklist items, ticking 45 out of 52 possible steps (86.5%), followed by Assessors 5 and 12, who each ticked 44. These participants are highlighted in green. Conversely, Assessor 11 was the harshest, ticking 29 out of 52 steps (55.8%), followed by Assessor 8 with 33 (63.4%). These assessors are highlighted in red.

Table 4.10 *Checklist item responses for each participant*

Video:	P01NG (out of 14)	P02BP (out of 12)	P03BP (out of 12)	P03NG (out of 14)	Total steps (out of 52)	Percent possible steps
Participant 1	11	12	7	11	41	79%
Participant 2	13	12	8	10	43	82.7%
Participant 3	14	12	8	11	45	86.5%
Participant 4	11	9	10	5	35	67.3%
Participant 5	12	11	10	11	44	84.6%
Participant 6	11	11	10	8	40	76.9%
Participant 7	11	11	11	10	43	83.7%
Participant 8	13	6	12	2	33	63.4%
Participant 9	12	10	9	10	41	78.9%
Participant 10	11	10	9	11	41	78.9%
Participant 11	8	7	6	8	29	55.8%
Participant 12	11	12	12	9	44	84.6%
Mean	**11.5**	**10.25**	**9.33**	**8.83**	**39.91**	**76.85%**
SD	**1.5**	**2.01**	**1.87**	**2.8**	**4.98**	

Table 4.11 details the summated responses for the three global questions. As noted above, these responses were coded on a scale from 0-3. As such, for each assessor participant, there was a total of 12 possible marks awarded for each global item (a score of 12 indicates that the assessor gave an “Excellent” to all four videos) and a total potential mark of 36. By this measure, Assessors 1 and 3 were the most generous assessors, while Assessors 4 and 11 were the harshest.

Table 4.11 *Global item responses for each participant (per question)*

Question:	Communication	Self as Patient	Overall	Total marks awarded	Percent possible
Participant 1	9	10	11	30	83.30%
Participant 2	7	7	8	22	61.10%
Participant 3	12	10	11	33	91.20%
Participant 4	3	5	5	13	36.10%
Participant 5	10	8	8	26	72.20%
Participant 6	7	8	8	23	63.90%
Participant 7	8	10	4	22	61.10%
Participant 8	8	5	5	18	50%
Participant 9	10	4	7	21	61.10%
Participant 10	7	9	8	24	66.70%
Participant 11	9	3	4	16	44.40%
Participant 12	6	8	8	22	61.10%
Mean	**8**	**7.25**	**7.25**	**22.5**	

These two tables allow for an identification of which assessors were harshest or most lenient compared to their colleagues, for both the checklist items and global questions. Researchers writing about OSCE design and development have long noted the importance of examining score data in order to identify whether any assessors consistently award scores that are notably distinct from those of their peers (Bartman et al., 2013). Such assessors may threaten the reliability of OSCE scores, and the validity of decisions made on the basis of these scores, as they increase the chance that an awarded score is influenced by the assessor, rather than the true skill level of the student. By the measures described above, Assessors 4, 8 and 11 could be described as "harsh", while Assessors 1, 3, 5 and 12 could be described as "lenient". However, merely identifying these outlier assessors does not allow for an in-depth exploration of what specifically caused them to award such high or low marks compared to the general sample. This is the focus of section 4.6, in which data from a selection of these assessors will be examined in greater detail, to address whether any tentative links can be drawn between the cognitive processes they went through while judging the recorded OSCE performances, and the scores they awarded.

4.5.5 *Inter-rater reliability per item*

The collected data also allow for a calculation of IRR on a per-item basis, to determine which items in the marking guide were more likely to lead to disagreement between assessors. Tables

4.12 and 4.13 detail the IRR statistics for each item in the two OSCEs. As documented in the tables above, possible values for this statistic range from 0.5 (least possible agreement) and 1 (complete agreement). A value of 1 means that all assessor participants agreed that the student performed completed the skill required by the item correctly (or all agreed that she did not). A score of 0.5 means that half of the assessors agreed that she had performed the skill correctly, while the other half did not. Within each table, values of 0.75 and below are highlighted in red, while values of 0.75 – 0.8 are highlighted in orange, following the classification by Gwet (2014).

Table 4.12 *Inter-rater reliability for blood pressure measurement OSCE*

Item	PA in P02BP	PA in P03BP	Overall PA
1. Follow infection control guidelines in relation to: ‘bare below the elbows’, jewellery, hair, nails	0.75	0.83	**0.79**
2. Follow infection control guidelines regarding hand hygiene. State you would perform this.	0.83	0.75	**0.79**
3. Explain the procedure to your patient and gain consent.	0.75	1	**0.88**
4. Patient to be seated comfortably in chair.	1	1	**1**
5. Restrictive clothing removed from patient’s upper arm before applying cuff. Arm / hand in correct position.	1	0.92	**0.96**
6. Arm correctly positioned at heart level and supported.	1	0.92	**0.96**
7. Cuff bladder correctly placed on arm.	0.92	0.83	**0.88**
8. Locate brachial artery.	0.58	1	**0.79**
9. Brachial check to establish an estimated systolic pressure. This estimated number will be the number prior to the addition of the ’20-30mmHg’.	0.83	0.58	**0.71**
10. Cuff inflated to 20-30 mmHg above estimated systolic pressure you stated previously.	0.92	0.58	**0.75**
11. Cuff deflated in a controlled manner.	0.75	0.75	**0.75**
12. Systolic & diastolic blood pressure reading stated to video.	0.91	1	**0.96**
Mean	**0.85**	**0.85**	**0.85**

Table 4.13 *Inter-rater reliability for naso-gastric tube insertion OSCE*

Item	PA in P01NG	PA in P03NG	Overall PA
1.Wears correct full uniform - hair tied up/off shoulders.	0.67	0.92	**0.8**
2. Clearly states that handwashing has been completed prior to gloves and preparation of equipment.	1	1	**1**
3. Patient to be in position before starting video, assume consent.	0.92	0.83	**0.88**
4. Pre-arrange a "stop" signal patient can use i.e. raise their hand (articulate on video).	0.83	0.83	**0.83**
5. Select distance mark on tube i.e. length to be passed: distance from nose to ear lobe, then ear lobe to lowermost section of breastbone. (Must be clearly visible on video).	0.83	0.92	**0.88**
6. Lubricate proximal tip of tube with KY jelly. Check nostrils patent.	0.75	0.75	**0.75**
7. Insert tube into nostril, sliding backwards and inwards along floor of nose to nasopharynx.	1	0.58	**0.79**
8. As tube passes nasopharynx patient can be asked to swallow/take sips of water.	0.83	0.75	**0.79**
9. Advance tube through pharynx until predetermined mark is reached (articulate on video).	1	0.75	**0.88**
10. Check tube position is accurate using aspiration method. (Flush tube with 20mls air to clear, aspirate 2 – 10mls of stomach contents, test with pH indicator strips. (Articulate on the video)	0.92	0.67	**0.8**
11. Marks NG tube at exit point from nostril with indelible marker (immediately above or below adhesive tape).	0.83	0.58	**0.71**
12. Secure tube in place with adhesive tape/dressing.	1	0.92	**0.96**
13. Attach spigot.	1	1	**1**
14. Clearly states patient comfort is offered.	0.58	0.75	**0.67**
Mean	**0.86**	**0.8**	**0.83**

In the blood pressure measurement OSCE, three items on the marking guide (Appendix H) resulted in a percent agreement of 0.75 or less across the two videos:

- Item 9: *Brachial check to establish an estimated systolic pressure. This estimated number will be the number prior to the addition of the '20-30mmHg'.*
- Item 10: *Cuff inflated to 20-30 mmHg above estimated systolic pressure you stated previously.*
- Item 11: *Cuff deflated in a controlled manner.*

A further three items resulted in a percent agreement figure of between 0.75 and 0.8:

- Item 1: *Follow infection control guidelines in relation to: 'bare below the elbows', jewellery, hair, nails.*
- Item 2: *Follow infection control guidelines regarding hand hygiene. State you would perform this.*
- Item 8: *Locate brachial artery.*

In the naso-gastric tube insertion OSCE, three items resulted in a percent agreement figure of 0.75 or less across the two videos:

- Item 6: *Lubricate proximal tip of tube with KY jelly. Check nostrils patent.*
- Item 11: *Marks NG tube at exit point from nostril with indelible marker (immediately above or below adhesive tape).*
- Item 14: *Clearly states patient comfort is offered.*

A further three items resulted in a percent agreement figure of between 0.75 and 0.8:

- Item 7: *Insert tube into nostril, sliding backwards and inwards along floor of nose to nasopharynx.*
- Item 8: *As tube passes nasopharynx patient can be asked to swallow/take sips of water (if permitted).*
- Item 10: *Check tube position is accurate using aspiration method. (Flush tube with 20mls air to clear, aspirate 2 – 10mls of stomach contents, test with pH indicator strips. (Articulate on the video).*

These tables indicate that there were six items across the two OSCEs which resulted in an "unacceptable" level of disagreement between assessors (Gwet, 2014). Researchers writing about OSCE reliability have noted that calculating IRR across an entire OSCE, while useful for building a validity argument, does not allow for an identification of what specific items lead to lower levels of IRR. As such, it is necessary to break down the IRR statistics on a per-item basis (e.g., Cazzell & Howe, 2012), the approach taken here. The implications and key takeaways from these results can be found in section 4.5.6 below.

4.5.6 *Conclusions*

This section addressed the second research question: *What level of inter-rater reliability is evident in undergraduate nursing OSCEs?* As has long been noted in the field of assessment, conducting reliability analyses is a key step in the validation process of tests, in order to ensure that decisions made on the basis of awarded scores are defensible (Bandalos, 2018). This is particularly important when high stakes are associated with these decisions; for example, whether a student is allowed to progress to the next year of study. This study responded to calls from researchers in the field of nursing assessment, who have recently noted that published studies describing the development of OSCEs have often failed to include an explicit calculation of inter-rater reliability (Navas-Ferrar et al., 2017; Goh et al., 2019). One of the most common ways for IRR to be investigated is for multiple assessors to grade the same series of performances (e.g., Cazzell & Howe, 2012) and for the scores they award to be examined, the approach taken in the present study.

The results of the quantitative element of this study revealed considerable variations in how the 12 participants graded the same four OSCE performances, both in terms of the binary checklist items and the global questions. This suggests that the scores awarded to the students was contingent on the assessor who happened to be judging them, which violates the core assessment principle that a student's ability level should be the sole determinant of their score (Khan et al., 2013a). For three of the four videos, variation was present at the pass/fail decision regarding the student's overall performance: in videos P02BP, P03BP and P03NG, the student may have failed the OSCE as a result of the assessor. Given the wide range in cognitive processes that the assessors employed when judging the four performances (detailed in section 4.4), it is perhaps unsurprising that there was such variation in the scores they awarded. Nonetheless, such variation is concerning from an assessment perspective and needs to be unpacked further.

There are several key takeaways from the results outlined in this section. Firstly, the wide variation in the scores awarded to the four videos indicates the need for a stringent moderation procedure to be in place. As noted in section 4.4, the 12 assessors reported discrepancies in how moderation takes place within the School – the lack of scoring consistency evident in the present sample is perhaps indicative of this. From a validity perspective, it would clearly be indefensible not to allow a student to progress to the next year of study as a result of failing an OSCE she would have passed with a different assessor (as was the case with three out of the four videos here). The literature on OSCE design and development has described numerous approaches to moderation that can be employed, both before and after the OSCE takes place (Khan et al., 2013b). The 12 assessors who took part in the study indicated that it would be unlikely that a student would fail an OSCE based on the decision of a single assessor (in other words, the decision to fail a student is usually made in conjunction with the module coordinator). This is one form of moderation. However, such a process only improves the validity of the pass/fail decision, and only takes place after the OSCE has finished. Having moderation practices such as standard setting meetings before the OSCE is administered should reduce the need for these post hoc moderation procedures and bring about higher score reliability levels when the OSCE is administered (Rushforth, 2007).

Secondly, the results of the present study add to the debate about the role of subjectivity in performance assessments such as the OSCE. As noted in Chapter Two, researchers in both medical and nurse education have called into question the strict objectivity that the OSCE is designed to bring about (Mitchell et al., 2009; Hodges, 2013), arguing that it reduces the ability of expert assessors to make informed subjective judgements about students. The scores awarded by the 12 assessors in this study are unlikely to be interpreted as an endorsement for increasing subjectivity in OSCE assessment. The wide variation in how the four recorded performances were graded indicates that there is likely too much subjectivity at play, and that coordinators of these OSCEs should be trying to increase the level of objectivity in assessors' decision-making. If these results were to be replicated in other university nursing departments, it would indicate that this lack of subjectivity is common in undergraduate nursing OSCEs.

Thirdly, the relatively poor performance of the third-year student (videos P03BP and P03NG) is notable. As documented in Chapter Three, students from three different years of study were selected for participation as it was assumed that they would demonstrate differing ability levels, with the third-year student being the most competent at the two assessed skills. This turned out not to be the case, however the reason for this is unclear. One possible explanation is that the expectations regarding the assessed skills are different in the OSCE than they are in the world of clinical practice. The third-year student had significantly more real-world experience than the other two, and this may have affected how she performed the assessed skills. In particular, several participants remarked that this student felt comfortable skipping the measurement of an estimated systolic blood pressure in the BP OSCE, as this likely would not be performed in practice. This resulted in her losing marks from most of the assessors for items 9 and 10 in the BP OSCE. As such, this may indicate a potential theory/practice gap regarding what is expected in the OSCE and what is expected in clinical practice, which calls into question the strict alignment between the OSCE and the "real world" that is key to the validity of the OSCE (Hodges, 2003).

Finally, calculating IRR on a per-item basis allowed for an identification of which specific items within each OSCE were causing the most disagreement between assessors. This question has been investigated extensively in medicine (e.g., Brannick et al., 2011) and to a lesser extent in relation to undergraduate nursing OSCEs. Cazzell and Howe (2012) found that checklist items in the affective domain resulted in lower levels of IRR than those testing psychomotor skills. The results of the present study do not show a clear pattern regarding this distinction. Of the six items across the two OSCEs which resulted in a percent agreement statistic of 0.75 or less, only one (item 14 on the NG OSCE, *clearly states patient comfort is offered*), could be considered as assessing the affective domain, while the other five assessed psychomotor skills. However, this item did result in the lowest score of all 26 items in the two OSCEs, at 0.67. As such, it is not possible to determine conclusively that either psychomotor skills or affective skills are more likely to cause problems in terms of IRR. Indeed, the present study suggests that IRR problems may arise across all types of items.

4.6 *RQ3: Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?*

4.6.1 *Introduction*

So far, this chapter has examined the results of the qualitative (section 4.4) and quantitative data (section 4.5) separately, in order to determine the cognitive processes that the sample of assessors employed when judging the four recorded OSCE performances (qualitative) and the scores they awarded (quantitative). In section 4.4, it was demonstrated that assessor participants used a range of different mechanisms while grading the videos, and that these mechanisms were often idiosyncratic, in that they were not shared across the entire assessor pool. Perhaps unsurprisingly given this level of idiosyncrasy, the quantitative results detailed in section 4.5 revealed substantial variation in the scores awarded to the four videos. This variation was present in both the checklist and global items; and for three of the four videos resulted in a situation where the student was awarded a passing grade by some assessors but a failing grade by others. In the literature on assessor cognition and OSCE score reliability, most studies have adopted an exclusively qualitative (e.g., Yeates et al., 2013) or quantitative (e.g., Dunbar, 2018) approach, which parallels the structure of the present chapter so far.

For the final section of this chapter, the qualitative and quantitative data were analysed together, to investigate whether specific ways of judging the recorded performances could be linked with outcomes in terms of awarded scores. The issue of matching cognitive processes to awarded scores is one which has received minimal attention in the literature. Notable studies discussed in Chapter Two by Gingerich et al. (2014b; 2017) and Chahine et al. (2016) have indicated that, in a given assessor pool, notable "clusters" of assessors can be identified who approach the task of assessment in a similar way (for example, by consistently valuing communication as being the most important element of a performance) and award similar scores to each other. These studies allowed for a pinpointing of how IRR may be reduced as a result of assessor idiosyncrasy. However, none of these studies used nursing assessors in their sample, and only the study by Chahine et al. (2016) used OSCEs, with the research by Gingerich using postgraduate performance assessments of medical students. As such, published literature that seeks to explore the connections between cognitive processes and score reliability as they relate to undergraduate nursing OSCEs is scarce. This section sought to address this gap in the research.

The results of the quantitative element of the study revealed that there were several assessors that consistently awarded higher or lower scores than the rest of the sample (see tables 4.10 and 4.11). Given that the identification of extreme cases is an important step when conducting reliability analyses (Khan et al., 2013b), it was deemed logical to focus on these cases in order to explore how they approached the task of assessment, and whether a tentative explanation could be made for why they awarded particularly high or low scores. This approach is similar to "outlier sampling" described by Teddie & Yu (2007, p.81), which involves "selecting cases near the 'ends' of the distribution of cases of interest" for further examination, with the expectation that these cases "yield especially valuable information about the topic of interest". Such an approach is especially appropriate when the object of study is score reliability, as outlier assessors pose a particular threat to IRR (Bartman et al., 2013).

As such, the remainder of the chapter focuses on two assessors from the sample of 12 who participated in the study: one harsh assessor (who awarded consistently low scores to the four recorded performances) and one lenient assessor (who awarded consistently high scores). Firstly, there is a brief discussion of how the quantitative data were used to select these two assessors. Secondly, for each assessor there is an in-depth exploration of both the qualitative and quantitative data that was collected from them, in order to address whether any links can be drawn between how they approached the task of assessment, and the resultant scores they awarded. As such, while sections 4.4 and 4.5 represented an attempt to showcase the full breadth of data collected, the present section represents a different approach, by fully mixing the data in order to explore RQ3 in as much depth as possible. The section ends with a summary of the findings, and their relation to existing research into assessor cognition (section 4.6.5).

It is important to note at this juncture that the present phase of analysis necessitated a partial re-coding of some of the qualitative data. As noted by Braun and Clarke (2022), coding is an iterative and evolving process, and it is always possible for the researcher to return to his codes in order to glean more meaning from them. The coding strategy that informed section 4.4 focused largely on the semantic meaning of the data. The purpose of re-coding was to allow the researcher to identify latent codes; namely, those which explained the underlying assumptions and ideas that informed the assessment strategies of these two assessors.

4.6.2 *Outlier assessor selection*

There are numerous quantitative measures that could be used to identify particularly harsh or lenient assessor participants. A summary of these measures can be found in the table below. For each criterion, the harshest assessor is labelled in red, while the most lenient is labelled in green.

Table 4.14 *Criteria for selecting outlier participants*

Criterion:	Checklist items ticked (/52)	Global marks awarded (/36)	Overall "Fails"	Overall "Excellents"
Assessor 1	41	30	0	3
Assessor 2	43	22	0	2
Assessor 3	45	33	0	3
Assessor 4	35	13	1	0
Assessor 5	44	26	0	0
Assessor 6	40	23	0	0
Assessor 7	43	22	0	0
Assessor 8	33	18	2	1
Assessor 9	41	21	0	0
Assessor 10	41	24	0	0
Assessor 11	29	16	2	0
Assessor 12	44	22	0	0

The first potential measure is the overall number of checklist items ticked across the four videos. By this metric, Assessor 11 was the harshest, while Assessor 3 was the most lenient. Another measure is the allocation of scores for the global questions at the end of each marking guide. For each video, assessor participants answered three questions on a four-point Likert scale. The responses were coded from 0-3, meaning for each video there were a total of nine "points" available, for a total of 36 overall. By this measure, Assessor 4 was the harshest, while Assessor 3 was the most lenient. Finally, in terms of the classification awarded in the final global question (which assessed the student's overall performance), Assessors 4, 8 and 11 were the only ones to award a "Fail" to students, with the latter two each awarding this twice. In contrast, only four assessors awarded an "Excellent", with Assessors 1 and 3 each doing this three times. Using these measures, Assessor 3 was clearly the most lenient assessor, and was chosen for inclusion in this stage of analysis. Assessor 11 was the harshest on two out of the three criteria (checklist items ticked and overall "Fails" awarded), and the second harshest on the third (global marks awarded) and was chosen for inclusion.

4.6.3 *Assessor 11*

Assessor 11 has been employed in the School for over five years, and has been involved with OSCE assessment the entire time, including helping to assess both the blood pressure measurement OSCE and naso-gastric tube insertion OSCE in each of these years. She rated herself as an “expert” assessor. Examining this assessor’s transcript in detail uncovered several factors which may have affected the scores she awarded, thus potentially explaining why her grades were consistently lower than those of her colleagues. These were: *the use of safety as a criterion informing assessment decisions, confidence* in decision-making, *marking guide misalignment,* and *differences between video and face-to-face assessment*. These four factors are discussed in turn below.

4.6.3.1 *Safety as an assessment criterion*

When discussing her general assessment practices at the beginning of the interview, Assessor 11 described her role as an assessor as being one where she has to apply the marking guide to a student performance in order to award that student a grade:

> We normally have an assessment tool that we would follow. So there's each step that the student has to follow. So we would just make sure that they've performed each step of the assessment tool.

However, following this comment that her role is to “just make sure” the student has completed each step on the guide, she then mentioned safety as being the most important criterion that informs her judgements:

> The main thing is that they're safe. They're demonstrating safe practice. If they're not demonstrating safe practice, I just won't pass them.

This is in spite of the fact that “safety” is not an explicit assessment criterion in either of the marking guides used in the present study. As noted in section 4.4., the idea of safety is one that was frequently evoked by assessor participants when discussing their assessment practices. However, this has implications in terms of reliability, as different assessors may not share the same idea of what constitutes safe practice. Indeed, when asked to speak further about how her idea of safety is informed, Assessor 11 said the following:

> It would kind of be an overall sense that they don't know what they're doing. As I said, nerves do play a part coming into it. So if the student made an error, I would say to them at the end, “Is there anything you would have done different?” And if they said “no, everything was fine”, that's not OK, whereas if they say "oh, yeah, I should have done X, Y and Z". Well, then that's okay, they recognise it.

This quote indicates that her idea of safety is informed by two factors. Firstly, her idea of unsafe practice is “an overall sense” that the student is not able to complete the skill. Secondly, it is informed by reflexiveness on the part of students, in the sense that if a student did not perform well, but expressed awareness of their mistakes, this would be taken as a sign of nerves, and would be deemed by this assessor as “okay”, in contrast to a lack of awareness which is “not OK” and presumably unsafe. This discussion of safety is in contrast to what was described by other assessor participants when asked about their ideas of safety. Most commonly, assessors mentioned that their conception of safety within the OSCE has to do with the student not harming the patient, as discussed here by Assessor 4:

> Safety would be a big one. With safety, it’s that they don't do any harm, any damage to a patient.

This potential harm to the patient is usually related to either breaking infection control guidelines or performing a specific task within the OSCE incorrectly, in a way that could hurt the patient (again from Assessor 4):

> So in terms of safety, one of them will be the hand washing anyway, preventing infection. There will be one. But in terms of... there will be certain areas that they have to do so that they don't cause any harm, depending on which you're assessing, that they don't cause any undue harm to the patient. So, for example, the blood pressure if they blew, pumped up the cuff and walked away from the patient or left it there too long.

As such, these discussions about safety suggest that, in common with some of her colleagues, Assessor 11 evoked the idea of safety as being an important criterion informing her decision-making. However, when asked how she would tell if a student was being safe or not, her description was notably distinct from that of her colleagues. The reliability implications of this are twofold. Firstly, “safety” is not an explicit assessment criterion on either of the two OSCE marking guides used in the present study. This suggests that Assessor 11 was happy to deviate from provisions of the guide if she felt that a student was being unsafe, a clear contrast to the idea she expressed earlier, that her role is to “just make sure that they've performed each step of the assessment tool”. As such, her awarded scores are likely to be distinct from those of her colleagues who did not invoke safety as a criterion. Secondly, the fact that her idea of safety was different from at least some of her colleagues means that she may award a lower mark to a student that is “unsafe” according to her own idiosyncratic idea of safety than her colleague might who does not share the same idea as her.

Indeed, the data indicate that safety seemed to be the ultimate criterion informing Assessor 11's decision as to a student's overall performance in the OSCE. For three out of the four videos, this assessor discussed safety when asked to award an overall score to the student in the video. Regarding P02BP, Assessor 11 used the idea of safe practice as a justification for awarding the student a "Fail" for her overall performance:

> To be honest with you, if I was an assessor, I'd probably be looking to fail her, but going by the template, she'd be a borderline pass... I think this is the type of student who, she's not demonstrating safe practice. So in light of that, I'd probably fail her. Yeah. She hasn't demonstrated safety.

Here, Assessor 11 admits that she is "looking to fail" the student, but only feels empowered to do once she evokes the idea of safety. In other words, she seemed to be about to reluctantly award the student a "Borderline pass", but then began speaking about safety, and subsequently decided to award a "Fail". It is notable that Assessor 11 was one of only two in the sample that awarded this student a "Fail" for her overall performance. This indicates that the Assessor evoked the idea of safety as a justification for awarding a failing grade – a link between the process of judgement formation and the score awarded.

4.6.3.2 *Confidence*

A second notable factor which may explain why Assessor 11 awarded lower grades than her colleagues is confidence. This assessor frequently expressed the idea (either explicitly or implicitly) that there was minimal chance her judgements could be wrong, and that she only rarely has to get another assessor to check a recorded video in order to affirm her decision:

> No, I'd be confident enough. Maybe the odd one or two that I might say to a colleague "Look, can you just look at this". Mostly it can be due to a camera angle, and the quality of the video sometimes doesn't be that good. So you're kind of guessing, did they or didn't they? Yeah, but I'd say nine out of ten, I would know. When I say, "yeah, they have either done it or they haven't", that's it.

This idea of confidence was discussed in semantic terms by Assessor 11, as in the above excerpt. Additionally, looking at how Assessor 11 judged the four recorded performances reveals this idea of confidence underpinning her decisions, particularly her decisions to award low grades, or not to tick one of the binary checklist items. An extract from her discussion of the video P02BP illustrates this:

> Can I just stop that [the video] there? To me, that's not where the brachial artery is. The brachial artery is up here where the antecubital fossa is. So she's saying "I've located the brachial artery" I would say no, she hasn't. It should be here [gestures to correct position on own arm], just around there, so she's way down there. Where it should be anywhere kind of around the crease of the elbow. So she hasn't located the brachial artery. She might say, "oh, I have", but I can tell she hasn't.

Here, Assessor 11 was initially unsure as to whether the student in the video had correctly located the patient's brachial artery. As a result, she chose to stop the video and rewind so she could look again. Having done this, she then felt confident that the student had failed to do this correctly, and so chose not to tick item 8 on the BP marking guide (*locate brachial artery*). It is notable that this decision positioned her in a minority in terms of the 12 assessors, seven of whom did award this mark. A similar situation took place while she was watching video P03BP:

> I can't really see... No. Just a bit low... Can I just go back? No, it's not in the correct position. It's not her antecubital... It's quite hard to tell... No it's too low down.

Again, Assessor 11 expressed ambivalence about whether the student had completed a step in the marking guide correctly (in this case, item 7, *cuff bladder correctly placed on arm*). After rewinding the video to look again, she then felt confident in not awarding the mark to the student. This decision meant that she was one of only two assessors in the sample who did not tick item 7 for P03BP. As such, in these two instances, Assessor 11 used repeated viewings of the video to build confidence in her decision, and then chose not to award the student a grade for a specific item in the marking guide, a decision which placed in her in a minority of assessors. Thus, the process by which she built up confidence in her judgements (in this case, by rewinding the videos) allowed her to justify a decision not to award marks in these two videos. This idea of confidence will be revisited during the discussion of Assessor 3.

4.6.3.3 *Marking guide misalignment*

The third notable factor which was identified from an examination of Assessor 11's transcript is the perceived limitation of the marking guide in allowing her to express her judgements about a student. Notably, this assessor frequently discussed how the structure of the guide forced her to award higher grades than she wanted to. These discussions took place during the semi-structured interview, in which she discussed her assessment practices in general, and during the think alouds, in which she expressed this idea in relation to specific recorded performances. Regarding her assessment practices more generally, Assessor 11 noted that:

> The video OSCE, you're very much guided by the template. The template says it's there. It's there. Some of them, you look at them, they're absolutely excellent, but some of them you are like "I don't know", but it's there. They've done it… So if that's the case, you have to go by the template and say unless... You have to get really good justification as to why you haven't given it [a passing grade]. But if it's there and they've done it.

Here, Assessor 11 discussed how the structure of the OSCE marking guides (specifically, the "checklist" format (Rushforth, 2007)) often results in students being awarded higher grades than she thinks they deserve. This is because an assessor is forced to tick the items the student has completed satisfactorily, even if they believe the student is performing poorly overall. Assessor 11 indicated that she feels forced to "go by the template" even if it results in awarding grades she feels are too high. As such, this results in an incongruence between how well she believes a student is performing, and the grade awarded to that student. This issue is one that emerged while Assessor 11 was completing the think aloud section of the study. The following two comments she made while watching video P02BP serve to illustrate this:

> I'm just looking here. It says here, "brachial check to establish"... She was asked to give what her estimate was and she did. However, the artery is not correct. But she did do what she was told… So I would have to tick that, even though I know that's not right.
>
> No, stethoscope placement is not correct… But there's nothing to check here to say that.

In the first instance, Assessor 11 discussed how she had to tick item 9 on the marking guide (*brachial check to establish estimated systolic pressure*), even though she thought that the student had failed to complete the previous step (*locate brachial artery*) correctly. As such, the assessor believed that the student's perceived failure to locate the brachial artery invalidated her completion of item 9. However, because all items on the checklist are assessed independently, she felt obliged to award the mark for item 9. In the second instance, Assessor 11 noted that the student had placed the stethoscope in an incorrect position on the patient's arm, but said that there was no item in the guide about this specifically, and so was unable to deduct any marks for it. In both cases, this suggests a misalignment between how Assessor 11 processed the recorded performance, and how the marking guide directed her judgements.

In spite of her perceived frustration at the structure of the marking guide, it is noted that this is likely to have positive implications in terms of reliability. One of the explicit purposes of a standardised marking guide is to ensure that all students are assessed according to the same criteria (Khan et al., 2013b). As such, Assessor 11 deferring to the provisions of the guide (even if this was done reluctantly) should in theory lead to improved outcomes in terms of IRR. However, the global items at the end of each marking guide used in the present study, which asked assessors to judge a student's overall performance (rather than on a step-by-step basis) allowed Assessor 11 to award the failing grade she wished to, even though she had been "forced" earlier to tick more checklist items than she wanted (again related to P02BP):

> She has done some of the steps, but my overall impression would be to fail her... But what she has done here, I don't know what the marks are, but she has 1, 2, 3, 4, 5, 6 ticks. So she has done some parts correctly. But I know that she hasn't got the right pressure.

From this, it is suggested that the inclusion of the global items at the end of each marking guide allowed Assessor 11 to award a grade that aligned with her opinion of the student's performance in a way that the checklist items did not. However, given that she was one of only two assessors who awarded a "Fail" in terms of overall performance, this likely led to worse outcomes in terms of IRR.

4.6.3.4 *Differences between video and face-to-face assessment*

A final factor which may have led to Assessor 11 awarding low marks in the present study is the differences between face-to-face and in-person OSCE assessment. As discussed in section 4.4, this was a latent theme that was identified from the collected qualitative data. When discussing her assessment practices in general, Assessor 11 mentioned numerous ways that students completing face-to-face OSCEs could compensate for poor performance. One possible avenue for these students is through expressing awareness of any mistakes that they made:

> And very often throughout the procedure, if they've realised they've made a mistake and they recognise it there and then, "oh, the hell, I should have done that, I should have done this". I'd say that's fine.

Here, Assessor 11 indicated that she may refrain from awarding low marks to a student who did not complete a task in the correct way, as long as that student realised their mistake straight away and made it clear to the assessor that they knew they had done something incorrectly.

She also mentioned allowing a student to repeat a procedure if she wasn't sure the student had completed a task correctly, or else intervening in the OSCE to ask a theoretical question to the student about the scientific rationale behind what they were doing:

> So if that happened, I'd maybe ask a student to repeat the procedure or say to them "can you just explain to me what you're doing and why you're doing it" and give the rationale and if it's sufficient, well, fair enough. That's fair enough.

As such, it can be inferred from these comments that, when Assessor 11 is assessing face-to-face OSCEs, there are several different ways that students can mitigate poor performance: by expressing awareness of mistakes, repeating their performance of a task, or explaining the scientific underpinning of a procedure. However, when OSCE are assessed via video (as was the case in this study), there is no interaction between the student and the assessor. As such, the potential for students to improve their grades in the ways described above is removed. This was illustrated in the following comment about video P02BP:

> So there's no rationale. Like if you spoke to the student and said, well, “why is the cuff on that low?” There'd be no [answer] - "I don't know."

Here, Assessor 11 indicated that if she was assessing this performance in-person, she would intervene in the OSCE and ask the student to explain the scientific rationale that informed her positioning of the cuff bladder. However, because she was unable to do this through video, she instead makes an inference about the student’s lack of scientific knowledge, and uses this inference as part of her justification to award a “Fail” to that student for her overall performance.

4.6.4 *Assessor 3*

Assessor 3 has been employed in the School for less than five years, and has been involved with OSCE assessment for two administrations of OSCEs. She is involved with teaching both blood pressure measurement and naso-gastric tube insertion to students, but at the time of interview had not been involved directly with the assessment of these OSCEs. She rated herself as a “competent” assessor. Examining this assessor’s transcript in detail uncovered several factors which may have affected the scores she awarded, thus potentially explaining why her grades were consistently higher than those of her colleagues. These were: *the use of communication as a criterion informing assessment decisions, confidence* in decision-making and *marking guide misalignment.* These three factors are discussed in turn below. It is noted that this juncture that these factors are directly related to the issues discussed regarding Assessor 11, but had the opposite effect on awarded scores. For example, while Assessor 11 mentioned feeling confident enough to award low scores, Assessor 3 described not feeling confident enough to award low scores, and thus giving students the "benefit of the doubt". The implications of these divergences in terms of reliability are discussed throughout the section.

4.6.4.1 *Use of communication as an assessment criterion*

When discussing her general assessment practices at the beginning of the interview, Assessor 3 mentioned how she uses the Royal Marsden standardised procedures to guide her judgements of students:

> A lot of our skills are just standard. So I suppose I know by that, whether or not I kind of know them off by heart, I suppose being an experienced nurse as well and just kind of keep myself up to date with my standards. That's probably what guides me with doing it right.

These procedures informed the development of both marking guides used in the present study, as the module coordinators for both the BP and NG OSCEs based the marking guides on these procedures. As such, Assessor 3's comment can be understood as her saying that her role as an assessor is to apply these standardised criteria. In spite of this comment that the standards are what guides her judgements of students, she then mentioned communication as being the "first thing" she looks out for while assessing OSCEs:

> When I'm assessing, even though we use mannequins or use each other, that they're actually talking to the person or talking to the mannequins, so they remember they're working with people... You have to tell the person exactly what you need. Consent - it's the 21st century. They have the right to say no and everything like that. So that's probably the first thing. Like I said, they're actually talking to the mannequin.

This comment suggests that Assessor 3 places a particularly high level of importance on the extent to which a student completing an OSCE is able to communicate with the "patient" (usually a mannequin). When asked how she would be able to tell if a student was displaying good communication, Assessor 3 said this:

> Just being professional, saying hello, stating your name, stating that you're a student nurse, tell the patient exactly what's going to happen, what to expect, and just

> conversation... They have just that holistic approach to the situation because we're looking at people, you can learn a skill, but you can't really learn that dynamic.

What is notable in this comment is that Assessor 3 conceives of effective communication as being "holistic". As such, while an item on an OSCE marking guide may assess communication (e.g., item 3 on the BP OSCE used in this study, *explains procedure to the patient and gains consent*), Assessor 3 seems to value communication across the entire performance, such that a student who demonstrated effective communication with the patient would be perceived by Assessor 3 as being superior to a student who did not, even if they each completed the same number of steps on the marking guide. The importance that Assessor 3 places on communication is further emphasised by her response when asked what would distinguish a borderline pass performance from a failing performance:

> I would say their communication with the patient... If their skills were like some marks off, if they were really nice to the patient, I would [pass them].

In this comment, Assessor 3 indicated that a student could compensate for missing some of the required steps in an OSCE if they demonstrated good communication, particularly if they were "really nice to" the patient. As such, it can be suggested that for Assessor 3, the ultimate criterion for deciding whether to pass or fail a student is that student's communication skills. This is in marked contrast to the opinion expressed by Assessor 11, who noted that safety was the criterion that was most important to her while deciding on a student's grade. As such, while both Assessors 3 and 11 indicated that they used either the marking guide or the Royal Marsden standardised procedures (which inform the marking guide) to guide their judgements when assessing OSCEs, in fact they both seem to be guided by other criteria: either safety or communication.

The implications of this in terms of reliability are twofold. Firstly, given that the function of a marking guide is to ensure that multiple assessors grade a performance in the same way, the fact that Assessors 3 and 11 freely admitted to judging students against criteria that are not explicitly contained within the guide is likely to lead to divergence in how they would score the same performance. Secondly, given that these two criteria they used to inform their judgements are different, it is possible that an OSCE performance that was considered "safe" by Assessor 11, but where the student had minimal interaction with the patient, would be awarded a passing grade by Assessor 11 but not by Assessor 3. Indeed, this was made explicit in the following comment by Assessor 11:

> You might have a student that comes in that's all chatty and great with the talking to the mannequin, but they might not demonstrate safe practice, where you might have the student that comes in and doesn't open his or her mouth. But they have demonstrated safe practice [so they would pass].

As such, it can be inferred that these two assessors' different opinions about the most important aspect of an OSCE performance would lead to divergent outcomes for students in terms of awarded scores.

4.6.4.2 *Confidence in decision-making*

The issue of confidence when making assessment decisions is another theme that was identified in Assessor 3's transcript. However, in contrast to Assessor 11, who frequently invoked her own confidence in order to justify awarding low marks, or choosing to fail a student, Assessor 3 mentioned not yet having the confidence to do this. Speaking of how she expects her assessment practices to change as she gains more experience, Assessor 3 noted that:

> Yeah, I suppose I think the more I do, maybe the more confident... I do feel competent.... So I suppose just being a bit more definite of how I feel rather than being like "aww" and overthinking it and replaying it and thinking yes or no. Just deciding. No, he didn't meet it.

Here, Assessor 3 is explicit that, at the moment, she often does not feel "definite" about the grades she awards, and seems to be in two minds frequently about the grade a student deserves. In this case, she links this lack of confidence specifically with her reluctance to award low grades to students. This idea emerged again when discussing the marking guide for the BP OSCE used in the present study. Regarding the cuff placement (item 7 on the marking guide), she noted that:

> Yeah. I'd have to focus in to see that little line. I have to take their word that they found it.

Here, she discussed how she finds it difficult to tell through video whether a student had placed the cuff in the correct position (in other words, whether the "line" on the cuff was in the right place on the patient's arm). As a result, she would have to "take their word" that they had put it in the correct place, and presumably award the mark to them, even though she might actually not be sure that it was correct. Assessor 3 made a similar comment regarding the extent to which she would be able to tell if a student had correctly located a patient’s brachial artery (item 8 on the marking guide):

> In an OSCE they could go, “oh yeah, I hear it [the Korotkoff sounds]. Yeah. That's 120” or whatever…They're telling me that's what they heard… So I have to take that. So yeah, I wouldn't know. I'd have to take what they said.

Again, Assessor 3 described how she might not be able to tell if the student had correctly located the patient’s brachial artery (and could therefore hear the patient’s pulse), and as a result would choose to award them the mark for that item. This decision-making process was evident while Assessor 3 watched the video P02BP, and was asked by the researcher whether she thought the student had correctly located the patient’s brachial artery:

> Yes… It is usually more on the inside of the arm, but if she feels it there. She might have felt it there… Well, I'll take her word for it.

Assessor 3 seemed to be ambivalent about whether the student had located the artery correctly, admitting that it is “usually” on a different part of the arm to where the student had located it, and noting that she “might have” felt the patient’s pulse. In spite of these apparent reservations, she chose to award the mark to the student (item 8 on the BP OSCE). This is a notable contrast to what happened when Assessor 11 was watching the same video (discussed in section 4.6.3.2 above). Assessor 11 was also unsure initially as to whether the student had located the artery, and chose to rewind the video to watch it again, ultimately becoming confident in her decision that the student had failed to find the artery, and so not awarding her the mark for item 8. This suggests a clear discrepancy in how confident these two assessors were regarding their assessment decisions, and how this led to divergence in the mark they awarded for item 8 on the BP OSCE. This may partially explain the large division in how these two assessors graded the P02BP video: Assessor 3 awarded 12/12 marks for the binary checklist items and three “Excellents” for the global items, while Assessor 11 awarded 7/12 binary checklist items, a “Good” for the student’s overall communication, and a “Fail” for the final two global items.

4.6.4.3 *Marking guide misalignment*

A final issue that was identified from Assessor 3’s data was a misalignment between the provisions of the marking guide, and her own judgements of students. This idea was noted in Assessor 11’s data as well. However, while Assessor 11 mentioned that the marking guide often forces her to be more generous in terms of awarded marks than she wishes to be, the opposite was true for Assessor 3. In other words, Assessor 3 frequently expressed that the structure of the marking guide often compels her to award lower grades to students than she thinks they deserve. Assessor 3 noted that she often wants to be lenient to students - in terms of the grades she awards to them - due to their perceived inexperience, but is not allowed to do this because she has to follow the marking guide:

> A lot of them are just out of Leaving Cert… So I do have that element of, give them a little bit of leeway… But then obviously we have the standard… So I'm conflicted with it myself, but I have to go as per the rubric and if they don't hit it, I can't give it to them.

As such, the fact that she has to judge students on the basis of the marking guide is a source of “conflict” for her, seemingly because it means that she has to award lower grades to student that she wishes to. This idea of being constrained by the marking guide recurred throughout the think aloud portion of the interview. Regarding P03NG, Assessor 3 said the following:

> So she didn't do that. I don't think the spigot was on [item 13] and at the end she didn't offer the comfort [item 14]. So that's how I'd mark her kind of step by step. But if I was giving her feedback, that was excellent, though… It's just because I have to measure it to this [marking guide] for fairness to everyone.

Here, Assessor 3 discussed how she thought the student was “excellent”, even though she failed to complete the final two steps on the binary checklist. She indicated regret at being forced not to award marks for these two steps, noting that it would be unfair if she chose to give the student marks for these two items. This idea emerged again while she was watching video P03BP:

> Yes. You do need to get the estimated systolic blood pressure [items 9 and 10]. So therefore you know how high to inflate the balloon. Yes. In practice, I don't do that… if I was looking after a patient, I'd do it exactly like that... So I have to X her in that, even though that's not wrong. But I couldn't mark her with a tick and someone else actually did it. Clearly. It wouldn't be fair.

In this extract, Assessor 3 remarked on the fact that the student failed to take an estimated systolic blood pressure reading from the patient (items 9 and 10 on the BP marking guide). However, she invoked her own experience as a justification for the student’s actions, saying that if she were to take a blood pressure measurement, she also would not complete this step. Ultimately, she had to not award the marks for these two items, even though she perceived what the student did as being “not wrong”. This suggests again that Assessor 3 wished to award a higher grade to the student than the marking guide compelled her to do, recognising that acting outside the provisions of the guide would not be fair to other students completing the same OSCE. As such, this represents a notable divergence from the ideas expressed by Assessor 11, who commented frequently that the structure of the marking guide compelled her to be more generous in terms of the marks she awarded to students.

4.6.5 *Conclusions*

This section sought to address the third research question, *Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?* This is an issue which has received minimal attention in assessor cognition literature thus far, particularly in nurse education. The analysis approach taken was outlier sampling, with data from the harshest and most lenient assessor analysed in depth, in order to suggest reasons for the large divergence in the scores they awarded to the four recorded videos used in the study. This phase of analysis revealed four notable factors which may have contributed to this score variance.

Firstly, in spite of Assessors 3 and 11 professing to using either the marking guide or the Royal Marsden standardised procedures to inform their judgements, both invoked other criteria that seemed to be the most important factor affecting their assessment decisions: a student's communication skills (in the case of Assessor 3) and whether the student had demonstrated safe practice (Assessor 11). This is concerning from an assessment standpoint, as deviating from the provisions of the marking guide is likely to reduce score reliability (Khan et al., 2013b). The phenomenon of assessors admitting to judging students against criteria which are not explicitly contained within the marking guides has been noted elsewhere in literature on assessor cognition. In interviews with 25 nursing assessors, East et al. (2014, p.463) determined that "overwhelmingly assessors determined students' competence subjectively". Indeed, in their study, they found that considerations as to whether the student had demonstrated safe practice, as well as whether the student had displayed adequate communication, were two of the most important factors informing assessors' judgements. The findings from the present study further indicate the apparent importance of these two factors in undergraduate nursing OSCEs. However, given that these were not explicit criteria in either marking guides, this indicates a disconnect between what assessors perceive to be important in an OSCE performance and the content of the marking guides used to assess that performance. In light of the noted tendency for assessors to be willing to judge students outside the provisions of the marking guide if they deem it appropriate (e.g., Hyde et al., 2022), the findings of the present study suggest that issues with score reliability are (at least partially) caused by this disconnect.

A second factor which appeared to explain some of the divergence in scores awarded by Assessors 3 and 11 is confidence in decision-making. Assessor 11 had over five years of experience assessing OSCEs, and rated herself as an "expert" assessor. Throughout her interview, she appeared able to leverage this experience to confidently award low marks to students, believing there was little room for ambiguity in her decisions. This is in contrast to Assessor 3, who had less than five years of experience as an assessor, and rated herself as "competent". She frequently expressed being unsure as to whether a student had completed a checklist item correctly, and usually chose to give the student the "benefit of the doubt" and award them the mark. As noted by Govaerts et al. (2013), experienced assessors can use their "task-specific" knowledge in order to assess performance more accurately than their inexperienced counterparts. In order words, experienced assessors may develop expertise related to specific tasks which inform their judgements. This finding is notable in light of the divergence in how Assessors 3 and 11 judged the video P02BP, particularly regarding item 8 on the checklist (*locate brachial artery*). Both assessors expressed some level of ambivalence in relation to whether the student had completed this item correctly. However, after rewinding the video, Assessor 11 became confident that the student had missed the artery, and did not award her the mark, while Assessor 3 chose to award the mark even though she admitted there was a chance the artery had not been located correctly. This suggests that Assessor 11 has developed confidence related to her task-specific knowledge of assessing blood pressure measurement, and was able to use this to justify not awarding the mark for item 8. Indeed, across the sample of 12 assessors, the five who chose not to award the mark for item 8 (Assessors 4, 7, 8, 9 and 11) all had more than five years of experience assessing OSCEs, while the three assessors with less than five years of experience all awarded this mark. As such, these findings indicate that IRR may be negatively affected due to the differing levels of experience in a sample of assessors, and the resultant divergence in feelings of confidence in awarding low marks.

Thirdly, there were notable differences in how Assessors 3 and 11 engaged with the marking guides used in the present study. Assessor 3 frequently discussed how the marking guide forced her to be harsher (i.e., award lower grades) to students than she wished to be, while Assessor 11 described the opposite phenomenon. The issue of a misalignment between how assessors judge students and the marking guides that are used to direct their assessments has been widely noted in the literature on medical assessment (e.g., Tavares & Eva, 2013). Put simply, research in assessor cognition has consistently indicated that assessors judge performance in a holistic or global way (i.e., related to a student's performance as a whole). As such, the process of having to assess students on the basis of a series of checklist items (as in the current study) may be cognitively demanding, as assessors have to "translate" their holistic judgements in order to fit within the structure of the guide (Yeates et al., 2013). The findings from the present study are novel in that they suggest a similar phenomenon is at play in undergraduate nursing OSCEs. Both Assessor 3 and Assessor 11 seemed to judge the recorded performance in holistic terms, and then had to adjust their judgements to fit the structure of the guide. For example, regarding P03NG, Assessor 3 remarked that the performance was "excellent" even though the student failed to complete several items on the checklist. From a reliability standpoint, this misalignment may serve to increase agreement between assessors on the binary checklist items, given that it compels them to judge students according to their completion of these items. However, this is not the case regarding the global items, for which assessors can draw on any information they see fit when awarding a grade. This may explain why, for video P02BP, Assessors 3 and 11 awarded similar marks for the binary checklist items (8/12 and 6/12) but considerably divergent scores for the global items ("Excellent", "Good" and "Good" compared to "Good", "Fail" and "Fail"). As such, the results of this study indicate that the process by which assessors are forced to adhere to the structure of the marking guide may increase IRR for binary checklist items but not necessarily for global items.

Finally, the fact that assessors in the present study had to judge student performances via video is likely to have affected the reliability of the scores they awarded. Research on conducting performance assessments such as the OSCE through video has proliferated since the COVID-19 pandemic. However, published work thus far has largely been concerned with logistical issues inherent in running an assessment (e.g., Major et al., 2020), rather than focusing specifically on how the use of video affects assessors' cognitive processes. The present findings indicate that the process of judging an OSCE performance through video cannot be assumed to be the same as judging face-to-face. Assessor 11 noted several ways that the use of video limited her ability to judge students as she wanted to, for example by preventing her from intervening in the OSCE to ask students if they understand the scientific rationale behind what they are doing. This is likely to have contributed to the relatively low scores she awarded in this study: for video P02BP, she indicated that she would have asked the student a question related to blood pressure measurement. The fact that she was unable to do this meant that the student had no opportunity to compensate for the perceived poor aspects of her performance, and so received a low grade from Assessor 11. All in all, the present study highlights that the use of video in undergraduate nursing OSCEs may have unexpected outcomes in terms of score reliability, and should be investigated further.

Chapter Five

Discussion

5.1 *Introduction*

This research investigated the cognitive processes that assessors go through when judging undergraduate nursing OSCEs, and explored the links between these processes and the inter-rater reliability of the OSCEs. This chapter summarises and contextualises the findings from the study in relation to the broader research landscape (which can be found in Chapter Two). Section 5.3 details the practical implications of the present findings in terms of OSCE assessment, both for the School and beyond. As noted in section 3.4, the use of a case study approach means that the findings of the current study are directly relevant to the School, but are likely to be pertinent to other organisations who wish to develop, or are currently using, nursing OSCEs. Section 5.4 details the theoretical implications of the current study, while section 5.5 outlines the strengths and limitations. The chapter ends with a summary of recommendations for future research (section 5.6), as well as an epilogue (section 5.7).

5.2 *Summary of research*

OSCEs have become increasingly popular in undergraduate nursing education, driven by the recognition that robust, well-designed assessments are an important tool in ensuring that newly qualified nurses are prepared for the world of clinical practice (Goh et al., 2019). However, as with all tests, it is imperative that the core assessment principles of validity and reliability are upheld (AERA et al., 2014). This is especially true when high stakes are associated with the outcomes of such assessments, for example the decision to allow a student to progress to the next year of study, as is the case with many undergraduate nursing OSCEs (Goh et al., 2019).

The importance of examiner consistency as a key factor affecting score reliability has long been noted by researchers in nurse education (e.g., Rushforth, 2007). This focus on assessor consistency has also been noted in more recent literature: in a review of 28 studies on the topic of challenges facing nursing assessment designers, Anim-Boamah et al. (2021, p.352) reported that "the strongest predictor of a student passing their clinical skills assessment was the leniency… of the examiner". In other words, a student's score on an assessment may be determined by the assessor who happens to be grading them, rather than by their own ability level. This highlights the ongoing need to determine how variance between assessors arises and suggest means by which it can be accounted for.

Responding to these trends, this thesis investigated the following three research questions:

> RQ1: *What are the cognitive processes assessors employ when judging undergraduate nursing OSCEs?*
>
> RQ2: *What level of inter-rater reliability is evident in undergraduate nursing OSCEs?*
>
> RQ3: *Is there evidence that assessors' cognitive processes lead to specific outcomes in terms of awarded scores?*

In order to address these questions, a mixed-methods approach was employed. The researcher filmed six videos of three students each completing two OSCEs: blood pressure measurement (BP) and naso-gastric tube insertion (NG). There was one student each from first- (P01), second- (P02) and third-year (P03) of a General Nursing programme. The BP OSCE was performed on a real person, while the NG OSCE was performed on a mannequin. Four of these videos were selected for inclusion in the study: the first-year student completing the NG OSCE (P01NG), the second-year student completing the BP OSCE (P02BP) and both third-year student videos (P03BP and P03NG).

After the recording of the videos, 12 assessor participants underwent a semi-structured interview in which they were asked about their role as assessors and how they make decisions about students. Additionally, they completed a cognitive interview relating to their interpretations of both OSCE marking guides. Finally, they participated in a think-aloud protocol, in which they watched the four recorded videos and vocalised their thoughts as to how well or poorly the students were performing. These formed the qualitative data used as part of the study. Participants also completed the marking guide for each of the four videos, for which they had to decide whether to tick each binary checklist item (12 for the BP OSCE and 14 for the NG OSCE) they perceived the student to have completed successfully, as well as answering three global questions at the end, relating to the student's overall performance. These formed the quantitative data used in the study. A copy of the marking guides used for the study can be found in Appendix G. Detailed information about the research methodology can be found in Chapter Three.

5.3 *Key findings*

5.3.1 *Summary of findings*

Regarding RQ1, thematic analysis of the collected qualitative data was used to determine the cognitive processes that assessors go through while judging undergraduate nursing OSCEs.

This is an issue which has been investigated in research on assessment of medical students (e.g., Kogan et al., 2011; Yeates et al., 2013, Boursicot et al., 2021), but has yet to receive significant attention in the literature on nursing assessment. The present study found that assessors' cognitive processes could be grouped according to the three stage framework of *observation*, *processing* and *integration* outlined by Gauthier et al. (2016), with additional factors, labelled as *assessor development*, affecting assessors' processes of judgement formation in the long-term. In the traditional psychometric view of high-stakes assessment, it is important that assessors approach the task of assessment in broadly similar ways, such that any difference in awarded scores can be ascribed to differences in the "true" skill level of the test-taker (Khan et al., 2013a). The results of the present study suggest that this is far from the case. The prevalence of potentially subjective factors has the potential to threaten the reliability of awarded scores, as they increase the chance that a student's score on an OSCE will be overly affected by the assessor who happened to be grading them.

Regarding RQ2, the quantitative data were analysed to calculate the level of inter-rater reliability present across the four videos. This analysis responded to calls in the literature on nursing OSCEs for researchers to include an explicit calculation of IRR, with authors such as Navas-Ferrar et al. (2017) and Goh et al. (2019) noting that such a calculation is often absent from published work in the area. The results of the present study revealed considerable variation in how the four performances were graded by the 12 assessors. This variation is, in a sense, unsurprising, given the range of idiosyncratic judgement mechanisms identified. Nonetheless, such variation is concerning from an assessment standpoint, as it threatens the validity of decisions made on the basis of assessment scores. This is particularly notable in light of the finding that, for three out of the four videos, the pass/fail decision was contingent upon the assessor; the students in these videos would have passed the OSCE with some assessors but failed with others.

Regarding RQ3, outlier sampling was used to explore the links between the cognitive processes that assessors went through while judging the videos, and the scores they awarded. This section builds on work by Gingerich et al. (2014b, 2017) and Chahine et al. (2016), who sought to uncover the specifics of how issues with IRR arise due to the idiosyncratic ways that assessors approached the task of assessment. In the present study, qualitative and quantitative data from the harshest and most lenient assessors were contrasted in order to suggest reasons why they awarded such divergent grades. The results suggest four factors that could account for this: *the use of safety and communication as assessment criteria*, *confidence in decision-making,* *marking guide misalignment*, and *differences between video and face-to-face assessment*. Although this phase of data analysis was exploratory and only used data from two assessors, it nonetheless suggests a tentative explanation for specifically how score variance arises.

5.3.2 *Implications of key findings*

The use of a case study approach, for which OSCEs administered in the School were used to investigate assessors' cognitive processes and score reliability, means that the results of the present study have implications for assessment design within the School. However, there are also implications for the broader assessment landscape beyond the School. OSCEs have been incorporated into nursing assessments at postgraduate level, for example the UK's Nursing & Midwifery Council Test of Competence for nurses who trained abroad and wish to practice in the UK (Nursing & Midwifery Council, 2022). These implications are discussed below, grouped under the relevant headings of *marking guides, assessor idiosyncrasy, use of video in OSCE assessment*, and *moderation practices*.

It is noted at this juncture that researchers writing about OSCE design and development within undergraduate nursing programmes have emphasised that the administration of OSCEs (and other performance assessments) is costly and time-consuming for staff (Solà-Pola et al., 2020). As such, any recommendations or suggestions are made with full awareness that they might not be realistic within the constraints of the School. Indeed, the five staff members interviewed in Phase 1 of the study noted on several occasions that the OSCEs as currently administered represented the best possible approach in light of these constraints.

Additionally, it is important to emphasise that the primary aim of the present study was to de-privatise the cognitive processes that assessors go through while judging OSCEs, rather than to evaluate how effective or ineffective these OSCEs are. In order to facilitate this de-privatisation, several adjustments were made to the OSCEs included in the study (particularly the inclusion of the three global items at the end of each marking guide). As a result, the following implications are discussed with awareness that the OSCEs used in this study are not direct copies of the ones which are actually administered in the School, even if there is a large degree of overlap.

5.3.2.1 *Marking guides*

There are numerous findings from all three phases of data analysis undertaken in the present study that relate to the structure and content of the marking guides used to assess undergraduate

OSCE performances. A notable finding that was identified in both the thematic analysis of assessors' judgement processes (section 4.4.4.1), as well as the outlier analysis of Assessors 3 and 11 (sections 4.6.2.1 and 4.6.3.1), was the use of criteria not contained within the marking guides to assess the recorded performances. In particular, the concepts of safety and communication were highly important in informing assessors' judgements as to how well or badly the students were performing. This raises the question of how the content of OSCE marking guides is devised, and the resultant effects on validity and reliability.

The issue of what elements are contained within the marking guide has implications in terms of content validity, defined as "the relationship between the content of a test and the construct it is designed to measure" (AERA et al., 2014, p.14). The focus by assessors in the present study on the issues of safety and communication echoed the results of a 2014 study by East et al., who found that these two concepts were among the most important factors affecting assessors' decisions about OSCE performances. As such, it is suggested that nursing assessors believe that these two ideas are central to effective OSCE performance, and should be included in OSCE marking guides in order to maximise the content validity of the assessment.

The lack of explicit criteria pertaining to safety also has implications for both the face validity and reliability of the OSCE. Face validity refers to the perception that an assessment is measuring what it is designed to measure (AERA et al., 2014). The emphasis by assessor participants on safety as an important element of effective nursing practice suggests that an OSCE marking guide which does not include this criterion may be perceived by assessors as less valid. As discussed in the previous chapter, this has implications for score reliability, as assessors may choose to openly defy the marking guides and judge students according to what they (the assessors) perceive to be the most important aspects of a particular skill. Indeed, as documented in a recent study by Hyde et al. (2021), assessors who did not believe that an OSCE was measuring what it was supposed to measure were explicit that they would use their own criteria in order to reward students they thought were performing well (and punish those who were not).

The question of what institutions such as the School can do to account for these issues with content and face validity is an important one. One possible solution is including assessors in the drafting of marking guides, in order to bring about greater alignment between the content of the guides and what assessors actually believe to be important (Tavares & Eva, 2013). A study by Hyde et al. (2020) is illustrative of such an approach. They recruited 12 medicine assessors to devise an OSCE marking guide from scratch, and found that there were five aspects of performance (including "Safety") that the assessors believed to be central to effective practice. They speculated that such an approach would reduce the instances of assessors "deviating" from the marking guides due to "the belief that overlooked significant factors are at play" (p.19). However, it is also important to stress that the content of undergraduate nursing OSCE marking guides (particularly the list of steps required in order to complete a specific procedure) is often derived from standardised procedures such as those published by Royal Marsden (Dougherty et al., 2015). As such, a potential approach could be to develop a modified marking guide that combines these standards with the additional criteria, such as safety, that assessors believe to be important. Such marking guides may have greater buy-in from assessors than those currently in use, which should reduce instances of assessors acting outside the guide when awarding grades to students.

Another notable finding related to the marking guide is that, of the items that are included in the guide, assessors have idiosyncratic ideas about which are the most important (section 4.4.2.1). This was true even when (as was the case in the present study) all items are weighted equally. Notably, the present study revealed that certain assessors mentioned psychomotor tasks as being the most important, while others mentioned affective tasks. This has clear implications for score reliability, as a student who is strong in the affective domain but weak in the psychomotor domain may receive a higher score from an assessor who values affective skills. The link between assessors having idiosyncratic ideas about the relative importance of items on the marking guide, and the resulting scores they award, has been explored by Chahine et al. (2016). They found that assessors could be grouped into two "clusters", and that assessors in each cluster awarded similar marks to each other.

Relatedly, assessors in the present study also discussed the idea that certain items within OSCE marking guides were so important that a student should not be allowed to pass the exam without completing them successfully (section 4.4.4.2). In other words, specific items on the checklist were considered by some assessors to be so instrumental to performing a procedure that the rest of the items were meaningless if these important steps had not been completed. These "red line" items varied depending on the assessed OSCE (and, indeed, depending on the assessor). Four of the participants noted that the OSCEs administered in the School used to contain such items; however, at the moment there are no items within OSCE marking guides awarded such importance.

The issue of what the School, and other institutions devising OSCE marking guides, can do to account for assessors' opinions about the relative importance of different items within the marking guides is a notable one. One possible solution is simply to introduce a weighting system, such that different items on the guide are allocated marks depending on their perceived importance, an approach that is common in medicine OSCEs (Khan et al., 2013b). Such an approach would require active input from assessors into the drafting of the guide, in order to determine this weighting, as well as any potential "red line" issues. However, as noted by Rushforth (2007), the choice of mark allocation and identification of "red line" issues is a subjective choice. Indeed, the results of the present study suggest that assessors may not share the same ideas about which items on the marking guide are the most important. In any case, seeking input from assessors is likely to be a starting point for facilitating such discussions.

Finally, assessors' comments about the marking guides used in the present study revealed that they had different strategies for translating their checklist scores into a global score at the end (sections 4.4.4.2 and 4.6.5). This was particularly notable for the video P03BP, to which Assessors 3 and 11 awarded similar checklist scores (8/12 and 6/12) but considerably divergent global scores ("Excellent", "Confident" and "Good" compared to "Good", "Not at all confident" and "Fail"). Indeed, the relationship between the checklist and global scores was discussed explicitly by Assessor 11, who was ambivalent about choosing to fail the student even though she had completed half of the checklist items successfully.

The use of checklist and global items on OSCE marking guides has long been the subject of debate within the nursing OSCE literature (Rushforth, 2007). Rushforth noted that using marking guides which contained only binary checklist items was assumed to increase the IRR of OSCE scores. However, empirical studies of IRR have found that binary checklists do not necessarily lead to improved IRR outcomes in every case, and the addition of global or holistic items can result in high levels of IRR and additionally allow for a more nuanced description of a student's performance (Rushforth, 2007). It was with the aim of uncovering more nuances of assessors' decision-making processes that three global items were included in the marking guides used in the present study.

Additionally, the use of only checklist items on a marking guide may have an effect on the consequential validity of the assessment, defined as the "consideration of negative consequences of test use" (AERA et al., 2014, p.11). Researchers writing about the use of OSCE in both medicine (e.g., Hodges, 2013; Khan, 2017) and nursing (e.g., Mitchell et al.,

2009) have expressed concerns that the popularity of OSCEs at undergraduate level may lead to a situation where students can pass the exam simply by completing a series of discrete steps, rather than by displaying appropriate holistic performance. The incorporation of some global items into the marking guide allows assessors to identify and reward students who perform well in an overall sense, in addition to completing many of the required steps. As such, the inclusion of global items on the marking guide may have positive effects on validity, as students who wish to receive a good score will have to complete the required steps but without coming across as "robotic" (Khan, 2017, p.2), which better prepares them for the world of clinical practice. Ultimately, institutions such as the School should consider the extent to which the inclusion of global items within the marking guide is appropriate, bearing in mind that (at least based on the results of the current study) assessors tend to process OSCE performances in a global sense anyway.

5.3.2.2 *Assessor idiosyncrasy*

The results of the present study suggest multiple ways in which assessors' processes of judging and grading undergraduate OSCE performances can be thought of as idiosyncratic (Gingerich et al., 2014a). Findings from the thematic analysis indicate that assessors who participated in the present study may lack a fixed sense of what constitutes "good" practice, and so judge student performances against those previously seen, or what they would do themselves if they were completing a specific skill (section 4.4.3.1). This is a finding that has been widely noted in previous studies of assessors in medicine, most notably by Yeates et al. (2013) who described this phenomenon under the heading of *criterion uncertainty*. The potential effects of these comparisons in terms of score reliability are notable, as a student's grade may be affected by whoever happened to be judged before them, instead of by their own skill level.

The issue of assessors using comparisons to inform their judgements of student performances raises the question of what can be done by those administering OSCEs to mitigate the potential negative effects on reliability. One potential solution is the development of detailed scoring rubrics, defined by Jonsson and Svingby (2007, p.131) as "a scoring tool for qualitative rating of authentic or complex student work". A well-designed scoring rubric not only specifies the elements that constitute a larger task, but also details the level of performance required in each of these elements in order for a student to be awarded a specific grade. As documented in Chapter Three, assessors who participated in the present study reported that the marking guide is generally the only document they receive which guides their grading of OSCE performances.

For example, item 3 on the binary checklist for the BP OSCE states *explain the procedure to the patient and gain consent.* Assessors described varying approaches that students could take while completing this item: as a result, there was a wide variety in what students could do that would be perceived as satisfactory by different assessors, leaving open the possibility that a student's process of gaining consent could be deemed as appropriate by one assessor but not another. The use of a scoring rubric would detail what specifically students have to do in order to complete this item correctly, reducing the possibility for unwanted score variance between assessors. Indeed, "increased consistency of judgement when assessing performance" (Jonsson & Svingby, 2007, p.132) is one of the most widely noted advantages of using rubrics. As such, organisations who wish to bring about greater levels of IRR should consider the development of detailed scoring rubrics which contain not only the steps that are required by students, but also a description of how to complete these steps.

Another potential method of reducing assessor reliance on performance comparisons is the use of recorded videos in order to facilitate discussion as to what constitutes effective practice. There are multiple strategies that could be adopted in order to implement this. One possibility is the use of scripted videos, in which a panel of expert assessors develop a series of scripts which portray performance at varying levels. For example, in the study by Yeates et al. (2013), the authors devised scripts for a clinical scenario in which a trainee doctor deals with a patient suffering from an unexplained loss of consciousness. Three videos were recorded, one each representing "excellent", "borderline" and "fail" grades. A similar approach for undergraduate nursing OSCEs would allow for assessors (especially those new to the process of assessing OSCEs) to have a visual representation of different practice levels before they begin assessing. As a result, they would be able to "anchor" their interpretations of student performances to these videos. This would likely reduce their reliance on idiosyncratic performance comparisons, which should result in higher levels of IRR. The use of scoring rubrics and video exemplars both have the aim of cultivating a shared understanding between assessors of what effective performance at an assessed skill looks like.

Another notable finding in terms of assessor idiosyncrasy is their reliance on a range of different mitigating factors that are used to account for why a student did not complete a specific step (or series of steps) on an OSCE correctly (section 4.4.3.2). These explanations have to do with an assessor's "educational perspectives" (Roberts et al., 2020, p.9): namely, the extent to which they believe their interpretations of an OSCE performance should be affected by reflexiveness on the part of students (i.e., expressing awareness of potential mistakes), the artificiality of the OSCE environment, and the perceived inexperience of the student. This has potential implications for score reliability, as these perspectives were not shared between all 12 assessors. A student who interacts awkwardly with a mannequin in the NG OSCE would likely be awarded a higher grade by the ten assessors who noted that they would factor in the artificial nature of such interactions when awarding a grade to the student than they would the two who did not mention this. Notably, the use of video seemed to compound this issue in the case of Assessor 11. As documented in section 4.6.3.4, Assessor 11 reported that if an OSCE were taking place in person, she would ask follow up questions once the student had completed the procedure, to determine whether the student was aware of mistakes she had made. The fact that video OSCEs do not allow such an interjection is likely to have accounted for the particularly low grades awarded by Assessor 11 in the present study.

The issue of what institutions such as the School can do regarding such idiosyncrasy depends on the ultimate aim of the OSCE. Based on the quantitative results of the present study (section 4.5), it is suggested that one of the central aims of OSCE development should be to bring about a greater level of score reliability (particularly IRR). If that is the goal, one possibility to bring this about is through a rigorous program of assessor training, which has long been noted as a means by which unwanted score variance can be addressed (Khan et al., 2013b). As noted by Gingerich et al. (2014a), this approach is based on the assumption that score variance arises because assessors do not apply the assessment criteria in the correct way, or make unjustified interventions into the assessment (as was the case in the present study). A program for assessor training would instruct assessors in how to apply the marking guide in a way that is consistent. Such an approach could emphasise the potential negative effects in terms of reliability of, for example, unilaterally asking follow-up questions to students after the OSCE is complete, and awarding students higher grades if they expressed awareness of their mistakes. However, it is not the case that implementing rater training will automatically reduce unwanted score variance, and numerous authors have noted the importance of tailoring assessor training programs to suit institutional needs, as well as evaluating these programs to ensure they are having the desired effect (Preusche et al., 2012; Gingerich et al., 2014a). The results of the present study, therefore, could be used as the basis for a targeted assessor training program within the School. A future analysis of awarded scores could then be conducted to determine whether such a program had the intended effect of increasing reliability between assessors.

Two related issues that were identified in the present study are expertise on the part of assessors (section 4.4.5.2) and the use of inferences to inform assessment decisions (section 4.4.2.2). As documented by Benner (1982), expert nurses are able to quickly identify the salient aspects of a specific situation, and can rely on their own experiences, rather than “abstract principles” (p.402), when making sense of that situation. As noted in the previous chapter, the experienced assessors who participated in the current study often invoked the idea of a “gut instinct” informing their assessments of students, and were much more likely to express confidence that their assessment decisions are rarely incorrect. However, this reliance on expertise may result in discrepancies regarding how students are graded, as two assessors may not share the same “instinct” about a student, and will likely be relying on their own unique experiences. This was particularly relevant to Assessor 11, and was discussed in section 4.6.3.2 as being a potential reason why the grades she awarded in the present study were so low compared to the rest of the sample. Similarly, the use of inferences by assessors has notable implications for reliability, due to the fact that not all assessors made inferences, and that the effects of inferences on awarded scores were inconsistent (i.e., some assessors used inferences to justify awarding students a grade, while some made a similar inference but chose not to award a grade).

The issue of what assessment designers should do regarding the issues of expertise and inferences is a contentious one. Some authors have written that expert assessors should be given scope to exercise subjective judgements about students, as they can use their expertise to identify salient aspects of a student’s performance beyond what can be captured within an OSCE marking guide (Eva & Hodges, 2012). This would allow for a more detailed description of a particular student’s strengths and weaknesses. Similarly, researchers have noted that expert assessors can make “high level” inferences about students that accurately explain why they performed in a certain way within the assessment (Govaerts et al., 2011). Indeed, such inferences were present in the results of the current study, particularly regarding the student in the P03BP video: multiple assessors correctly determined that this student had experience working in clinical practice, and this was why she skipped performing an estimated systolic blood pressure reading. Seen this way, assessor idiosyncrasy can be thought of as “meaningful” (Gingerich et al., 2014a), in that it allows for informed judgements about students. As such, both expertise and inference formation represent a potential tension in OSCE assessment: namely, the need for reliable scores and the validity that results from accurate and detailed decisions about students’ ability levels.

Ultimately, the question of what to do regarding these two related issues again has to do with the purpose of the OSCE. The OSCEs used in the present study are relatively simple ones, used as part of an assessment battery to determine whether students in first- or second-year of a nursing program have demonstrated sufficient mastery of the curriculum in order to progress to the next year study. When OSCEs are used for summative purposes such as these decisions, it is important that scores demonstrate a sufficient level of reliability. Seen this way, assessors leveraging their expertise to make potentially idiosyncratic judgements, or making inferences about students, may threaten the validity of these decisions, as they increase the possibility that a student's score was influenced by the nature of the assessor who happened to be grading them (Kogan et al., 2011; Gingerich et al., 2014a). If the goal of organisations such as the School is to improve IRR in order to make defensible decisions, then assessors should be discouraged from exercising this idiosyncrasy.

However, if the OSCE is used for formative purposes, the value of assessor idiosyncrasy may be increased, particularly when the assessment is conducted through video. Authors such as Lewis et al. (2020) have noted the potential of video OSCEs in terms of formative assessment, as students are able to replay their own performances in order to identify areas for improvement. Assessors in these types of OSCEs may be able to give detailed feedback to students that they can use to improve their skills in the future. Seen this way, the use of "expert judgement" (Eva & Hodges, 2012, p.917) can be useful, as it may result in the student receiving more information about their own performances. In these cases, subjectivity on the part of assessors has positive effects on the validity of the OSCE, as it leads to improved outcomes in terms of student learning. As such, organisations such as the School need to be explicit about the goals of the OSCE and how these will interact with the effects of expertise and inference formation.

5.3.2.3 *Moderation practices*

The findings from the quantitative element in the present study (section 4.5) revealed significant variation in how the four recorded OSCE performances were graded by the 12 assessors. This variation was present for both the checklist and global items, and for three of the videos resulted in a situation where the pass/fail decision was contingent on the assessor. There are three implications of these findings for the School in terms of what to do to account for this variance. These implications are discussed below, in the order they would take place: *standard setting* (before a batch of OSCEs is administered), *cross-moderation* (during the administration of a batch of OSCEs) and *reliability analyses* (after the batch of OSCEs).

Firstly, the current findings suggest the potential utility of adopting a process of standard setting prior to the OSCE. The purpose of standard setting is to determine the grade required in order to pass the OSCE (Khan et al., 2013b). Because nursing OSCEs usually have the ultimate goal of determining whether students are sufficiently competent at performing a specific procedure, the OSCEs are normally criterion-referenced. In a criterion-referenced assessment, a student should receive the same score regardless of the overall level of the cohort of students, based on their completion of a series of criteria (AERA et al., 2014). When administered in the School, the OSCEs used in the present study are criterion-referenced: the decision as to whether a student has passed is based on a simple calculation of the percentage of steps they completed correctly. This has the benefit of simplicity (both for students and assessors); however, in other contexts, OSCEs frequently employ more sophisticated procedures of criterion-referenced standard setting, which take into account what assessors believe to be important within a given skill, and how they conceptualise different levels of performance at that skill, particularly at the pass/fail distinction (e.g., Traynor & Galanouli, 2015). There are multiple different methods of doing this, the most notable being the Angoff procedure. Ultimately, institutions using OSCEs could consider adopting a more complex standard setting procedure: doing so should increase IRR when OSCEs are administered, as assessors' opinions will have been factored into the standard setting process.

Secondly, the results of this study indicate the need to ensure that there are rigorous moderation processes, specifically cross-moderation of performances, in place when OSCEs are used for summative purposes. This is particularly true for OSCEs that are single-station, as a student's score may be determined by a single assessor. As a result, the impact of a particularly harsh or lenient assessor is increased (Rushforth, 2007; Khan et al., 2013b). Participants in this study reported varying levels of moderation in the School: some mentioned a less formal process whereby any potentially failing grades were double-checked by the module coordinator, while others mentioned a systematic procedure in which 20% of all videos were marked by two assessors to ensure consistency. While it is encouraging to see moderation in place, it is considered best practice for institutions using OSCEs to have standardised procedures in place for all OSCEs. The latter approach has been noted in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) as being an acceptable method of maximising consistency between assessors, and ensuring that the grades awarded to test-takers are not contingent on the nature of the assessor.

Thirdly, the high level of score variance found in the present study suggests the need for designers and administrators of new performance assessments to conduct thorough IRR analyses as part of their validity argument. This is a well-documented step when any assessment is set up, and is particularly important when high-stakes decisions are made on the basis of awarded scores; for example, in the field of certification and licensure (NCCA, 2014). These decisions can only be considered valid if the assessment designers can provide evidence that students would have received broadly the same score regardless of what assessor(s) was judging them (Fraenkel & Wallen, 2006).

Additionally, this study highlights the need to conduct reliability analyses on performance assessments that have already been in place for several years. The two OSCEs used in this study have been in place (in some form) for over a decade. As noted by numerous researchers writing about score reliability (e.g., Thompson & Vacha-Haase, 2017), reliability is not a property that an assessment "has" indefinitely; rather, each new set of scores produced should be examined to determine whether there are threats to reliability. While this may not always be possible given the resource and time constraints inherent in nursing departments (Solà-Pola et al., 2020), it is nonetheless considered best practice, and should be adhered to whenever possible.

The present findings suggest two potential approaches for such an analysis. The first approach is to use descriptive statistics to identify outlier assessors, namely those that consistently award high or low scores compared to the general cohort of assessors (as documented in section 4.6.2 of the previous chapter). As detailed by Bartman et al. (2013), such assessors pose a particular threat to IRR. The method used to detect outliers will depend on the specific cohort of assessors and the needs of the institution administering the OSCE: as noted by Kardong-Edgren et al. (2017), the issue of identifying and accounting for outlier assessors has not received significant attention in the nursing assessment literature. As such, institutions such as the School have to determine what approach is best for them. One relatively simple procedure that has been described by Bartman et al. (2013) is to identify assessors whose mean scores are more than three standard deviations above or below the mean of the overall sample of assessors. However, they note that this figure of three is arbitrary, and could be adjusted based on the situation. If such an approach was applied to the small cohort of 12 assessors who participated in the present study, two assessors (Assessors 8 and 11) would have been identified as awarding checklist items more than one standard deviation below the mean, while a single assessor (Assessor 3) awarded scores more than one standard deviation above the mean.

Subsequent to this identification, the next step would be to determine what to do with these outlier assessors. An obvious starting point would be for a sample of these assessors' graded performances to be marked again by a different assessor (preferably one whose awarded scores were close to the mean), particularly any performances that were awarded a failing grade. If there were notable discrepancies in how the new assessor graded the performances compared to the original assessor, they could discuss the performance together and decide on a final grade. While some authors suggest simply removing outlier assessors from the assessor pool once they have been identified (e.g., Kardong-Edgren et al., 2017), this may not be feasible given the limitation in terms of available staff in institutions such as the School. If this is the case, these outlier assessors could be given targeted training or professional development in order to better align their grades with those of their colleagues for future OSCE administrations (Dunbar, 2018).

A second approach for IRR analysis concerns the identification of checklist items which cause particular problems in terms of IRR. As documented in section 4.5.5 of the previous chapter, six items across the two OSCEs used in the present study resulted in a percent agreement figure of 0.75 or below, which is classed as “unacceptable” according to Gwet (2014). Once these items have been identified, OSCE designers can then facilitate targeted training or discussions with assessors in order to bring about greater alignment in how these items are scored by assessors. For example, the lowest IRR reported in the present study was item 14 on the NG OSCE, *clearly states patient comfort is offered*. Assessors could be asked to talk about what they expect of students in order to complete this item satisfactorily. Such a discussion should improve IRR outcomes the next time the OSCE is administered.

5.3.2.4 *Use of video in OSCE assessment*

Findings from the present study have implications regarding the use of video in performance assessments such as the OSCE, both within the School and beyond. As noted by authors such as Lara et al. (2020), the incorporation of video into these assessments had taken place before 2020, but was rapidly accelerated due to the onset of the COVID-19 pandemic in March of that year, which precluded many educational institutions from conducting in-person assessments. The two OSCEs which were used in the present study (blood pressure measurement and naso-gastric tube insertion) had been adapted for remote administration by module coordinators in the School, and were conducted remotely during the pandemic (and, in the case of the BP OSCE, before the pandemic). However, as with all assessments, it is important to consider the effects on reliability and validity when the modality of the assessment is changed (AERA et al., 2014).

Assessor participants' discussions of the use of video OSCEs suggest potential issues regarding the reliability of awarded scores. Many participants were positive about the use of video in terms of scoring consistency, specifically mentioning the ability to pause and rewind videos as preventing them from having to guess as to whether a student had completed a step on the guide correctly (section 4.4.5.3). However, assessors also expressed that they had difficulty seeing some aspects of the recorded performances on video (section 4.4.2.3). As discussed in the outlier analysis, different assessors had different strategies for what to do when something was not clearly visible to them: some would fail to award a mark for an item that they could not see clearly, while others would choose to give students the benefit of the doubt and award the mark. As such, while assessors were almost uniformly positive about the effects on reliability of the use of video, a deeper analysis of the data revealed potential negative effects on IRR.

These findings suggest the need for institutions such as the School to conduct reliability analyses when assessments such as the OSCE are moved online. Researchers writing about video OSCEs have emphasised that score reliability cannot be assumed to correlate between face-to-face and video OSCEs (Lara et al., 2020), and that the assumption of adequate levels of IRR in a remote OSCE requires the collection of evidence. One such method is that described by Dagnaes-Hansen et al. (2018), who recruited an assessor to grade a series of cystoscopies in-person, which were also recorded. A month later, the same assessor graded the same series of recorded performances, and the scores were found to demonstrate high levels of intrarater reliability, indicating that the scores awarded between the two assessment modalities (in-person and video) were comparable. Such an approach represents the "gold standard" in collecting evidence vis-a-vis the reliability of awarded scores when OSCEs are administered through video. While this method may not be feasible for nursing faculties such as the School, the current findings nonetheless indicate that some kind of reliability check should be carried out to determine the effects of video OSCE administration on IRR.

The results of the present study also speak to the validity implications of administering video OSCEs. Two assessors discussed the potential negative consequences of conducting video OSCEs, mentioning the fact that students can film themselves completing the assessed skill a number of times until they are happy with the result (section 4.4.5.3). While this process is likely to have positive effects on skill acquisition, it reduces the realism of the assessment, as students may not have as many attempts to complete the same skill in the clinical setting. It has long been noted that a key advantage of the OSCE in terms of validity is the fidelity of the exam environment to the real world (Hodges, 2003). Seen this way, the use of video OSCEs may undermine validity, and it is incumbent upon OSCE designers to consider the validity effects of video OSCEs. Due to the COVID-19 pandemic, such an approach was likely the only feasible way of conducting any practical assessment of students' skills (Major et al., 2020). However, the return to in-person teaching and assessment provides an opportunity for nursing faculties to consider the use of video OSCEs, and their place within the broader battery of assessments.

5.4 *Theoretical implications*

This research has responded to calls for continued investigation into the cognitive processes undergone by assessors of performance assessments in the health sciences as they judge student performances (Boursicot et al., 2021), as well as score reliability in nursing OSCEs more generally (Goh et al., 2019). As such, the present mixed methods study explored the issue of how assessors make judgements about students, and the results of these decisions in terms of awarded scores. In addition to the practical implications discussed above, this study has several theoretical implications for the field.

Firstly, the current research is among the first to unpack the various factors which influence assessment decisions at the observation, processing and integration stages of decision-making (Gauthier et al., 2016). In the traditional psychometric view of assessment, raters should be trainable to perceive the same situation in the same way, and any variance between two raters assessing the same thing is counted as error (Gingerich et al., 2014a). However, researchers in medical education and (to a lesser extent) nurse education have problematized this view of assessors, noting that because all assessors are individuals, they bring their own unique set of knowledge and experiences into the assessment environment (Mitchell et al., 2009; Govaerts et al., 2013). Seen this way, assessor variance is no surprise, as it would be unrealistic to expect multiple assessors to be "blank slates" who can leave aside their own opinions and ideas in order to assess students in an objective way.

The present study has demonstrated that (at least in the context in which the study took place), nursing assessors do indeed bring with them a range of individualised judgement mechanisms which influence their interpretations of student performances. While assessors did note that they were trying to judge students against the fixed criteria of the guide, the findings also indicated that they used additional, exogenous criteria when judging students, and lacked a unified sense of what "good" performance entailed (Yeates et al., 2013). Because of this, they ended up interpreting the recorded videos in a number of ways, influenced by their own idiosyncratic judgement processes.

Given the variation in how assessors in this study approached the task of assessment, it is perhaps unsurprising that there was a significant amount of variance recorded in terms of the scores awarded for the recorded OSCE performances. As such, the present study has indicated that existing models of assessor decision-making (e.g., Kogan et al., 2011; Gauthier et al., 2016), underpinned by a social cognitive approach (Greifeneder et al., 2018), are useful in explaining how nursing assessors "interpret and construct their own personal reality of the assessment context" (Govaerts & van der Vleuten, 2013, p.1169). Given the increasing popularity of OSCEs in undergraduate nursing programmes worldwide (Goh et al., 2019), it is expected that the issue of assessor cognition will only become more important, and researchers should seek to replicate the current study in different contexts.

Secondly, this study has demonstrated the usefulness of adopting a mixed-methods approach to investigating the issue of inter-rater reliability. Large-scale studies of reliability in both medical and nursing assessment have tended to use quantitative methods in order to determine whether there is evidence that decisions made on the basis of awarded scores can be deemed valid (Brannick et al., 2011; Navas-Ferrar et al., 2017). At the same time, initial studies into assessors' judgement processes have used a qualitative approach, to dig deep into the "black box" (Kogan et al., 2011) of decision-making. The present study is one of a handful of studies which have sought to combine these approaches, in order to "link" assessors' decisions to measurable outcomes in terms of awarded scores (Gingerich et al., 2014b; 2017; Roberts et al., 2021). Although the results of the present study are not sufficient to conclude any definite link between processes and outcomes, the data analysis nonetheless identified notable differences in how the harshest and most lenient assessors in the present sample approached the task of OSCE assessment, and how these differences led to variance in the scores they awarded. As such, this study has exemplified the need to centre assessors themselves in studies about IRR, in order to gain a better understanding of how score variance between assessors arises.

Thirdly, the present study adds to the ongoing debate about the role of subjectivity in assessment. As noted by Hodges (2013), subjectivity was long seen as a pejorative concept in performance assessments, equated with bias and unfairness. Researchers working in the field of medical education have written of the need to "rehabilitate" subjectivity, noting that incorporating some level of subjectivity into assessment might allow for experienced assessors to identify and reward students if they display talent or flair that goes beyond what can be captured in the marking guide (Govaerts & van der Vleuten, 2013). This worry about extreme objectivity has been echoed in the nursing literature, with researchers expressing concerns that the aim of standardising all aspects of an assessment (a central goal of the OSCE) removes the complexity and context-specificity that is central to nursing practice (Mitchell et al., 2009). Seen this way, an over-emphasis on reliability actually threatens, rather than strengthens, the validity of the OSCE (Rushforth, 2007).

The results of this study are unlikely to be interpreted as an endorsement of introducing more subjectivity into the assessment process, at least at undergraduate level. While the idea that assessors should be allowed to exercise some level of autonomy with regard to their decisions is a reasonable one, the wide variation in scores awarded to the same four videos in this study indicate that there is perhaps too much subjectivity at play. Indeed, the considerable variation in scores suggests that there may be a notable threat to the validity of decisions made on the basis of OSCE scores. If the findings from this study were to be replicated elsewhere, it would suggest that undergraduate nursing OSCE assessors need to become more objective in their judgements, rather than given more autonomy regarding their decisions. However, as noted in section 5.3, the extent to which subjectivity and objectivity have a place in OSCE assessment is dependent on the ultimate goal of the OSCE.

5.5 *Strengths and limitations*

There are several strengths to be noted regarding this research study. Firstly, at the time of writing, this study is among the first to investigate the cognitive processes of undergraduate nursing OSCE assessors. Previous work in assessor cognition has taken place largely in the field of medicine (e.g., Kogan et al., 2011). When this project was conceived, it was unclear whether the mechanisms that nursing assessors employ when judging students are the same as those employed by their medical counterparts. The results of this study suggest that there is indeed a large overlap between these two groups.

This study is distinguished from the relatively small number of similar studies which have taken place in the field of nurse education. A study by East et al. (2014) documented how nursing assessors form judgements about students; however, their work employed exclusively a semi-structured interview approach. Participants in their study were therefore interviewed in the abstract about how they approach the task of assessment. In the present study, the use of the think aloud protocol allowed for a determination of how assessors actually grade students in practice. Additionally, the mixed methods approach employed here allowed for an exploration of the link between how assessors form judgements and the quantitative results of these judgements. This is distinct from previous work which has largely adopted either a qualitative (East et al., 2014) or quantitative (Dunbar, 2018) approach to the issue of assessor consistency.

Finally, this study benefitted from the use of a strong framework of assessors' decision-making, based on a paper by Gauthier et al. (2016). The use of a framework has long been urged by researchers working in the field of assessor cognition (Gingerich et al., 2011). Using this framework allowed the researcher to isolate the different stages of decision-making, and identify the cognitive processes involved at each stage. As noted by Gauthier et al. (2016), future research in assessor cognition can use this framework to devise targeted studies (e.g., focusing specifically on the *observation* stage of decision-making, to determine whether nursing assessors in other contexts are prone to making inferences about students).

However, there are also several limitations of this study that need to be addressed. The first is to do with the size and specificity of the sample. Clearly, a sample of 12 assessors working in the same institution is insufficient to allow the generalisation of results beyond this setting. Future research should seek to use the same decision-making framework applied in this study in order to investigate whether the cognitive processes employed by these participants are replicated in other cohorts. Findings from the field of medical assessment have noted a large degree of overlap in how assessors in different environments make decisions about students (Kogan et al., 2011; St-Onge et al., 2016); a clear next step for nurse education researchers is to determine whether the same overlap is present across undergraduate nursing programmes.

Secondly, the OSCEs used in this study are relatively "simple" ones, administered to nursing students in their first or second year of study. Previous research in assessor cognition has tended to use more complex, multi-station OSCEs (or other performance assessments) in order to investigate the issue of assessor cognition (e.g., Yeates et al., 2013, Gingerich et al., 2014a). While the use of first and second-year OSCEs in this study is relatively novel from a research standpoint, it is unclear from this study whether assessors engage with the task of assessment differently when they are judging more advanced skills. Investigating assessors' cognitive processes as they grade a more complex OSCE would allow for a meaningful comparison between these two conditions.

Finally, it is noted that the constraints faced by the researcher in terms of the limitation on how long each interview could be (DCU guidelines stipulate they should be approximately one hour), meant that each participant only judged four videos. While this was enough to determine variation in how they graded these videos, it did not allow for the calculation of more complex measures of IRR such as Cohen's kappa (Gwet, 2014) or Intraclass Correlation Coefficient (Koo & Li, 2016). These measures of reliability are generally stronger when *less* assessors rate *more* cases; in this study, the opposite approach was employed. As such, the comparability of these findings vis-a-vis other studies (e.g., Cazzell & Howe, 2012) is limited. Future research could use the same methodology applied in this study in order to get less raters to watch more videos.

5.6 *Recommendations for future research*

There are four notable directions for future research on the basis of the present study, informed in part by the limitations noted in the previous section. Firstly, researchers should focus on attempting to replicate the findings of the current study in other institutional contexts. The exploratory, case study approach taken here does not allow for results to be generalised beyond the context of the School. As such, it remains unknown whether assessors outside of this context are affected by the same factors when making decisions about student performances. Systematic reviews of assessor cognition research in medical assessment have noted the presence of similar judgement patterns shared by assessors across different institutions (e.g., St-Onge et al., 2016). It is entirely possible that nursing assessors in other universities in Ireland (or indeed, outside the country) share similar judgement processes with those who took part in this study; however, this will remain unknown unless explicitly investigated.

Secondly, it is impossible to ascertain from the present study whether the judgement processes demonstrated by assessors are specific to the two OSCEs which were included. Nursing researchers have long noted the context-specificity of effective nursing practice (Mitchell et al., 2009). The task specificity of assessors' judgement formation has been documented by researchers in medical assessment (Govaerts et al., 2013; Gauthier et al., 2016). It is possible that a skill such as blood pressure measurement is liable to cause assessors to (for example) make inferences in a way that other skills would not. Indeed, as noted previously, other studies in assessor cognition have often used more complex skills in order to look at the issue of assessor decision-making (e.g., Gingerich et al., 2014b). Future research should seek to investigate assessors' cognitive processes using different OSCEs, in order to determine whether the same decision-making strategies are employed by assessors.

Thirdly, future research should seek to employ larger sample sizes, so that more complex calculations of IRR can be made. The present study lacked the statistical power to calculate measures of IRR such as Intraclass Correlation Coefficient and Cohen's kappa, using instead the relatively crude calculation of percent agreement. As such, it is difficult to situate the work in the context of other studies of IRR in nursing OSCEs, such as that by Cazzell and Howe (2012), who only had two assessors participate in their work, but had them watch 207 videos each. The present study, with its use of 12 assessors, allowed for a capturing of the breadth of assessors' decision-making processes; however, given that each assessor only watched four videos, it did not allow for a "deep" calculation of IRR.

Finally, researchers should seek to build on the final phase of analysis in this study, in which an attempt was made to determine whether assessors who awarded similar scores as each other employed similar decision-making processes to reach those scores. Two studies by Gingerich and colleagues (2014; 2017), used ANOVA to determine the extent of score variance that can be explained due to an assessor belonging to a specific "cluster", while a study by Chahine et al. (2016) used Ordinal Logistic Hierarchical Linear Modelling to investigate the same issue. Future research in nurse education (with larger sample sizes) could employ a similar approach, by using methods such as Latent Partition Analysis to divide assessors into clearly defined groups, and calculating the effect on score variance of belonging to a specific group. Such an approach would go beyond what was feasible in this study.

5.7 *Conclusion*

5.7.1 *Making summative OSCEs fit for purpose*

Since their inception in 1975, OSCEs have become popular in undergraduate nursing programs, and are currently performed in over 30 countries worldwide (Goh et al., 2019). Given the documented use of OSCEs as a mode of summative assessment, in which students' scores are used to make high-stakes decisions (Rushforth, 2007), it is crucial that those developing and administering OSCEs provide evidence pertaining to the validity of these decisions. Only when this evidence is documented can the decisions be deemed defensible. As OSCEs require the use of human assessors to evaluate student performances, a key source of validity evidence is the consistency, or inter-rater reliability (Gwet, 2014), of assessors' judgements. However, the outcomes of the study presented here indicate that there may be notable threats to this consistency. As such, this study has implications for those using OSCEs as a mode of summative assessment. Based on the findings, there are a number of steps that stakeholders can take to ensure that summative OSCEs are fit for purpose.

Firstly, there should be significant thought put into the content and structure of the marking guides used to assess student performances. Results from the present study consistently indicated that threats to IRR arose due to assessors judging students on the basis of idiosyncratic criteria that were not explicitly contained within the guides. As such, it can be determined that IRR may be reduced if the content of marking guides does not align with what assessors believe to be important (Hyde et al., 2021). Additionally, the present study found that, of the criteria which are included in the guide, assessors had differing ideas as to which were the most important. Several assessors noted that there were some elements within the guides that should result in an automatic fail if not completed correctly, and suggested that the weighting of marks within the guide should reflect the relative importance of the different criteria (Chahine et al.. 2016). On the basis of these results, it is suggested that marking guides should be re-conceptualised with active input from the assessors who will be using them. Although this still represents a subjective decision (ten Cate & Regehr, 2019), it should nonetheless reduce instances of disagreement between assessors, as some of the nuances in how they form judgements will be captured within the guide *prima facie*.

Secondly, the present study highlights the importance of conducting thorough reliability analyses when an OSCE is designed and administered at undergraduate level. Assessment researchers (AERA et al., 2014) as well as those writing about the development of OSCEs specifically (Khan et al., 2013b), have long noted the necessity of this step; however, the results of this thesis suggest that it is being overlooked, at least within the context where the study took place. This is unsurprising given the significant constraints often faced by undergraduate nursing departments in terms of resources and staffing (Solà-Pola et al., 2020). Nonetheless, the considerable variance in awarded scores documented in this thesis is alarming from an assessment perspective, particularly given the fact that the OSCEs used in the present study are administered for summative purposes. As such, it is strongly advised that those involved in administering undergraduate nursing OSCEs factor in the time and money needed to conduct this vital step. As discussed throughout this chapter, the use of video may provide a convenient method for cross-checking assessors' awarded grades and identifying notably harsh or lenient assessors.

Finally, the present study calls into question the usefulness of the OSCE as a mode of summative assessment unless significant steps can be taken to ensure scoring consistency. If it was replicated across a general administration of an OSCE (for example, across an entire year-group of students), the level of divergence recorded in the present study would likely preclude defensible decisions being made on the basis of OSCE scores. However, this does not mean that the OSCE should be abandoned as a method of assessment entirely. Indeed, numerous authors have noted the potential for OSCEs to be used as a method of formative assessment, with assessors' comments about student performances used as the basis for students to identify potential areas where they can improve their practice (e.g., Rushforth, 2007). This is particularly relevant in the context of video OSCEs, as students can watch and review their own performances, and approached described in a paper by Lewis et al. (2020). As such, if organisations do not have the resources to document and improve IRR, they should consider whether OSCEs are an acceptable mode of summative assessment, and whether they might better be used for formative purposes only.

5.7.2 *Implications beyond the healthcare sector*

Although the present study was conducted with undergraduate nursing OSCEs, the results have implications for the performance assessment landscape beyond the healthcare sector. Human judgement of performance is a feature of many different assessments across the education sector more broadly. For example, recent research has investigated the use of "multiple mini-interviews", in which test-takers undergo a circuit of short interviews with the aim of evaluating their readiness to enter teacher education programs (Klassen & Kim, 2021). This assessment format has clear parallels with the classic multi-station OSCE format, and as such a key consideration is how to ensure consistency across assessors, so that assessment scores have adequate levels of IRR. Notwithstanding the concept of task specificity that has recurred throughout this thesis (Govaerts et al., 2013), results from the present study can be used to inform the following set of principles for the design and administration of performance assessments in education.

1. Recognition of the complexity of human cognition

The present study has demonstrated the range of often subjective and idiosyncratic factors that affect how assessors form judgements about test-taker performances, and the potential threats to reliability that may develop as a result. From a social cognitive point of view, the presence of such factors in the assessment context is unsurprising, as they relate to the same cognitive mechanisms that individuals use when navigating the world more generally (Gilbert, 1998). Indeed, there is an increasing recognition in the performance assessment literature that the complete removal of assessor idiosyncrasy is neither possible nor necessarily desirable (Gingerich et al., 2014a). As such, those designing performance assessments should begin by recognising the complexity of human cognition and its effects on the assessment process. Such an approach would allow for the development of assessments that have the goal of ensuring greater alignment between the demands of the assessment on assessors and how they process information (Tavares & Eva, 2013). If done effectively, this should reduce instances of assessor inconsistency when the test is administered.

2. Strong assessment blueprinting

Best practice for assessment development posits that one of the starting points for the design of any assessment is the decisions that will be made on the basis of assessment scores (AERA et al., 2014). These decisions influence all stages of the development of the assessment, particularly the preparation stage where assessors are given instructions and training. The results of the present study indicated that assessors seemed to be unclear as to the ultimate goal of the OSCEs that were used: although the OSCEs are ostensibly implemented for summative purposes, some participants spoke about the OSCE in formative terms, mentioning that the aim of the OSCE was to allow students to develop their skills. This led to idiosyncrasy in how assessors engaged with the task of assessment, with some reporting that they would make interventions to the assessment (such as allowing students to begin again if they were nervous) that had the likely effect of reducing IRR. As such, developing a strong assessment blueprint, which foregrounds the purpose of the assessment, and communicating this to assessors, should be the starting point for the development of any performance assessment (Khan et al., 2013b). This process should reduce instances of idiosyncrasy, and hopefully bring about greater agreement between assessors.

3. Re-thinking assessment content

The present study indicated that assessors were liable to judge students against holistic criteria that were not explicitly contained within the scoring guides, based on what they perceived to be important. As discussed in the previous chapter, this is likely to have led to divergences in the scores they awarded. Recent research from Hyde et al. (2020) noted that, when asked to devise a new marking guide, assessors' ideas about what should be contained within the guide were different from the guides that were already in use: assessors preferred the inclusion of holistic performance domains such as safety and fitness to practice. This is in line with research from both social cognition (e.g., Gilbert, 1998) and assessor cognition (Yeates et al., 2013) which indicates that people process behaviour in holistic terms, Consequently, assessments that are constructed to assess a series of discrete binary items (as was the case with the OSCEs used in the present study) might be inherently incongruous with human cognition, and may lead to divergence between assessors. To that end, re-thinking how marking guides are constructed and structured, with the potential to incorporate holistic competencies which assessors believe to be important, has great potential as a way of mitigating assessor inconsistency.

4. Assessor moderation practices

The impact of outlier assessors on IRR was particularly striking in the current study. The findings suggest that experienced assessors may have developed ways of assessing students that lead to notable divergences in the scores they award compared to those of their less-experienced counterparts. As discussed throughout the thesis, this is an issue of validity as well as reliability, as these assessors may have unique insights and experiences which affect how they grade student performances (Hodges, 2013). However, if the goal of a performance assessment is to make summative decisions, this idiosyncrasy poses a distinct threat to reliability. One method of addressing this is for quantitative score data to be used to calculate whether certain assessors are consistently “hawks” (harsh) or “doves” (lenient) (e.g., Dunbar, 2018). When a performance assessment is delivered every year, there may be an assumption that experienced assessors are less likely to judge performances in an idiosyncratic way, and therefore do not pose a threat to IRR. However, this study indicates that this may not be the case. As a consequence, it is imperative that that routine reliability checks be carried out, so that outlier assessors can be identified and the relevant adjustments made to the assessment process.

5. Assessor training/professional development

As discussed above, the findings of the present study suggest that at least some variance between assessors is caused by assessors acting in ways that are at odds with the summative goals of the assessment. An example of this is when assessors perceive that students are nervous and allow them to restart an OSCE if they do not do well initially. With this in mind, it should be possible to improve IRR through a well-designed and rigorous process of assessor training (Gingerich et al., 2014a). Such a process could use both qualitative and quantitative data (following the approach described in this thesis) in order to specify potential areas for improvement. As noted by researchers writing about assessor training programs (e.g., Preusche et al., 2012), there is no one-size-fits-all approach, and programs are more likely to be successful if they are tailored to the specific needs of an institution. Therefore, those developing performance assessments should consider developing iterative training programs for assessors, which incorporate new issues as they emerge from each administration of the assessment. Such a flexible approach would ensure that training and professional development remain relevant to the assessment, which should reduce instances of variance between assessors.

5.8 *Epilogue*

The importance of effective assessments in contributing to the development of healthcare professionals has long been recognised (Harden et al., 1975; Rushforth, 2007). While the OSCE has long been touted as the “gold standard” of undergraduate assessment, the present study indicates that the involvement of human assessors in the process may cause issues in terms of reliability and validity. However, during a time when the influence of technology and Artificial Intelligence in assessment is increasing (O’Leary et al., 2018), the results of this study should not be interpreted as a call for the removal of people from the assessment process. Instead, the identification of potential issues regarding the use of human assessors can be used to “sharpen the people” involved in OSCE assessment (van der Vleuten et al., 2010, p.712).

The present study is small in scale, local in context, and somewhat theoretical, yet it speaks to a very significant issue: the education of nurses who care for people when they are at their most vulnerable. It is incumbent upon everyone involved in undergraduate nursing assessment to ensure that student nurses are afforded the maximum possible opportunity to develop into effective clinical practitioners. This study took place against the backdrop of the COVID-19 pandemic, a time during which almost all organisations had to re-think their assessment practices. This period of change has been difficult for everyone, but has also provided an opportunity to “re-assess” what we want from our assessments, and what we owe to test-takers. As stakeholders across the testing world – universities, credentialing organisations, professional bodies - return to in-person assessment provision, the challenge of developing and maintaining rigorous performance assessments is as relevant as ever. This study provides food for thought for anyone wishing to harness the power of human assessors in performance assessment contexts while at the same time maximizing fairness for all students

Reference List

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards For Educational And Psychological Testing*.

Anim-Boamah, O., Christmals, C. & Armstrong, S. (2021). Clinical nursing competency assessment: a scoping review. *Frontiers of Nursing, 8*(4), 341-356. https://doi.org/10.2478/fon-2021-0034

Bandalos, D. (2018). *Measurement Theory and Applications for the Social Sciences.* New York, USA: Guilford Press.

Bartman, I., Smee, S. & Roy, M. (2013). A method for identifying extreme OSCE examiners. *Clinical Teacher, 10*(1), 27-31. https://doi.org/10.1111/j.1743-498x.2012.00607.x

Benner, P. (1982). From Novice to Expert. *American Journal of Nursing, 82,* 402-407.

Blythe, J., Patel, N. S. A., Spiring, W., Easton, G., Evans, D., Meskevicius-Sadler, E., Noshib, H., & Gordon, H. (2021). Undertaking a high stakes virtual OSCE ("VOSCE") during Covid-19. *BMC Medical Education*, *21*(1), 1–7. https://doi.org/10.1186/s12909-021-02660-5

Bouriscot, K. & Roberts, T. (2005). How to set up an OSCE. *The Clinical Teacher, 2*(1), 16-20. https://doi.org/10.1111/j.1743-498X.2005.00053.x

Boursicot, K., Kemp, S., Wilkinson, T., Findyartini, A., Canning, C., Cilliers, F., & Fuller, R. (2021). Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. *Medical Teacher*, 43(1), 58–67. https://doi.org/10.1080/0142159X.2020.1830052

Bowen, G.A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education*, *45*(12), 1181–1189. https://doi.org/10.1111/j.1365-2923.2011.04075.x

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

Braun, V. & Clarke, V. (2021). *Thematic Analysis: A Practical Guide.* Thousand Oaks, CA: Sage Publications.

Cant, R., McKenna, L., & Cooper, S. (2013). Assessing preregistration nursing students' clinical competence: A systematic review of objective measures. *International Journal of Nursing Practice, 19*(2), 163-176. https://doi.org/10.1111/ijn.12053

Castleberry, A., & Nolen, A. (2018). Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching & Learning, 10*(6), 807–815. https://doi.org/10.1016/j.cptl.2018.03.019

Cazzell, M., & Howe, C. (2012). Using Objective Structured Clinical Evaluation for Simulation Evaluation: Checklist Considerations for Interrater Reliability. *Clinical Simulation in Nursing, 8*(6), 219-225. https://doi.org/10.1016/j.ecns.2011.10.004

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

Chong, L., Taylor, S., Haywood, M., Adelstein, B. A., & Shulruf, B. (2017). The sights and insights of examiners in objective structured clinical examinations. Journal of Educational Evaluation for Health Professions, *14*, 34. https://doi.org/10.3352/jeehp.2017.14.34

Cömert, M., Zill, J. M., Christalle, E., Dirmaier, J., Härter, M., & Scholl, I. (2016). Assessing communication skills of medical students in Objective Structured Clinical Examinations (OSCE) - A systematic review of rating scales. *PLoSONE, 11*(3), e01527171. https://doi.org/10.1371/journal.pone.0152717

Cope D. G. (2015). Case study research methodology in nursing research. *Oncology Nursing Forum, 42*(6), 681–682. https://doi.org/10.1188/15.onf.681-682

Corcoran, A.M., Lysaght, S., LaMarra, D. & Ersek, M. (2013). Pilot test of a three-station palliative care Observed Structured Clinical Examination for multidisciplinary trainees. *Journal of Nursing Education, 52*(5), 294-298. https://doi.org/10.3928/01484834-20130328-02

Dagnaes-Hansen J., Mahmood O., Bube S., Bjerrum, F., Subhi, y., Rohrsted, M. & Konge, L. (2018). Direct observation vs. video-based assessment in flexible cystoscopy. *Journal of Surgical Education*, *75*(3), 671–677. https://doi.org/10.1016/j.jsurg.2017.10.005

Darling-Hammond, L. & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford Center for Opportunity Policy in Education. Available at: https://globaled.gse.harvard.edu/files/geii/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning-report_0.pdf

Denney, M., Freeman, A., & Wakefor, R. (2013). MRCGP CSA: Are the examiners biased, favouring their own by sex, ethnicity and degree source? *British Journal of General Practice,* 63(616), 718-725. https://doi.org/10.3399/bjgp13x674396

Doorenbos A. Z. (2014). ixed methods in nursing research: An overview and practical examples. *Kango kenkyu: The Japanese journal of nursing research*, *47*(3), 207–217.

Dougherty, L., Lister, S., & West-Oram, A. (Eds.). (2015). *The Royal Marsden Manual of Clinical Nursing Procedures*. John Wiley & Sons.

Doyle, L., Brady, A.M., & Byrne, G. (2009). An overview of mixed methods research. *Journal of Research in Nursing, 14*(2), 175–185. https://doi.org/10.1177/1744987108093962

Dublin City University (DCU). (2019). Research ethics [webpage]. https://www.dcu.ie/researchsupport/researchethics.shtml.

Dunbar, S. S. S. (2018). Consistency in grading clinical skills. *Nurse Education in Practice, 31*, 136–142. https://doi.org/10.1016/j.nepr.2018.05.013

East, L., Peters, K., Halcomb, E., Raymond, D., & Salamonson, Y. (2014). Evaluating Objective Structured Clinical Assessment (OSCA) in undergraduate nursing. *Nurse Education in Practice, 14*(5), 461–467. https://doi.org/10.1016/j.nepr.2014.03.005

Edwards, A. L. (1951). Balanced Latin-square designs in psychological research. *The American Journal of Psychology, 64*, 598–603.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://psycnet.apa.org/doi/10.1037/0033-295X.87.3.215

Eva, K., & D Hodges, B. (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Medical Education*, 46(9), 914–919. https://doi.org/10.1111/j.1365-2923.2012.04310.x

Fraenkel, J.R. & Wallen, N.E. (2006) *How to Design and Evaluate Research in Education.* New York, USA: McGraw Hill.

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: Review and integration of research findings. *Medical Education*, *50*(5), 511–522. https://doi.org/10.1111/medu.12973

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 89–150). McGraw-Hill.

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-Based Assessments as Social Judgements: Rethinking the Etiology of Rater Errors. *Academic Medicine*, 86(10), 1–7. https://doi.org/10.1097/acm.0b013e31822a6cf8

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014a). Seeing the “black box” differently: Assessor cognition from three research perspectives. *Medical Education*, 48(11), 1055–1068. https://doi.org/10.1111/medu.12546

Gingerich, A., Van Der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2014b). More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings. *Academic Medicine*, 89(11), 1510–1519. https://doi.org/10.1097/acm.0000000000000486

Gingerich, A., Ramlo, S. E., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education*, *22*(4), 819–838. https://doi.org/10.1007/s10459-016-9711-8

Goh, H. S., Zhang, H., Lee, C. N., Wu, X. V., & Wang, W. (2019). Value of Nursing Objective Structured Clinical Examinations: A Scoping Review. *Nurse educator, 44*(5), E1–E6. https://doi.org/10.1097/nne.0000000000000620

Goh, H.S., Ng, E., Mun L., Hui, Z. & Sok, Y.L. (2022). Psychometric testing and cost of a five-station OSCE for newly graduated nurses, *Nurse Education Today*, *112*., 105326. https://doi.org/10.1016/j.nedt.2022.105326

Govaerts, M., Schuwirth, L. W., Van der Vleuten, C. P., & Muijtjens, A. M. (2011). Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education: Theory and Practice, 16*(2), 151–165. https://doi.org/10.1007%2Fs10459-010-9250-7

Govaerts, M., Van de Wiel, M., Schuwirth, L. (2013). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Science Education*, *18*, 375–396. https://doi.org/10.1007%2Fs10459-012-9376-x

Govaerts, M., & van der Vleuten, C. P. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education*, *47*(12), 1164–1174. https://doi.org/10.1111/medu.12289

Greifeneder, R., Bless, H. & Fiedler, K. (2018). *Social Cognition: How Individuals Construct Social Reality*. London: Psychology Press.

Gwet, K.L. (2014). *Handbook on Inter-Rater Reliability.* Gaithersburg, MD, USA: Advanced Analytics.

Hamilton, D. & Carlston, H. (2013). The Emergence of Social Cognition. In D. E. Carlston (Ed.), *The Oxford Handbook of Social Cognition* (pp. 16–32). New York: Oxford University Press.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, *1*(5955), 447–451. https://doi.org/10.1136%2Fbmj.1.5955.447

Henderson, A., Nulty, D., Mitchell, M. L, Jeffrey, C. A, Kelly, M., Groves, M., Glover, P. & Knight, S. (2013). An implementation framework for using OSCEs in nursing curricula. *Nurse Education Today*, 33(12), 1459-1461. https://doi.org/10.1016/j.nedt.2013.04.008

Hodges, B. (2003). Validity and the OSCE. *Medical Teacher*, 25(3), 250-254. https://doi.org/10.1080/01421590310001002836

Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher, 35*(7), 564–568. https://doi.org/10.3109/0142159X.2013.789134

Howard, E. (2020). A review of the literature concerning anxiety for educational assessments. UK: Ofqual. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/865832/A_review_of_the_literature_concerning_anxiety_for_educational_assessment.pdf

Hyde, C., Yardley, S., Lefroy, J., Gay, S., & McKinley, R.K. (2020). Clinical assessors' working conceptualisations of undergraduate consultation skills: a framework analysis of how assessors make expert judgements in practice. *Advances in Health Science Education.* https://doi.org/10.1007/s10459-020-09960-3

Hyde, S., Fessey, C., Boursicot, K., MacKenzie, R. & McGrath, D. (2022). OSCE rater cognition – an international multi-centre qualitative study. *BMC Medical Education, 22*(6). https://doi.org/10.1186/s12909-021-03077-w

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher, 33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kahneman, D. (2011). *Thinking Fast and Slow*. New York: Farrar, Straus & Giroux.

Kardong-Edgren, S., Oermann, M. H., Rizzolo, M. A., & Odom-Maryon, T. (2017). Establishing inter- and intrarater reliability for high-stakes testing using simulation. *Nursing Education Perspectives*, *38*(2), 63–68. https://doi.org/10.1097/01.nep.0000000000000114

Khan, H. (2017). OSCEs are outdated: clinical skills assessment should be centred around workplace-based assessments (WPBAS) to put the 'art' back into medicine. *MedEdPublish*, *6*(4). https://doi.org/10.15694/mep.2017.000189

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013a). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher, 35*(9), 1437-46. https://doi.org/10.3109/0142159x.2013.818634

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013b). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Medical Teacher*, *35*(9), 1447-63. https://doi.org/10.3109/0142159x.2013.818635

Klassen, R.M., Kim, L.E. (2021). Developing Multiple Mini-Interviews for Teacher Selection. In: *Teacher Selection: Evidence-Based Practices*. Springer, Cham. https://doi.org/10.1007/978-3-030-76188-2_8

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education*, 45(10), 1048–1060. https://doi.org/10.1111/j.1365-2923.2011.04025.x

Lara, S., Foster, C., Hawks, M. and Montgomery, M. (2020). Remote assessment of clinical skills during COVID-19: A virtual, high-stakes, summative paediatric objective structured clinical examination. *Academic Paediatrics*, *20*(6), 760-761. https://doi.org/10.1016%2Fj.acap.2020.05.029

Lee, K. C., Ho, C. H., Yu, C. C., & Chao, Y. F. (2020). The development of a six-station OSCE for evaluating the clinical competency of the student nurses before graduation: A validity and reliability analysis. *Nurse Education Today*, *84*, 104247. https://doi.org/10.1016/j.nedt.2019.104247

Lewis, P., Hunt, L., Ramjan, L. M., Daly, M., O'Reilly, R., & Salamonson, Y. (2020). Factors contributing to undergraduate nursing students' satisfaction with a video assessment of clinical skills. *Nurse Education Today*, *84*, 104244. https://doi.org/10.1016/j.nedt.2019.104244

Lumley, T. & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Major, S., Sawan, L., Vognsen, J., & Jabre, M. (2020). COVID-19 pandemic prompts the development of a Web-OSCE using Zoom teleconferencing to resume medical students' clinical skills training at Weill Cornell Medicine-Qatar. *BMJ Simulation & Technology Enhanced Learning*, *6*(6): 376-377. https://doi.org/10.1136%2Fbmjstel-2020-000629

Masson, D. (n.d.). *Balanced Latin Square Generator.* Available at: https://cs.uwaterloo.ca/~dmasson/tools/latin_square/

Maxwell, J.A. (2005). *Qualitative Research Design: An Interactive Approach.* London: Sage.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.

McManus, I. C., Elder, A. T., & Dacre, J. (2013). Investigating possible ethnicity and sex bias in clinical examiners: An analysis of data from the MRCP(UK) PACES and nPACES examinations. *BMC Medical Education*, 13(1). https://doi.org/10.1186/1472-6920-13-103

Mitchell, M., Henderson, A., Groves, M., Dalton, M., & Nulty, D. (2009). The objective structured clinical examination (OSCE): Optimising its value in the undergraduate nursing curriculum. *Nurse Education Today, 29*(4), 398-404. https://doi.org/10.1016/j.nedt.2008.10.007

Mohr, C. D., & Kenny, D. A. (2006). The how and why of disagreement among perceivers: An exploration of person models. *Journal of Experimental Social Psychology, 42*(3), 337–349. https://doi.org/10.1016/j.jesp.2005.05.007

Mortsiefer, A., Karger, A., Rotthoff, T., Raski, B., & Pentzek, M. (2017). Examiner characteristics and interrater reliability in a communication OSCE. *Patient Education and Counseling, 100*(6), 1230–1234. https://doi.org/10.1016/j.pec.2017.01.013

Najjar, R. H., Docherty, A., & Miehl, N. (2016). Psychometric Properties of an Objective Structured Clinical Assessment Tool. *Clinical Simulation in Nursing*, *12*(3), 88–95. https://doi.org/10.1016/j.ecns.2016.01.003

National Commission for Certifying Agencies (2014). *Standards for the Accreditation of Certification Programs.* Institute for Credentialing Excellence. Available at: https://www.nasfaa.org/uploads/documents/2016_NCCA_Standards.pdf.

Navas-Ferrer, C., Urcola-Pardo, F., Subiron-Valera, A.B., & German-Bes, C. (2017). Validity and reliability of Objective Structured Clinical Evaluation in nursing. *Clinical Simulation in Nursing, 13*(11), 531-543. https://doi.org/10.1016/j.ecns.2017.07.003

Newton, P.E. (2017) *An approach to understanding validity arguments*. Office of Qualifications and Examinations Regulation. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/653070/An_approach_to_understanding_validation_arguments.pdf

OSCE. (2022, June 7). Nursing & Midwifery Council. https://www.nmc.org.uk/registration/joining-the-register/toc/toc-2021/osce/

O'Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018) The state-of-the-art in digital technology-based assessment. *European Journal of Education, 53*(2), 160-175. https://doi.org/10.1111/ejed.12271

Onwuegbuzie, A.J. & Leech, N.L. (2005). On Becoming a Pragmatic Researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375-387. https://doi.org/10.1080/13645570500402447

Paravattil, B., & Wilby, K. J. (2019). Optimizing assessors' mental workload in rater-based assessment: a critical narrative review. *Perspectives on Medical Education*, 8, 339-345. https://doi.org/10.1007/s40037-019-00535-6

Patrício, M. F., Julião, M., Fareleira, F., & Carneiro, A. V. (2013). Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher*, Vol. 35, 503–514. https://doi.org/10.3109/0142159X.2013.774330

Preusche, I., Schmidts, M. & Wagner-Menghin, M. (2012). Twelve tips for designing and implementing a structured rater training in OSCEs. *Medical Teacher, 34*(5), 368-372. https://doi.org/10.3109/0142159x.2012.652705

Purpora, C., & Prion, S. (2018). Using student-produced video to validate head-to-toe assessment performance. *Journal of Nursing Education*, *57*(3), 154–158. https://doi.org/10.3928/01484834-20180221-05

Roberts, R., Cook, M., & Chao, I. (2020) Exploring assessor cognition as a source of score variability in a performance assessment of practice-based competencies. *BMC Medical Education*, *20*(1), 168. https://doi.org/10.1186/s12909-020-02077-6

Robertson, T., Durick, J., Brereton, M., Vetere, F., Howard, S. Nansen, B. (2012). Knowing our users: scoping interviews in design research with ageing participants. *Proceedings of the 24th Australian Computer-Human Interaction Conference*. https://doi.org/10.1145/2414536.2414616

Robinson, O.C. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative Research in Psychology*, *11*(1), 25-41. https://doi.org/10.1080/14780887.2013.801543

Rushforth, H. E. (2007). Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. *Nurse Education Today*, 27(5), 481–490. https://doi.org/10.1016/j.nedt.2006.08.009

Schleicher, I., Leitner, K., Juenger, J., Moeltner, A., Ruesseler, M., Bender, B., … Kreuder, J. G. (2017). Examiner effect on the objective structured clinical exam - A study at five medical schools. *BMC Medical Education*, 17(1). https://doi.org/10.1186/s12909-017-0908-1

Scrimgeour, D. S. G., Cleland, J., Lee, A. J., & Brennan, P. A. (2019). Prediction of success at UK Specialty Board Examinations using the mandatory postgraduate UK surgical examination. *BJS Open*, 3(6), 865-871. https://doi.org/10.1002/bjs5.50212

Selim, A. A., Ramadan, F. H., El-Gueneidy, M. M., & Gaafer, M. M. (2012). Using Objective Structured Clinical Examination (OSCE) in undergraduate psychiatric nursing education: Is it reliable and valid? *Nurse Education Today, 32*(3), 283–288. https://doi.org/10.1016/j.nedt.2011.04.006

Setyonugroho, W., Kennedy, K. M., & Kropmans, T. J. B. (2015). Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Education and Counseling*, 98, 1482–1491. https://doi.org/10.1016/j.pec.2015.06.004

Smith, V., Muldoon, K., & Biesty, L. (2012). The Objective Structured Clinical Examination (OSCE) as a strategy for assessing clinical competence in midwifery education in Ireland: A critical review. *Nurse Education in Practice*, 12(5), 242–247. https://doi.org/10.1016/j.nepr.2012.04.012

Solà-Pola, M., Morin-Fraile, V., Fabrellas-Padrés, N., Raurell-Torreda, M., Guanter-Peris, L., Guix-Comellas, E., & PulpónSegura, A. M. (2020). The Usefulness and acceptance of the OSCE in nursing schools. *Nurse Education in Practice, 43*, 102736. https://doi.org/10.1016/j.nepr.2020.102736

St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education, 21*(3), 627–642. https://doi.org/10.1007/s10459-015-9656-3

Tavares, W. & Eva, K. (2013). Exploring the impact of mental workload in rater-based assessments. *Advances in Health Sciences Education,* 18(2), 291-303. https://psycnet.apa.org/doi/10.1007/s10459-012-9370-3

Teddie, C. & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal of Mixed Methods Research, 1,* 77-100. https://doi.org/10.1177/1558689806292430

Ten Cate, O. & Regehr, G. (2019). The power of subjectivity in the assessment of medical trainees. *Academic Medicine, 94*(3), 333-337. https://doi.org/10.1097/acm.0000000000002495

Terry, G., Hayfield, N., Clarke, V. & Braun, V. (2017). *Thematic Analysis.* In C. Willig, & W. Stainton-Rogers (Eds.), *The Sage Handbook of Qualitative Research in Psychology,* 17-37. London, UK: Sage.

Thompson, B. & Vacha-Haase, T. (2017) Reliability. In Secolsky, C. & Denison, D.B. (eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. Abingdon: Routledge.

Traynor, M., & Galanouli, D. (2015). Have OSCEs come of age in nursing education? *British Journal of Nursing*, *24*(7), 388–391. https://doi.org/10.12968/bjon.2015.24.7.388

United Nations Educational, Scientific and Cultural Organization (2020). *Education: From disruption to recovery*. Available at: https://en.unesco.org/covid19/educationresponse#schoolclosures

Van Der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice and Research: Clinical Obstetrics and Gynaecology*, *24*(6), 703–719. https://doi.org/10.1016/j.bpobgyn.2010.04.001

Walsh, M., Bailey, P., & Koren, I. (2009) Objective structured clinical evaluation of clinical competence: An integrative review. *Journal of Advanced Nursing, 65*(8), 1584-1595. https://doi.org/10.1111/j.1365-2648.2009.05054.x

Weaver, K. & Olson, J.K. (2006). Understanding paradigms used for nursing research. *Journal of Advanced Nursing, 53*(4): 459-469. https://doi.org/10.1111/j.1365-2648.2006.03740.x

Willis, G.B. (2015). *Analysis of the Cognitive Interview in Questionnaire Design.* Oxford: Oxford University Press.

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education*, *19*, 409–427. https://doi.org/10.1007/s10459-013-9453-9

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education : Theory and Practice*, *18*(3), 325–341. https://doi.org/10.1007/s10459-012-9372-1

Yeates, P., Moreau, M., & Eva, K. (2015). Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? *Academic Medicine : Journal of the Association of American Medical Colleges*, *90*(7), 975–980. https://doi.org/10.1097/acm.0000000000000650

Appendix A

Example Plain Language Statement

PLAIN LANGUAGE STATEMENT – Phase 3

Project title: Assessor cognition and score reliability in nursing Objective Structured Clinical Examinations (OSCEs)

Principal Investigator: Conor Scully (conor.scully9@mail.dcu.ie)

Supervisors: Prof. Michael O'Leary (michael.oleary@dcu.ie), Dr. Mary Kelly (mary.t.kelly@dcu.ie), Dr. Zita Lysaght (zita.lysaght@dcu.ie)

Research Overview

The purpose of this research, funded by Prometric, is to determine *how* assessors of performance assessments such as the OSCE make judgements about student performances, and whether an understanding of this can improve score reliability and assessment design.

What involvement will require

If you volunteer to participate in this project, you will be asked to participate in a series of procedures relating to your role as an OSCE examiner:

1. You will be asked to talk through the marking guide for two OSCEs (blood pressure measurement and naso-gastric tube insertion), and explain to the researcher how you understand the items in the guide.
2. You will be asked a series of questions about your role as an examiner.
3. You will view a series of videos of students completing these two OSCEs. You will be asked to award the students a grade, as if you were judging them "live" in a real OSCE. While watching the videos, you will be asked to vocalise your thought processes as to what the student is doing well or badly, and how you are interpreting their performance overall. Feel free to pause the video at any stage if you have something to say. After the video is over, the researcher might ask you to elaborate further on specific comments you made; he might rewind the video to aid your memory.

It is expected that the total participation time will be approximately one hour. The audio from the session will be recorded. The data will be transcribed by a third-party service, Happy Scribe. The researcher will edit the audio file before it is sent to be transcribed, such that your name is not included.

Potential risk

There is no potential risk beyond what is encountered in your day-to-day life. Your participation in this study is anonymous: your name will not be included, and any identifiable details about you will be removed.

Benefits

By participating in this project, you will be able to reflect on your own role as an assessor in undergraduate nursing OSCEs. The results of this phase of the study will be presented to participants, who might wish to use the data to inform future decisions about the OSCE, or to consider how their own practice as assessors might be improved.

Voluntary Participation

Participation in this study is **voluntary**. You are free to withdraw from participation any time you like, without having to provide a reason. You can decline to participate in the study at any time before it begins; you can end your participation at any time while watching the videos; and you can, after the study is over, ask that your data be withdrawn. If you decide to withdraw your data it will be disposed of permanently. Please note that while the principal investigator will do all he can to protect your anonymity, there are legal limitations to this. Collected data might be subject to a subpoena or freedom of information claim; in that case, the principal investigator might have to disclose data without informing the participant.

Personal Data Protection Notice (GDPR)

DCU Data Protection Officer: Mr. Martin Ward (data.protection@dcu.ie // Ph: 7007476)

- All data collected in this study will be securely stored on a password protected and encrypted computer. Only the principal investigator and his supervisors will have access to these data. The data will be permanently deleted after five years, or at the behest of the participant.
- Every participant has the right to lodge a complaint with the Irish Data Protection Commission.
- There are no consequences to failing to provide the personal data to any participants.

If you have questions about the study, please contact the principal investigator at: conor.scully9@mail.dcu.ie

If you have concerns about the study and wish to contact an independent person, please contact: The Secretary, Dublin University Research Ethics Committee, c/o Research and Innovation Support, Dublin City University, Dublin 9. Ph. 01-7008000. E: rec@dcu.ie

Appendix B

Example Informed Consent Form

Project Title: Assessor cognition as a means of improving score reliability in nursing Objective Structured Clinical Examinations

Principal Investigator: Conor Scully (conor.scully9@mail.dcu.ie), PhD candidate, Dublin City University

Principal Supervisor: Prof. Michael O'Leary (michael.oleary@dcu.ie), School of Policy and Practice, Dublin City University

This research project will examine the cognitive processes through which assessors go when judging students who undertake Objective Structured Clinical Examinations (OSCEs) in nursing. It is expected that a better understanding of these processes will improve the reliability of awarded scores and therefore improve assessment quality. This research is funded by Prometric through the Centre for Assessment Research, Policy and Practice in Education (CARPE) at Dublin City University.

Please circle Yes or No for each question

I have read the Plain Language Statement	Yes/No
I understand the information provided	Yes/No
I have had an opportunity to ask questions about the study	Yes/No
I received satisfactory answers to all my questions (if I had any)	Yes/No
I am aware of the procedure for participating in this study	Yes/No
I understand the potential benefits and harms inherent in my participation	Yes/No
I understand that my participation will be recorded	Yes/No
I understand the information related to data protection	Yes/No
I understand that my data will be stored for five years after completion	Yes/No
I am aware that there are legal limitations to data confidentiality	Yes/No
I am aware that I can withdraw from participation at any point	Yes/No

Signature:

I have read and understood the information in this form. My questions have been answered by the researcher, and I have a copy of this consent form. I therefore consent to participate in this research project.

Participant's signature:

Name in block capitals:

Date:

Appendix C

Interview Schedule: Phase 1

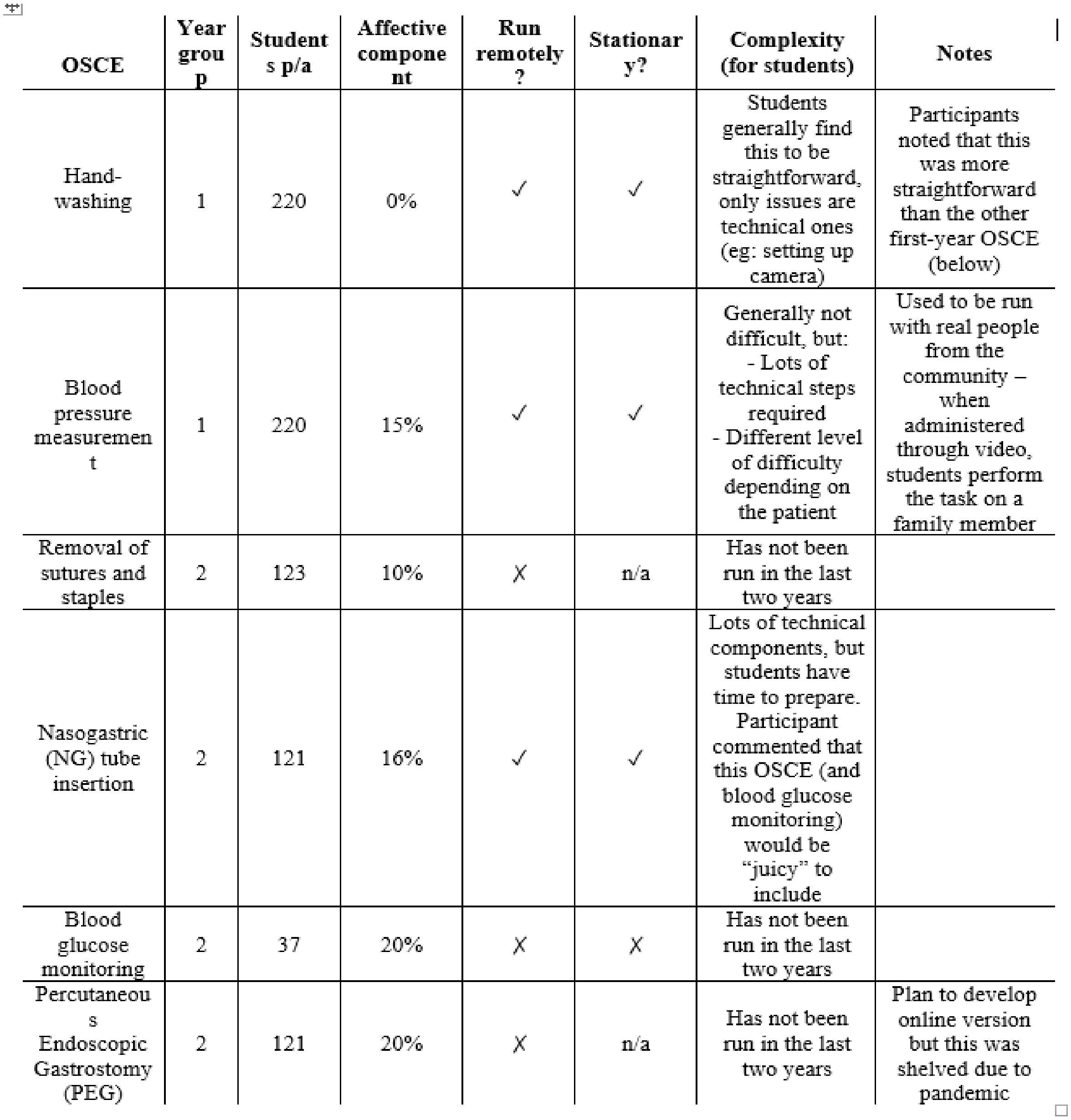
- Are there any OSCEs that you have personally found difficult to mark consistently? If so, please describe the OSCE and what makes it difficult.
- Please describe OSCEs that have tended to lead to divergence in how they are assessed. In other words, are there OSCEs where assessors are likely to have differing opinions about a student's performance level? What do you think causes this divergence?
- Are there any procedures in place which are designed to bring about consistency between OSCE assessors? If so, please describe them.
- Please describe the process(es) by which assessors are trained in how to examine the OSCE.
- Have you ever encountered a situation where you found that an assessor was being overly harsh or lenient? If so, why do you think this was the case? Please elaborate on what happened.
- Do you have any comments about the study?

Appendix D

Documentation of OSCE selection process

The table below details the six OSCEs, and the extent to which they met the inclusion criteria. This information in this table was derived from an analysis of the marking guides for each OSCE, the tables completed by module coordinators in advance of the interviews, and the data from the interviews (see section 3.6).

Table D *Criteria for OSCE selection*



OSCE	Year group	Students p/a	Affective component	Run remotely?	Stationary?	Complexity (for students)	Notes
Hand-washing	1	220	0%	✓	✓	Students generally find this to be straightforward, only issues are technical ones (eg: setting up camera)	Participants noted that this was more straightforward than the other first-year OSCE (below)
Blood pressure measurement	1	220	15%	✓	✓	Generally not difficult, but: - Lots of technical steps required - Different level of difficulty depending on the patient	Used to be run with real people from the community – when administered through video, students perform the task on a family member
Removal of sutures and staples	2	123	10%	✗	n/a	Has not been run in the last two years	
Nasogastric (NG) tube insertion	2	121	16%	✓	✓	Lots of technical components, but students have time to prepare. Participant commented that this OSCE (and blood glucose monitoring) would be "juicy" to include	
Blood glucose monitoring	2	37	20%	✗	✗	Has not been run in the last two years	
Percutaneous Endoscopic Gastrostomy (PEG)	2	121	20%	✗	n/a	Has not been run in the last two years	Plan to develop online version but this was shelved due to pandemic

Of these six OSCEs, three (removal of sutures and staples, blood glucose monitoring, and percutaneous endoscopic gastronomy (PEG)) had never been adapted for video. The coordinator of the removal of sutures and staples OSCE specified via email before the interview that it would not be suitable for inclusion. This OSCE was removed from the study at this stage.

The coordinator of the blood glucose monitoring OSCE similarly stated that it would not be suitable for inclusion, due to the fact that it required movement on the part of the participant:

> The blood glucose monitoring, there's a bit of moving around. So they are over by the sink and then they go to the trolley and ready to get their equipment and then go to the patient. So, I suppose, just having a very practical kind of thing would you be better off picking one that's just maybe they only stay in one place.

For this reason, this OSCE was removed from the study. The PEG OSCE was similarly removed from consideration due to the fact that it had never been adapted for video, and had not been administered at all in the last two years; this would reduce the number of assessors with experience examining the OSCE who could be invited to participate in the think aloud.

Of the three remaining OSCEs (hand-washing, blood pressure monitoring, and naso-gastric tube insertion (NG)), only the latter two contained any element of affective skills. Additionally, participants commented that both the blood pressure and NG OSCEs were more likely to provide rich data for the researcher.

In terms of the blood pressure OSCE, participants commented that it would be a good OSCE to include because there are a lot of steps involved in completing it successfully, which are often quite technically difficult for students (all emphasis added):

> “Blood pressure is probably one of those because it's *technically quite difficult*…”

> “I personally would have found, the students had difficulty with where I could see on the video the students who learned to do it, because *there's a lot of steps*, there's *a lot of smaller technical aspects*”.

Additionally, because the measurement of blood pressure has to take place on a real person (rather than a mannequin), this OSCE has an additional element of variability, due to natural differences between individuals. As noted by one participant:

> "There's *another variable that comes into play* is sometimes we bought relatives or people we knew come in from the community and they were the patients and then we would take the blood pressure on those people and these were some of them are retired people and did have heart complications and so forth. And *some of them are quite hard to take a blood pressure unless you're very skilled* and that comes with time and experience. So, *some of them are hard to actually get a pulse*, taking blood pressure, and that could add other pressure, pressure on the students."

Finally, the coordinator of the blood pressure OSCE stated explicitly that she had problems in terms of assessor consistency in the past:

> "Well, my template… *irrespective of what you put on the template assessors assessed it whatever way they felt*"

For these reasons, the **blood pressure OSCE** was chosen for inclusion in the study.

In terms of the NG OSCE, participants emphasised again that OSCEs which had a large number of technically specific steps were more likely to lead to divergence between assessors:

> "And I think there there's sometimes problems with *more complicated skills*, because you can't put each line down you need to assess because there has to be a little bit of allowing for slight variations because sometimes there are ways of doing things that are still right. And the problem there lies with the assessor, and *what they view as right or wrong*."

Subsequent to this comment, the researcher asked the coordinator of the NG OSCE to specify that this OSCE would be one that is classified as having "more complicated skills":

> Researcher: I know one of yours, isn't it *the gastric tube insertion*?
>
> Participant: Yeah, *that's a complicated one as there's lots of steps…the gastric tube one is the complicated one*.

In addition to this, participants in the interview had a discussion as to which OSCE(s) would be mostly likely to result in rich data in terms of assessor divergence. One participant commented that the skills assessed by another coordinator (blood glucose monitoring and NG) would be most likely to result in this divergence:

> "I think Alicia's [pseudonym] skills that she talks about would probably be the ones that *you would get the juiciest stuff*."

This participant explicitly compared these two OSCEs with the hand-washing OSCE administered to first-years, which she said would be "Not juicy enough!"

The blood glucose monitoring OSCE (for reasons cited above) was deemed not to be suitable for inclusion. As such, the decision was made to include the **NG OSCE**, due to its (relatively) high level of affective skills, the high number of technical steps required in order to achieve a good grade, and the comments from participants that it would be "juicy" to include.

Appendix E
Procedure for setting up and recording OSCEs
(Adapted from Boursicot & Roberts, 2005)

Before the OSCE:

1. Determining a venue: the OSCEs were staged in the Lab, a facility designed to help student nurses learn clinical skills. The researcher contacted the relevant coordinator to request access to the Lab and determine an appropriate date for the OSCE to be staged.

2. Recruitment of examiners: there were no examiners present in the Lab during the OSCE, as the judgement of the participant's performance took place at a later stage, during the cognitive interviews.

3. Determining the running order of the stations: With all three participants, the blood pressure OSCE was filmed first, followed by the NG tube insertion OSCE. Blood pressure measurement is a skill that is taught in first year, and it was expected that participants would be more experienced performing it. As such, it was completed first in case participants were nervous about filming.

4. Listing required equipment: the researcher's co-supervisor (Dr. Mary Kelly) ensured the day before filming that the equipment in the Lab was present and set up properly. The blood pressure OSCE required a sphygmomanometer and a stethoscope; while the NG tube OSCE required an NG tube, measuring tape, pH strips, orange juice, a marker and a syringe. Participants were also provided with gloves, aprons and hand sanitiser. Additionally, participants were provided with masks to wear while completing the OSCEs, in order to stop the potential spread of COVID-19.

5. Production and processing of marking guides: the participant was not graded while filming took place. As such, there was no need for marking guides to be present in the Lab.

6. Liaise with clinical skills staff: as noted by Boursicot & Roberts (2005, p.18), clinical skills staff can greatly assist with the administration of an OSCE. The researcher and Dr. Kelly

liaised with the clinical skills nurses to fix the date and time for the OSCE and ensure that the correct equipment was in place. The researcher also met with the technical director of the Lab the week before the OSCE, to be taught how to operate the recording equipment.

During the OSCE:

7. Briefing: the researcher spoke extensively (via email) with participants the week before filming. Participants were informed of the layout of the OSCE, the procedure for completing the OSCE, and had the opportunity to ask any questions. In line with the relevant ethical guidelines, the participants were informed that they were free to cease participation at any time, without having to give a reason.

8. Refreshments: The participants were provided with water during the OSCE, as well as a voucher to get lunch on campus when filming was completed.

After the OSCE:

9. Cleaning the Lab: the researcher and Dr. Kelly ensured that the Lab was left as it was before the OSCE was set up.

10. Collection of video recording: in line with the relevant data collection provisions, the researcher immediately transferred all video files to his encrypted laptop, and with backup copies on DCU’s secure Google Drive.

11. Debriefing: subsequent to the OSCE, the researcher sent debriefing information to the participant, thanking them for participation, informing them of the aims of the study, and inviting them to contact him if they have any questions, or would like to obtain recordings of their performance.

Appendix F

Interview Schedule: Phase 3

Semi-structured interview

Main questions

Please talk me through the process of assessing a student in an OSCE. Specifically, how you determine if a student has passed the exam?

Is there anything in particular you look out for when judging a student?

How do you determine if a student has completed an item on the checklist properly?

Please describe how you would imagine a “minimally competent” or “borderline competent” student. What distinguishes someone who has *just* passed from someone who has *just* failed?

How do you think you have developed as an assessor since you began assessing? Is there anything different about how you approach the task of assessment?

Has there ever been a time when you and another assessor disagreed about how a student performed? If so, how was this resolved?

Back-up/prompt questions

How do you determine if a student has completed an item on the checklist properly?

Has there ever been a time when you were unsure if a student has done something in the correct way?

Are there any items within the OSCE that you worry about marking in the same way as other assessors?

How do you think you have developed as an assessor since you began assessing? Is there anything different about how you approach the task of assessment?

Do you find it easier or more difficult to assess OSCEs than you used to?

Is there any aspect of student performance that you think has become more or less important since you began assessing? Is there anything you are more or less likely to look out for now than when you began assessing?

Cognitive interview

For each item in the marking guide, briefly explain how you would tell if a student has done this correctly or not.

Think aloud

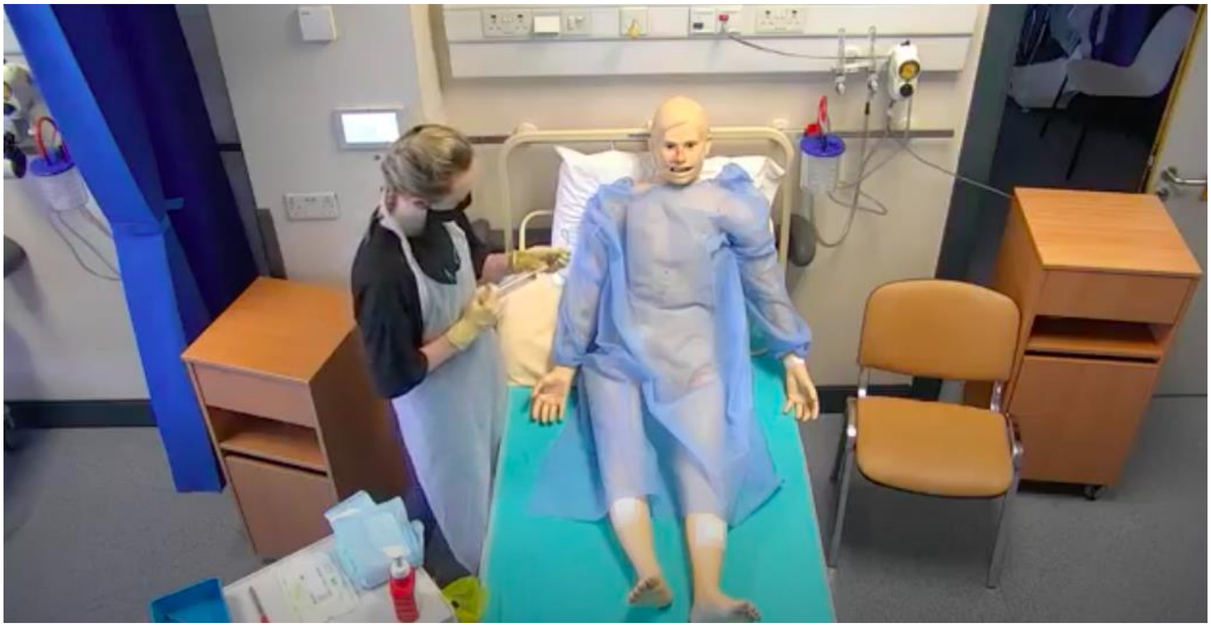
Imagine you are judging a series of OSCE performances for a summative, end-of-year assessment of second-year students. The OSCEs being administered today are blood pressure measurement and naso-gastric tube insertion. You don’t know any information about the student who is to perform. It is your job to assess the student’s performance and complete the marking guide provided to you.

When the video begins to play, please vocalise your thought processes in as much detail as possible. You are welcome to pause the video yourself, any time you like and as often as you like, if you need more time to elaborate. During this time, the researcher may ask you to discuss certain comments in more detail.

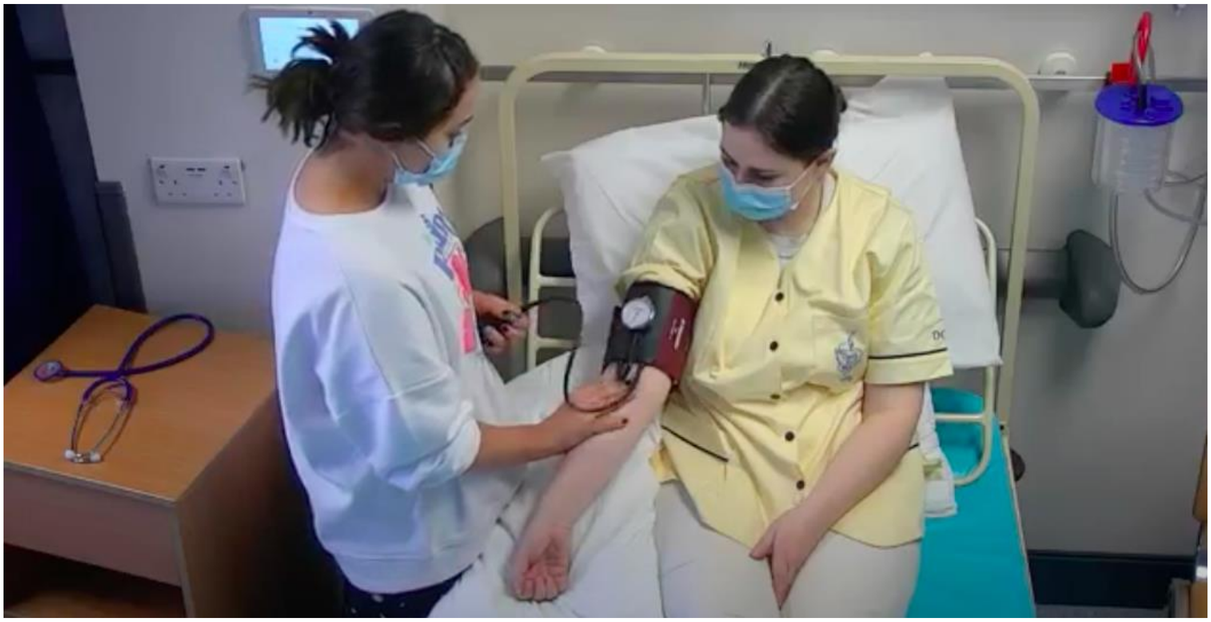
One thing to note: For the NG Tube OSCE, the mannequin doesn’t have a stomach, so the participants use orange juice to check the pH.

Appendix G

Stills from Recorded Videos



Still from P01NG: student takes an "aspirate" from the stomach



Still from P02BP: student uses her fingers to find a pulse

Appendix H

Marking Guides: Phase 3

Introductory questions

What is your name?

What is your age?

What is your role?

How long have you been in this role?

For how many years have you been involved in the assessment of OSCEs?

Approximately how many years have you assessed an OSCE for **blood pressure measurement?**

Approximately how many years have you assessed an OSCE for **naso-gastric tube insertion?**

How would you rate yourself as an assessor of OSCEs?

Advanced beginner / competent / proficient / expert

Blood pressure

Criteria	**Completed?**
Follow infection control guidelines in relation to: ‘bare below the elbows’, jewellery, hair, nails	
Follow infection control guidelines regarding hand hygiene. State you would perform this.	
Explain the procedure to your patient and gain consent.	
Patient to be seated comfortably in chair.	
Restrictive clothing removed from patient’s upper arm before applying cuff. Arm / hand in correct position.	
Arm correctly positioned, at heart level and supported.	
Cuff bladder correctly placed on arm.	
Locate Brachial artery.	
Brachial check to establish an estimated systolic pressure. This estimated number will be the number prior to the addition of the ’20-30mmHg’.	
Cuff inflated to 20-30 mmHg above estimated systolic pressure you stated previously.	
Cuff deflated in a controlled manner.	
Systolic & diastolic blood pressure reading stated to video.	

How would you rate the participant’s **communication** skills? Fail / borderline pass / good / excellent

How confident would you be allowing this student to perform this procedure **on you**? Not at all confident / Just about confident / Confident / Very confident

How would you rate this performance **overall**? Fail / borderline pass / good / excellent

Please complete the following sentence: “I think this is the type of student who…”

NG Tube insertion:

Criteria	**Completed?**
Wears correct full uniform - hair tied up/off shoulders.	
Clearly states that handwashing has been completed prior to gloves and preparation of equipment.	
Patient to be in position before starting video, assume consent.	
Pre-arrange a “stop” signal patient can use i.e. raise their hand (articulate on video).	
Select distance mark on tube i.e. length to be passed: distance from nose to ear lobe, then ear lobe to lowermost section of breastbone. (Must be clearly visible on video).	
Lubricate proximal tip of tube with KY jelly. Check nostrils patent.	
Insert tube into nostril, sliding backwards and inwards along floor of nose to nasopharynx.	
As tube passes nasopharynx patient can be asked to swallow/take sips of water (if permitted).	
Advance tube through pharynx until predetermined mark is reached (articulate on video).	
Check tube position is accurate using aspiration method. (Flush tube with 20mls air to clear, aspirate 2 – 10mls of stomach contents, test with pH indicator strips. (Articulate on the video)	
Marks NG tube at exit point from nostril with indelible marker (immediately above or below adhesive tape).	
Secure tube in place with adhesive tape/dressing.	
Attach spigot.	
Clearly states patient comfort is offered.	

How would you rate the participant’s **communication** skills? Fail / borderline pass / good / excellent

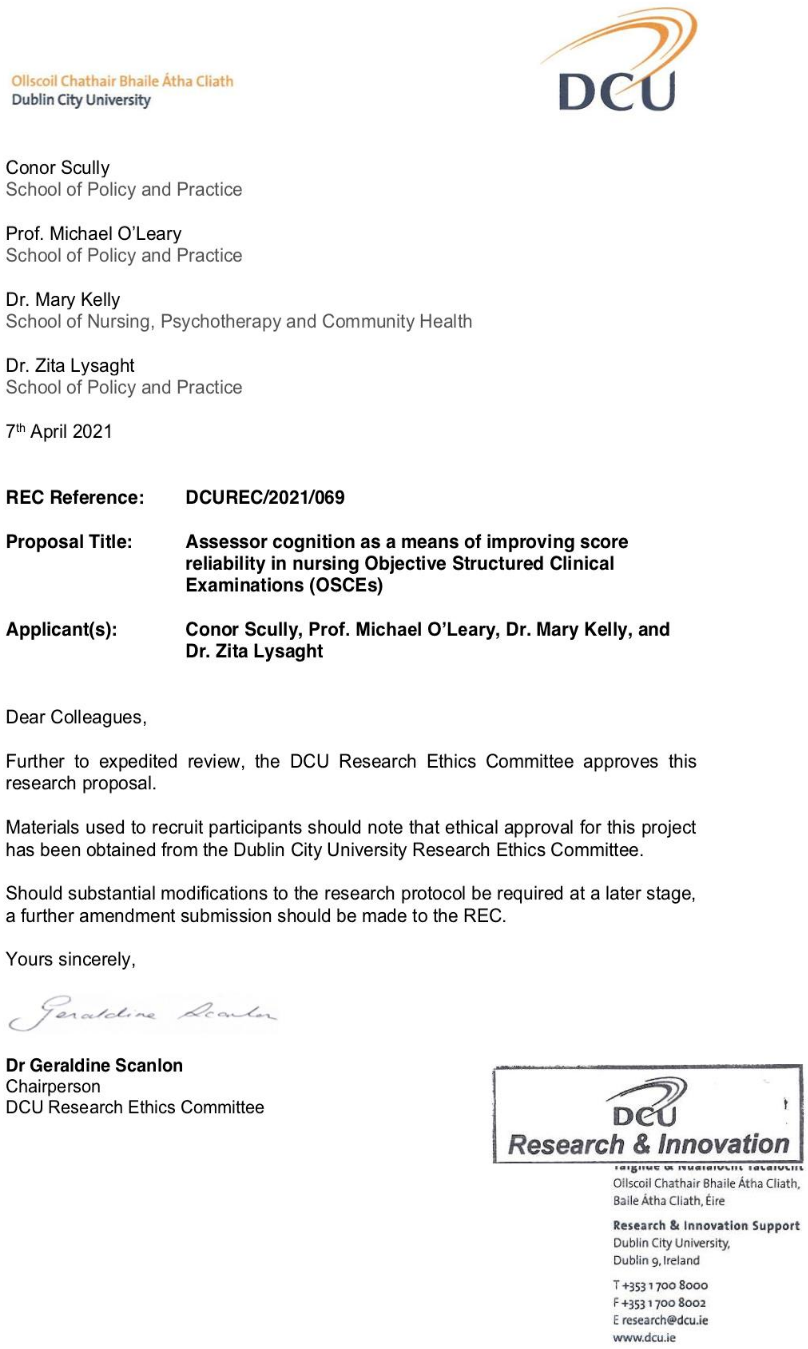
How confident would you be allowing this student to perform this procedure **on you**? Not at all confident / Just about confident / Confident / Very confident

How would you rate this performance **overall**? Fail / borderline pass / good / excellent

Please complete the following sentence: “I think this is the type of student who…”

Appendix I

Ethical Approval



Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

DCU

Conor Scully
School of Policy and Practice

Prof. Michael O'Leary
School of Policy and Practice

Dr. Mary Kelly
School of Nursing, Psychotherapy and Community Health

Dr. Zita Lysaght
School of Policy and Practice

7th April 2021

REC Reference:	**DCUREC/2021/069**
Proposal Title:	**Assessor cognition as a means of improving score reliability in nursing Objective Structured Clinical Examinations (OSCEs)**
Applicant(s):	**Conor Scully, Prof. Michael O'Leary, Dr. Mary Kelly, and Dr. Zita Lysaght**

Dear Colleagues,

Further to expedited review, the DCU Research Ethics Committee approves this research proposal.

Materials used to recruit participants should note that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee.

Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,

Geraldine Scanlon

Dr Geraldine Scanlon
Chairperson
DCU Research Ethics Committee

DCU
Research & Innovation

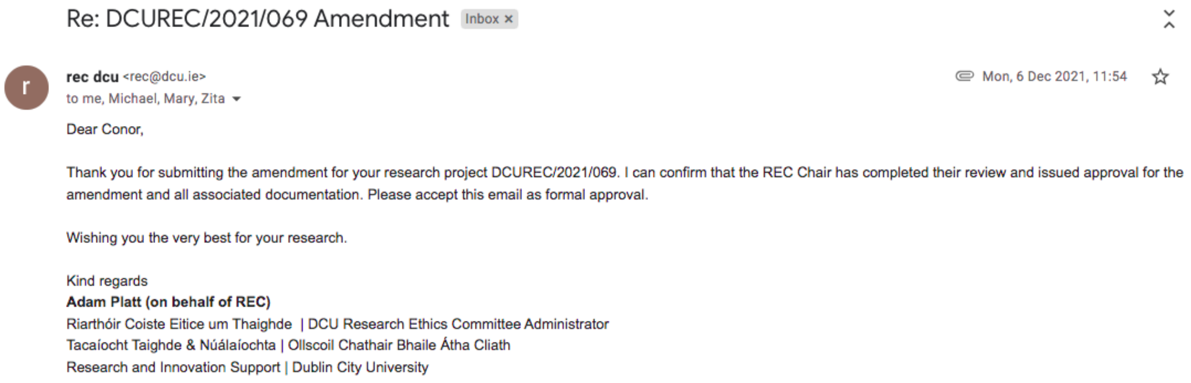
Taighde & Nuálaíocht Tacaíocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie

Appendix J

Ethics Amendment Approval



Re: DCUREC/2021/069 Amendment

rec dcu <rec@dcu.ie>
to me, Michael, Mary, Zita

Mon, 6 Dec 2021, 11:54

Dear Conor,

Thank you for submitting the amendment for your research project DCUREC/2021/069. I can confirm that the REC Chair has completed their review and issued approval for the amendment and all associated documentation. Please accept this email as formal approval.

Wishing you the very best for your research.

Kind regards
Adam Platt (on behalf of REC)
Riarthóir Coiste Eitice um Thaighde | DCU Research Ethics Committee Administrator
Tacaíocht Taighde & Núálaíochta | Ollscoil Chathair Bhaile Átha Cliath
Research and Innovation Support | Dublin City University