Quantitative Evaluations of Teaching Quality in Universities: The Impact of Students' Gender-Specific Expectations on Teaching Evaluations

by

VIKTORIA JAKCSIOVA

A thesis
submitted in partial fulfilment of the requirements
of the University of Greenwich for the Degree of
Doctor of Philosophy

June 2023

DECLARATION

I certify that the work contained in this thesis, or any part of it, has not been accepted in substance for any previous degree awarded to me or any other person, and is not concurrently being submitted for any other degree other than that of MPhil/PhD Psychology which has been studied at the University of Greenwich, London, UK.

I also declare that the work contained in this thesis is the result of my own investigations, except where otherwise identified and acknowledged by references. I further declare that no aspects of the contents of this thesis are the outcome of any form of research misconduct.

I declare any personal, sensitive or confidential information/data has been removed or participants have been anonymised. I further declare that where any questionnaires, survey answers or other qualitative responses of participants are recorded/included in the appendices, all personal information has been removed or anonymised. Where University forms (such as those from the Research Ethics Committee) have been included in appendices, all handwritten/scanned signatures have been removed.

Student Name: .……Viktoria Jakcsiova……

Student Signature: ……Jakcsiova…….

Date: ……21/06/2023…

First Supervisor's Name: ……Dr Jana Uher……

First Supervisor's Signature: …….Dr Uher…….

Date: ……22/06/2023…

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my first supervisor, Dr Jana Uher, for all her invaluable advice and guidance during my PhD studies. I appreciate your outstanding feedback and encouragement. Thank you for helping me progress as a researcher.

I am also very grateful to the rest of my supervisory team, Dr Rebecca Smith, and Professor Claire Monks, for their many insightful comments and suggestions, and kind support. Thanks again to the whole team for guiding me during this journey.

Thank you to the University of Greenwich for granting me the Vice Chancellor's Scholarship.

Many thanks to my PhD and EdD colleagues for creating friendly, encouraging, and productive atmosphere.

I would also like to thank my parents, other family, and friends. Special thanks go to my mother, Maria J., for her continuous encouragement. I also want to thank Katarina J., Zuzana K., and Barbora T., for their support.

ABSTRACT

Universities frequently use student evaluations of teaching (SETs) as a tool to evaluate teaching quality. I conducted two online multi-method empirical studies to explore whether the methodological and methodical limitations of standardised rating scales such as SETs may contribute to rather than impede the manifestation of gender biases in students' evaluations of their lecturers. These limitations include constraints in humans' abilities to generate quantifications as well as potential variations in students' interpretations and use of item statements and answer scale categories.

Study 1 involved surveys with open-ended questions and standardised rating scales, on which participants (N = 181) rated anonymously their "best" or "worst" lecturers from real-life contexts. In Study 2, participants (N = 336) read scenarios describing fictitious lecturers performing teaching behaviours based on my findings from Study 1.

My findings revealed extensive variations in the ways in which student participants interpreted and used item statements and answer scale categories. For every item statement and answer category, the samples generally constructed large fields of meanings or reasons for picking the categories. However, on average, each participant listed only one to two reasons for their rating or choice of the answer category. I found significant gender differences in how frequently participants mentioned certain teaching behaviours, and in scores awarded to "best" male versus "best" female lecturers. In both studies, participants weighted several teaching behaviours differently in their ratings of lecturers of a different gender, although this only occurred in a minority of cases.

Therefore, students may, in some cases, apply gendered schemas when evaluating lecturers and the methodological and methodical limitations of standardised rating scales might enable rather than reduce the manifestation of gender biases. I discuss practical implications for students, universities, and lecturers, and recommend that institutions should re-consider using SETs as the predominant method of teaching evaluation.

CONTENTS

TABLES AND FIGURES

List of tables

List of Appendix tables

List of Figures

Chapter 1: Introduction

In this chapter, I establish the background and contextual foundations of this research. I define student evaluations of teaching (SETs) and briefly describe different types of these surveys as used in the U.K. Afterwards; I discuss how SETs impact universities, students, and lecturers. I present the overall aim and importance of this research, as well as the over-arching research question. This chapter concludes with the outline of the structure of my entire thesis.

1.1 General background of this research

Student evaluations of teaching (SETs) are worldwide one of the most frequently applied assessments of teaching quality (Goos & Salomons, 2017) or teaching effectiveness (Stark & Freishtat, 2014). Teaching quality can be defined as a quality of instruction and teaching practices that stimulate, engage, and challenge students and enable them to develop new knowledge and skills (Darling-Hammond, 2009; TEF, 2016; *see also Chapter 2*).

SETs frequently act as a major factor in decisions about lecturers‘ careers (Hornstein, 2017). In this thesis, I focus specifically on SETs that employ a format of standardised rating scales.

Importantly, students may hold stereotypical beliefs and expectations about their lecturers due to certain obvious yet unrelated to teaching characteristics (e.g., gender, age), which may be reflected in SET results. Specifically, this can result in differences in lecturers’ ratings even when there are no differences in teaching quality. Although different biases and stereotypes may influence students‘ judgements, I focused on gender stereotypes.

Using standardised rating scales, such as SETs, also raises several methodological and methodical concerns. These concerns include relying on students‘ memories and interpretive decisions, as well as structure, wording and format of SETs. Another concern is that individual differences of students may influence the ways in which they interpret key elements of rating scales, item statements, and answers scale categories.

In this research, I explored whether these two problems, the potential influence of gender stereotypes on SETs, as well as the methodological limitations of standardised rating scales, are intertwined. Specifically, I examined whether the standardised format of

SETs may encourage rather than impede potential influences of gender stereotypes on student evaluations of lecturers.

1.2 Student evaluations of teaching in the U.K.

SETs can be defined as feedback designed to capture students' experience of a class and evaluate their lecturer (Ching, 2019). Many crucial findings about SETs originate from research conducted in the countries outside the U.K. (mostly the U.S.) and are considered and included in this thesis. However, the present PhD research focuses on SETs within the U.K.

There are different types of student evaluation surveys. In the U. K., this includes *national surveys*, such as the National Student Survey (NSS), the Postgraduate Research Experience Survey (PRES), and the Postgraduate Taught Experience Survey (PTES). Another type are *institutional surveys*, such as module or course evaluations and surveys specific to institutions.

Each university in the U. K. participates in the NSS, a survey focused on undergraduate students in their final year of study. This survey's purpose is to assess different areas of student experience, such as teaching on the course, assessment and feedback, academic support, organisation and management, and overall satisfaction. The NSS also contains optional open-ended questions about positive and negative aspects of the course (National Student Survey, 2019). Universities in the U.K. are generally obligated to participate in the NSS, although they are no longer obligated to promote this survey to their students. For example, all universities in England that are regulated by the Office for Students must take part in the NSS as an ongoing condition of their registration (National Student Survey, 2023). Universities generally attach high importance to NSS results (Thiel, 2019), and allocate resources to address any perceived problems (Sabri, 2013). In terms of other national surveys, the PRES is the survey directed on postgraduate *research* students, whereas the PTES has a similar format to PRES, but focuses on postgraduate *taught* students. Each of the above-mentioned national surveys generally contains a certain number of core items on a rating scale that are standardised across all universities and several optional items that are specific to each institution.

Institutional surveys, such as module and course evaluations, may slightly differ across institutions, with numerous universities in the U. K. currently applying the EvaSys Online system. Electric paper "Evaluationssysteme" is a supplier of automation software for exams, questionnaires, and evaluation surveys. Generally, students of all years can

participate in module evaluations, and these are usually administered, online or offline, at the end of the term, although this may differ across universities.

Both national and institutional surveys are filled by students anonymously. Students, therefore, cannot be held accountable for responses they provide in these surveys.

1.3 Student evaluations of teaching and their impact

SETs initially served formative purposes, such as for enhancing the quality of teaching. But in the 1970s, it became common to use them as a summative tool, for evaluating lecturers' overall performance (Hornstein, 2017) or to assist in making personnel decisions (Galbraith et al., 2012). Therefore, SETs can form a major factor in decisions about lecturers' careers.

In the U. K., SETs influence a university's position in league tables. This position is considered in the Teaching Excellence Framework (TEF) and can affect staff and student recruitment (Holland, 2019). The TEF takes into account students' responses and views, which are collected through the National Student Survey (Office for Students, 2019). Universities with a TEF award may charge up to the maximum tuition fee of £9,250 per student per year for their full-time programme, compared to £9,000 for universities with no award. Therefore, universities attempt to achieve a good standing in the TEF.

Apart from universities, SETs may also affect students and their learning. As SET scores may impact lecturers' careers, it is reasonable to assume that some academics may feel pressured to "teach to SETs". Academics reported grading more leniently, simplifying content or scheduling an entertaining class right before SETs in hopes of obtaining higher ratings from students (Boring et al., 2016; Simpson & Siguaw, 2000). Focus on student satisfaction may, therefore, come at the expense of good teaching (Bedggood & Pollard, 1999), with some lecturers dedicating more time to activities they believe students perceive as enjoyable instead of activities that enable students to learn (Braga et al., 2014). It can ultimately lead to students misperceiving how proficient in the subject they really are. Students can face challenges in subsequent modules when encountering lecturers who grade justly and challenge students to an appropriate level (Stroebe, 2016).

Finally, the group probably most impacted by SET results are university lecturers themselves, as SETs frequently form an important factor in decisions concerning their

promotion, retention, and salary (Benton & Cashin, 2014; Emery et al., 2003). Lecturers may also be affected by offensive comments that some students provide. These comments might negatively impact lecturers' well-being and mental health (Heffernan, 2023). In sum, it is crucial to ensure that SETs are valid, accurate, not misinterpreted by administrators and that lecturers are treated equally, such as in terms of gender.

1.4 The general aim and importance of the present research

I aimed to identify the ways in which students use and interpret student evaluation scales and to explore how students generate their rating decisions. My further aim was to investigate whether students hold their lecturers to gendered expectations. I examined this problem through scrutinising the methodological and methodical limitations of standardised rating scales such as SETs.

To ensure that the methods applied to evaluate lecturers are accurate and gender equitable, I explored what information (teaching behaviours, attitudes) students consider when making their rating decisions. I also examined possible underlying patterns of stereotypical beliefs influencing students' perceptions of teaching behaviours and teaching evaluations. I investigated on the aspects influencing student judgements in their interpretations of SET items, answer scale categories, teaching behaviours and inferred teaching attitudes. Identifying possible gender-related stereotypical biases can help develop teaching evaluation methods that could enhance accuracy, minimise implicit biases and improve gender equality.

1.5 Organisation of this thesis in relation to my PhD research

In Chapter 1, I provided a broad overview of my PhD research, discussed key terms relevant to SETs and their general background. In Chapter 2, I outline factors unrelated to teaching quality that may influence SETs. To explore how potential gender stereotypes may influence students‘ judgements of lecturers, I also discuss relevant theoretical frameworks from social psychology used in this research. In Chapter 3, in order to scrutinise SETs from the methodological point of view, I apply several methodological concepts from the Transdisciplinary Philosophy of Science Paradigm for Research on Individuals (TPS-Paradigm), which provides conceptual frameworks for analysing research methods across sciences (Uher, 2015a, 2018a, 2018b, 2019, 2020). In Chapter 4, I outline the rationale of this PhD research and present all research aims and research questions. I also justify the general methodology of my two studies as well as my use of a moderate version of social constructionism as my epistemological stance. I

present two online multi-method empirical studies in Chapter 5 (Study 1) and Chapter 6 (Study 2). In the final Chapter 7, I summarise and analyse the key findings and highlight key patterns identified in my data. I conclude by outlining the original contributions of this PhD research and providing relevant recommendations for practice.

1.6 General summary of this chapter

This chapter provided a synopsis of this PhD research. I explained the general context in which SETs are applied and I outlined different types of SETs used in the U.K., such as national and institutional surveys. I highlighted general impact of my research on students, universities, and lecturers, and presented an over-arching research aim. The next chapter discusses problems related to SETs, such as inconsistency of data that students enter into SETs with verifiable information, factors unrelated to teaching quality that may influence SET ratings, as well as gender biases and stereotypes in educational context.

Chapter 2: Gender biases and stereotypes in SETs and other influences on SETs unrelated to teaching quality

SETs aim to evaluate teaching quality or teaching effectiveness. However, these evaluations may be influenced by factors independent of the lecturers, such as students‘ approaches towards SETs and the ways in which students complete SETs.

Students may also hold stereotypical beliefs about their lecturers, which may be reflected in results obtained from SET forms. This can result in differences in lecturers’ ratings even when there are no differences in teaching quality. Although various types of biases can influence evaluations, this chapter focuses on potential gender bias and gender stereotypes. The first part of this chapter outlines and discusses students‘ approaches towards SETs. The second part examines gender bias and gender stereotypes in educational settings and SETs and applies relevant theoretical framework to explain empirical findings.

2.1 Students’ approaches towards SETs

2.1.1 Data entered into SETs may not reflect verifiable information: empirical findings and theoretical background

Because SETs are worldwide one of the most frequently applied assessments of teaching quality (Goos & Salomons, 2017), universities and researchers commonly assume that students put effort into filling out evaluations and complete them accurately. For example, there is rarely a discussion about the consistency of students’ comments with verifiable information in research on SETs (Clayson & Haley, 2011). However, even though 80-85% of student participants in an Australian study did not attend previous classes, they still answered the question “Did classes start on time?”. Only 15-20% of participants[1] could, therefore, give an accurate answer, but even participants not attending lectures still provided ratings (Bedggood & Pollard, 1999). Similarly, when asked to evaluate a fictitious professor who never delivered any lectures, 66% of participants studying Medicine rated this professor and even included several qualitative comments, both positive and negative, about the lectures. Even when the fictitious lecturer’s name

[1] Unless otherwise specified, “participants” always refers to student participants in this chapter.

was supplemented with a random photograph of a young model, 49% of participants still rated the lecturer (Uijtdehaage & O'Neal, 2015).

In another study, participants answered two items on SETs about lab work and whether they found it beneficial. Surprisingly, participants in courses without labs frequently failed to mark the item as "not applicable". In one course, only 12 out of 32 participants did so, while the remaining 20 participants rated non-existent labs (Emery et al., 2003). In a different study, participants rated the "The library services and resources are good enough for my needs" with "neither agree nor disagree" (on a 5-point scale) but then commented they have never even used the library. Participants, therefore, provided ratings instead of choosing a more appropriate "not applicable" option. This was also the case with other items, such as IT resources or enhancement of communication skills (Ashby et al., 2011).

At a private university, the instructor realised they had been coming to every class a few minutes late. However, SETs again failed to reflect that, as all the students reported the instructor was showing up to class on time (Clayson & Haley, 2011). A similar case study involved a professor recounting they had a habit of coming to every class five minutes early and never missed a class. However, on a "reliability of meeting class" item, they obtained a rating of only 4.46 from one class and 4.04 in another class, while the college average was 4.76 (Emery et al., 2003).

Similarly, even though marks were returned to students already in the very next class after the examination (i.e., at the earliest possible date), student evaluations failed to reflect this (Sproule, 2000). On the item "Work returned reasonably promptly", only 50% of students gave a maximum rating of 5, whilst 27.7% rated 4, and 22.2% provided a rating of 3.

These surprising findings suggest that students either struggled with knowing how to fill the questionnaires appropriately or exerted little effort into completing them. Students could have also assumed that they simply must answer all the items. However, the surveys in all three studies (Ashby et al., 2011; Emery et al., 2003; Uijtdehaage & O'Neal, 2015) included a "non-applicable" answer category.

Alternatively, when participants were asked to rate items that were non-applicable, they may have assumed they had misremembered relevant information. This seems possible because the general number of participants who provided ratings dropped considerably when researchers included a photograph of the lecturer. Furthermore, because attending the lectures was voluntary and participants had access to approximately

75% of lectures' content through podcasts, some participants may not have attended classes at all and thus not know the identity of the lecturer (Uijtdehaage & O'Neal, 2015). This may, however, indicate that some students may complete evaluations without experiencing any lectures.

Participants may have also completed SETs "mindlessly", with little effort or thought (Uijtdehaage & O'Neal, 2015). These "mindless" ratings are more likely to occur if a) students perceive evaluations as a routine and unpleasant task, b) students do not see the potential impact of their evaluations or c) the activity is cognitively taxing and potentially difficult for students to engage in after weeks of classes (Dunegan & Hrivnak, 2003; Uijtdehaage & O'Neal, 2015).

Regarding students' perceptions of SETs, most participants reported they see SETs positively (e.g., Kite et al., 2015; Spooren & Christiaens, 2017) and did not mind completing evaluations (Spencer & Schmelkin, 2002).However, students who participated and completed these studies may have not represented general student population.

In terms of perceived impact of SETs, participants indeed stated that their primary motivation for completing SETs is a belief that their feedback may lead to subsequent improvement in teaching (Y. Chen & Hoshower, 2003), or that lecturers use or at least consider this feedback (Iqbal et al., 2016). However, participants also reported doubts about whether administrators and instructors consider their evaluations (Spencer & Schmelkin, 2002; Surratt & Desselle, 2007). Considering that individuals are likely to exert the least effort unless they have the motivation to be precise (Heilman, 2012), this could explain why some students may put less effort into filling out the questionnaires.

Regarding cognitive demands related to SETs, Dunegan and Hrivnak (2003) highlighted that students tend to complete SETs at the end of the term. Students likely have to prepare for exams and write their assessments during this time and may not dedicate as much effort to completing SETs as they would during a quieter time. In terms of empirical evidence, there is a lack of research addressing the amount of effort that students put into completing their evaluations (Wachtel, 1998). In an older study, over half of participants reported they attempt to be precise and spend sufficient time when completing SETs (Marlin, 1987). However, this again only accounts for those participants who reported this information. Perhaps some students find SETs more cognitively taxing compared to others and therefore complete them with less accuracy.

In sum, the information that students entered into SETs was inconsistent with verifiable evidence. This could be explained either by lack of perceived impact of SETs or

students' effort; the difficulty of the task, potentially due to completing SETs during the cognitively taxing time; or combination of all these factors. In either case, students, perhaps accidentally, encoded inaccurate information. When even ratings that are traceable to verifiable evidence fail to be accurate, thus it could be argued that the validity of other judgements, which cannot be as easily evidenced, should be questioned (Sproule, 2000).

All the explanations for the findings covered in this sub-section assume students entered the information inconsistent with verifiable evidence accidentally and were unaware of doing so. In the next sub-section, I discuss empirical findings about the information entered into SETs that may be intentionally inaccurate or exaggerated (Clayson & Haley, 2011; *see 2.1.2*).

2.1.2 Altered information in SETs as reported by students

Other studies investigated whether students consider any factors, which are unrelated to teaching, during completing of SETs. In a survey of U.S. participants from three universities, 60% of the participants reported they altered their evaluations because of the personality of the lecturer and 30% of the participants because of the grades they received or according to the difficulty of the exams (Clayson, 2005).

This issue was explored further with 219 participants with marketing business backgrounds, who answered questions about their opinions of SETs (Clayson & Haley, 2011). Most participants (56%) reported that they knew someone who rated their lecturer with a higher or lower score than justified and more than a third (31%) admitted to doing so themselves. A considerable proportion of participants (41%) said that they knew someone who wrote untrue comments, and almost one-fifth of participants (19%) reported that they did so themselves. Overall, participants estimated that around one-third of the SET results might contain inaccurate (verifiable) information. Participants admitted to providing imprecise information in order to protect a lecturer (3.9%) or because they liked a lecturer (11.2%). They also admitted to writing untrue comments to hurt a lecturer (2.6%) or because they disliked a lecturer (12.9%; Clayson & Haley, 2011). The actual percentages could be even higher. Some participants may have under-reported "bad behaviour", such as lying on the SET forms, due to *social desirability bias*, which refers to participants amending their responses to appear more favourably (Fisher, 1993).

These findings indicate some students may intentionally provide inaccurate information when completing SETs (Clayson & Haley, 2011). However, participants did

not specify whether they thought they were answering accurately during the completion of SETs, but later, upon further reflection, realised their answers were inaccurate or whether they intentionally provided wrong information at the time of completion of SETs. The former option would mean that this inaccuracy was unintentional, whereas, in the latter option, participants would have been aware of providing inaccurate information.

According to Lindahl and Unger (2010), students usually complete SETs at the end of the year, frequently during states of heightened emotional arousal, under the assurance of anonymity. These conditions could lead to students experiencing *deindividuation,* which refers to people feeling like a part of a group and not feeling personally accountable for their behaviours (Zimbardo, 1969). Students may then be more likely to engage in less moral actions, such as writing cruel comments about their lecturers (Lindahl & Unger, 2010), exaggerating, or intentionally providing inaccurate information. Unlike anonymity, accountability can motivate people to be more accurate in their evaluations (Mero et al., 2003) and can even reduce their stereotype-based expectations (Heilman, 2012). Some support for deindividuation affecting SETs comes from the U.S. study, in which participants reported that the group's opinion of their lecturer influenced their rating, even if they personally felt differently (Mortenson & Sathe, 2017).

The research discussed in this sub-section focused solely on information that students reported as intentionally inaccurate, but much information that students put into SETs may be inaccurate due to errors or misperceptions (e.g., Clayson & Haley, 2011, *see 2.1.1*). Therefore, students may not even be aware of providing incorrect information.

2.1.3 Low response rate

Another crucial problem with SETs is the low rate of student responses. Generally, the reported response rates vary between 30% and 70% (Hoel & Dahl, 2019). The rate is especially low for electronic and online evaluations (Al-Maamari, 2015; Nulty, 2008). This may be highly relevant during the COVID-19 pandemic because numerous universities (e.g., in the U.K.) transferred to solely online teaching. Therefore, students must have completed SETs online.

Low response rates may lead to a lack of adequate information to assess teaching effectiveness. There is also a high risk of sampling error, which means student responses may not represent those of other students (Berk, 2012). Specifically, it is unclear whether student feedback represents general opinions of the class or only the perception of

students with extremely positive or negative opinions (Hoel & Dahl, 2019). This may apply especially when the class size is small (Chapman & Joines, 2017).

Lecturers may get penalised for a low rate of student responses, even though this low rate can be influenced by various factors, such as student absence from class or even technical issues (Berk, 2012; Hornstein, 2017). Student attendance in class – and consequently, participation in SETs if they are conducted in the classroom – can depend on the time of the lectures, which is often out of the lecturers' control. Alternatively, lecturers may have provided such detailed content on an online platform or "lecture capture" recording obligatory at some universities that students deem unnecessary to attend their lectures (Stark & Freishtat, 2014). If numerous students do not attend at the time of completing SETs, the response rate will be low.

Lecturers have a very weak influence on students' intentions to complete SETs and are the least effective predictor in SETs completion (Weng et al., 2014). Specifically, the influence of lecturers had the lowest effect on participants' intentions to complete online SETs compared to the other predictors (e.g., participants' attitudes towards SETs or peer influence). In sum, lecturers may be penalised due to the factors they cannot control, and even despite doing a thorough job (Stark & Freishtat, 2014).

2.2 Conceptualisation of teaching quality

2.2.1 Definitions of teaching quality

Researchers have not yet reached a clear consensus on a definition of 'teaching quality' (Cochran-Smith, 2021; Maruli, 2014) and definitions vary across different educational bodies and contexts (Darling-Hammond, 2021). In the Teaching Excellence Framework, 'teaching quality' is described in relation to teaching practices. It is assessed against the following criteria[2]: a) teaching that engages and challenges students, b) a course design that enables students to develop their knowledge and fulfil their potential, c) lecturers who provide effective assessment and feedback that supports students' progression (TEF, 2016). Darling-Hammond (2009) defines teaching quality as a strong instruction that enables student learning. Drawing from both definitions, I conceptualise 'teaching quality' similarly as the quality of instruction and teaching practices that engage and challenge students and stimulate their learning. However, I also consider the

[2] This framework includes criterion d), institutions recognise and reward excellent teaching. However, although this factor may influence teaching quality, students are unlikely to reflect on it in SETs and I do not consider it here.

effectiveness of the ways in which lecturers interact with their students. For instance, when completing SETs, often perceived as evaluations of teaching quality, students may focus on lecturers' interpersonal behaviours (e.g., lecturers' responses to students' questions). In this PhD research, I explore which specific teaching behaviours students consider salient when evaluating their lecturers (*see Chapter 5*).

2.2.2 Differences between definitions of 'teaching quality', 'teacher quality' and 'teaching effectiveness'

There is still confusion regarding terminology and differences between 'teaching quality', 'teacher quality' and 'teaching effectiveness' (Tikrity, 2023), which are frequently used interchangeably (Bradford et al., 2021). However, these terms differ in their meanings. For example, 'teaching quality' could be distinguished from 'teacher quality'. Whereas 'teacher quality' encompasses interpersonal characteristics of lecturers, 'teaching quality' emphasises practice and effectiveness of lecturers during lectures, without considering the qualities of lecturers outside the classroom (which are, in contrast, considered in 'teacher quality'; Cochran-Smith, 2021; Towers et al., 2023). 'Teacher quality' incorporates the ways in which lecturers develop materials or curriculums, and their own teaching practices. The term 'teacher quality' can, therefore, be seen as broader in scope compared to 'teaching quality' (Snoek, 2021; Towers et al., 2023). However, this may also involve aspects that lecturers cannot influence, such as specific circumstances or rules imposed by institutions (e.g., in terms of required curriculums). In sum, although both concepts are related, 'teacher quality' predominantly focuses on aspects of the "who" (lecturers; a focus is on interpersonal *characteristics* rather than specific practices) whereas 'teaching quality' emphasises 'how and what' (practice during lectures; Bradford et al., 2021). In comparison, 'teaching effectiveness' is often defined as a lecturer's ability to contribute to or facilitate student learning (Boring, 2017; Shadreck & Isaac, 2012) but could also be seen as a sub-construct of 'teaching quality' (*see also Chapter 3*). Regarding student evaluations conducted in university contexts, most researchers discuss 'teaching quality' or 'teaching effectiveness' whereas the term 'teacher quality' (frequently used in pre-university contexts) is relatively rare. These concepts are used in dynamic and context-dependent ways, which may hinder, if not prevent, establishing consistent, clear definitions across national or educational bodies (Cochran-Smith, 2021; Towers et al., 2023).

Recent research revealed that 'teacher quality' may include components that are crucial to student development but overlooked in formal policy definitions (Towers et al.,

2023). For example, instructors used creative ways to develop innovative and inclusive approaches to tackle challenges related to online learning during the COVID-19 pandemic. This suggests that the focus should be moved away from narrow definitions of quality in policies, framed as a problem to be solved, and instead emphasise enhancing instructors' skills in order to enrich students' learning experiences (Hardy et al., 2021; Towers et al., 2023).

2.3 Factors unrelated to teaching quality that influence SETs

2.3.1 Lecturers' characteristics

SETs may also be affected by other factors that are unrelated to teaching quality, such as lecturers' fluency, mannerisms, and charisma. When asked to watch one of two videos of a lecturer with either low or high fluency, participants perceived they have learned more when taught by a fluent lecturer. However, comparisons of learning achievements between these groups of participants revealed both groups learned the same amount of information regardless of lecturers' fluency (Carpenter et al., 2013). Similarly, participants who watched videos with fluent lecturers reported higher confidence in their learning, which was, however, unsupported by their actual progress in learning (Toftness et al., 2018). This suggests that students may not always be aware of what factors actually improve their learning. Instead, when producing their ratings, students may focus on superficial characteristics, which are easy to perceive.

Another factor that may impact SETs is the lecturer‘s personality, specifically, student perceptions of their lecturers' personalities. Although lecturers' self-perception of their traits mostly revealed either low or no relationship with SETs (Feldman, 1986), participants‘ perception of their lecturers' personalities significantly correlated with the SET scores that participants provided for their lecturers (Clayson & Sheffet, 2006; Feldman, 1986).

SET scales may even simply assess lecturers' likability. Participants‘ perception of a lecturer after five minutes of interacting with this lecturer significantly correlated with SET scores obtained several months later (Clayson & Sheffet, 2006). The analysis of an online website on which students evaluated their lecturers revealed lecturers were more likely to obtain high ratings if students perceived them as nice, caring, understanding and helpful (Davison & Price, 2009), and therefore, more likeable.

Furthermore, even lecturers' non-verbal behaviours, such as smiling or mannerisms, may affect SETs (Merritt, 2008). In the U.S., the factor analysis of responses

of over 42,000 students revealed their perceptions were most influenced by a factor defined by researchers as lecturers' charisma or popularity (Coats et al., 1972). Similarly, in a U.K. study with 199 undergraduate students, participants' perception of their lecturer's charisma (assessed by the item statement "The lecturer has charisma") explained 37% of the variation in the "module attributes" and 69% of the variation in "lecturer's ability" (Shevlin et al., 2000).

In the "Dr Fox" study, an actor taught a topic he had no expertise in, trying to apply a charismatic manner, although the content was meaningless. Participants, who consisted of both professionals from Psychology, Psychiatry and Philosophy backgrounds, as well as graduate university students, evaluated his lecture favourably (Naftulin et al., 1973). However, the actor delivered a lecture about Mathematics. The results may have been different if the taught topic would have been in the participants' areas of expertise. Nevertheless, the participants were all educated and familiar with academic standards (Naftulin et al., 1973). In contrast, students completing SETs are, in most cases, not experts in the subject. Hence, students may be influenced by the lecturers' charisma and rate more charismatic lecturers more highly than the lecturers with less charismatic manners.

Lecturers' enthusiasm seems to affect the SET scores even on item statements related to other factors. The lecturer attempted to teach the class in two consecutive semesters identically, but acted in a more enthusiastic manner in one of the semesters (Williams & Ceci, 1997). The lecturer's SET scores between the first and second semesters significantly differed, even scores on the items entirely unrelated to enthusiasm, such as organisation, knowledge, tolerance, and accessibility, even though the lecturer attempted to keep these elements identical between semesters. But an enthusiastic presentation style seemed to have no significant effect on student learning because their results in exams in both terms were nearly identical (Williams & Ceci, 1997).

2.3.2 Other influences on SET scores unrelated to lecturers

Even though lecturers' charisma or enthusiasm affects SET scores, they do not seem to influence student learning. Perhaps students find enthusiastic lecturers more interesting or motivational, which results in students awarding these lecturers with higher SET scores. Alternatively, it could be questioned whether assessments of students' learning (e.g., through exam scores) appropriately assessed what students actually learned.

However, even practices entirely unrelated to lecturers, student learning and teaching quality, such as weather or the type of subject taught, may affect SETs scores.

Students were more likely to provide lower ratings on cold and rainy days (Braga et al., 2014). SET scores may also be influenced by the type of the subject lecturer teaches. The analysis of archival data at the university in the U.S. revealed that students rated their Maths lecturers much lower than their English, History and Psychology lecturers (Uttl & Smibert, 2017). In a German study, undergraduate medical students (who did not know they were participating in this study) were randomly distributed into twenty groups, half of them were provided with cookies, and the other formed a control group (Hessler et al., 2018). Participants provided with cookies evaluated lecturers significantly higher than the control group. However, both the course content and the lecturer were the same. Similarly, offering chocolate to participants before they completed SETs increased lecturers' SET scores (Youmans & Jee, 2007). Perhaps participants felt an obligation to reciprocate or eating chocolate put them in a better mood (Hessler et al., 2018; Youmans & Jee, 2007).

All these factors are unrelated to lecturers, teaching quality or teaching effectiveness. However, these factors may contribute to students' enjoyment of the classes. A lecturer with high charisma or enthusiasm may make the class more interesting for students. Sunny days or lecturers who provide students with cookies may also invoke students' enjoyment of the classes.

Thus, a factor unrelated to teaching quality that impacts SET scores could be students' enjoyment of their lectures or simply student satisfaction. Importantly, student expectations may differ from academic values (Wong & Chiu, 2019). In the U.K., E.U. and U.K. undergraduate students pay tuition fees up to £9,250 per year; International students pay even more. However, tuition fees differ between Scotland, Wales, Northern Ireland, and England. Specifically, all Scottish students (and E.U. students who started in or before the 2020/2021 academic year), are not obliged to pay any fees for undergraduate degrees at universities in Scotland. Furthermore, in Wales, tuition fees were capped at £9000 (instead of £9250) for U.K. and E.U. students. Requiring high tuition fees may lead to universities treating students as consumers, which can increase students' demands and expectations. For example, in a recent qualitative study in London, most student participants reported they see themselves as customers and expect to be treated well (Adisa et al., 2023). Some student participants also reported they base their SETs on grades they received. Therefore, universities may prioritise student satisfaction instead of

student learning, which could have a detrimental effect on students and the quality of the educational system in general.

Experiences of first-year undergraduate students rarely aligned with their original expectations of university life and were unrealistically high (Smith & Wertlieb, 2011). In a cross-cultural study with over 500 Economics university students from ten different countries (e.g., Armenia, Germany, Mexico, Poland), over 60% of participants reported universities did not meet their expectations (Mamica & Mazur, 2020). In interviews with 30 lecturers in England, many of them reported they had to incorporate the role of an entertainer into their teaching (Wong & Chiu, 2019). However, the main goal of university teaching should not be to entertain students but to assist them in broadening their expertise in the subject (Marimon et al., 2020).

Hence, lecturers who dedicate time to activities that improve student satisfaction (even if entirely unrelated to learning) may obtain higher SET scores than those lecturers who prefer to challenge their students (Hessler et al., 2018). Lecturers who provide difficult exams and content may encourage deeper student learning but fail to make students satisfied, which may negatively affect their SET scores (Marks, 2000). These lecturers may, therefore, be unfairly penalised (Carrell & West, 2010; Simpson & Siguaw, 2000).

2.4 Gender bias and gender stereotypes in education and SETs

2.4.1 Gender and sex

Psychologists did not yet reach a clear consensus on terminology regarding sex and gender, which is complex and still controversial (Hyde et al., 2019). I adopt the terminology that uses *sex* to mean a biological sex assignment at birth, and *gender* to mean self-identity or roles, behaviours, feelings, and attitudes associated with one's sex (e.g., Morgenroth & Ryan, 2018)[3]. In my research, I explored how students perceived their lecturer's *gender*. Specifically, my assumption in line with the cognitive miser (Corcoran & Mussweiler, 2010; s*ee also 2.1.3.4*) is that students categorise their lecturers into groups based on their gender (e.g., man, woman). Therefore, I assume students may not be aware of the sex, i.e., referring to a biological sex assignment of their lecturers, and I do not examine it in this thesis.

[3] Some authors use „sex/gender" to highlight that biological and sociocultural factors cannot be fully separated and sex can also represent a social construct (Hyde et al., 2019; Morgenroth et al., 2021; Morgenroth & Ryan, 2021).

Furthermore, for a long time researchers conceptualised gender as binary (men versus women) and deriving directly from biological sex, but these views are slowly evolving, leading to more awareness and support for non-binary (not identifying as men or women exclusively) individuals (Hyde et al., 2019; Morgenroth et al., 2021; Morgenroth & Ryan, 2018, 2021). I adopted a simplified binary approach and considered only male and female lecturers because investigating stereotypes against non-binary lecturers was beyond the scope of this PhD research. However, I fully acknowledge the complexity of gender as well as non-binary gender identities[4]. I highlight this point again in the final discussion.

2.4.2 Gender bias and stereotypes in educational settings

Societal influences may impact individuals from an early age. When asked to draw a scientist, only girls produced a picture of a female scientist. Even though 49% of participants were girls, there were only 28 pictures of female scientists, which formed less than 0.58% of the sample (Chambers, 1983). This implies that children, especially boys, may imagine a typical scientist as a man. However, the participants consisted of children from 5 to 11 years. Perhaps these children did not yet encounter a variety of scientists at this age. Furthermore, this study was conducted 40 years ago, and children nowadays may hold different views.

Nevertheless, when examining SETs at a high school level, the SETs of biology, physics and chemistry teachers in the U. S. revealed male participants rated female teachers lower than male teachers in all three disciplines. In contrast, female participants rated female teachers less positively only in physics (Potvin et al., 2009). To check whether these differences may be due to different teaching styles, rather than discrimination on the basis of gender, the researchers controlled for numerous factors related to teaching style, such as a number of demonstrations and frequency of class discussions. But teacher gender as well as the interaction of student-teacher gender, remained significant. Specifically, male participants still rated female teachers significantly lower than male teachers in all disciplines, whereas female participants rated them lower in physics. A test for any differences in teaching effectiveness showed no differences in students' later performance in the subject in regards to their teachers' gender (Potvin et al., 2009).

[4] In line with this view, I included three categories for gender of participants (men, women, or non-binary participants).

These findings imply even children may be affected by gender stereotypes and are in line with the previous findings (e.g., Bian et al., 2017). These influences do not disappear with age. One of the societal influences perpetuating gender stereotypes could be a general representation of women in educational materials, such as textbooks (Peterson & Kroner, 1992).

An analysis revealed the prominent presence of gender bias, specifically portraying women in stereotypical, negative ways, and women being underrepresented in psychology textbooks. This included textbooks expanding on the work of female psychologists to a much lesser extent compared to the work of male psychologists and portraying women (versus men) significantly more frequently as clients and significantly less frequently as therapists (Peterson & Kroner, 1992). These findings are surprising because women are often over-represented at the undergraduate psychology level. For example, in the academic year 2021/2022, 81% of undergraduate Psychology students in the U.K. were women (HESA, 2023). Because Peterson and Kroner (1992) conducted their research over 30 years ago, it could be argued that the representation of gender in psychology textbooks has changed since then. However, a study undertaken fifteen years later yielded similar results (Collins & Hebert, 2008). The analysis of over 3000 images of individuals from psychology textbooks published between 1999 and 2007 revealed that men were depicted significantly more frequently (59%) compared to women (41%).

Even materials for lecturers may inadvertently promote gender bias. A striking example is a guide for lecturers on how to improve their SET ratings by simply advising them to "Be male" (Neath, 1996, p.1364). Specifically, female lecturers were encouraged to change their gender or, if opposed to doing so, at least change their behaviour to be less feminine in order to boost their ratings (Bray & Howard, 1980). The guide stated: "If you are female, do not be very demanding of your students" (Neath, 1996, p. 1365). This questionable approach takes for granted that male and female lecturers are treated differently for the same behaviour ("being demanding of their students"). Alternatively, the author perhaps intended to draw attention towards sexism in academia with a humorous remark. Considering this guide was published over 25 years ago, this again highlights that sexism in academia is an ongoing problem with historical roots. For example, in the U.K., women started entering academia only in the late 19th century. Other scientific institutions such as The Royal Society, the University of Cambridge or Harvard Medical School did not admit women on the same conditions as men until the 1940s

(Saini, 2017). To examine whether these problems persist in the current era, I also discuss more recent research related to sexism in academia.

For instance, a recently conducted study (Fisk et al., 2020) introduced a lightweight intervention in order to reduce gender bias against female lecturers in STEM. This intervention consisted of a female professor sending students an email with positive feedback. Although authors acknowledged that, ideally, women should not have to change their behaviour to conquer gender bias in SETs, they still introduced this method as a strategy to increase female lecturers' SET scores. Whereas these findings may be helpful for some female lecturers, the focus is on dealing with the symptom rather than the actual root of the problem. This strategy enforces the idea that female lecturers should conform to and comply with gendered expectations in lieu of students adhering to university standards. Furthermore, female lecturers are asked to perform additional work to obtain the same ratings as men. Recommendations like these put the responsibility for students' gendered attitudes solely on female lecturers, who are asked to alter their behaviours, instead of exploring alternative solutions and tackling the existing inequalities.

Not only students but also faculty may be biased in their behaviours towards lecturers, which implies that a wider scope of gender stereotypes may exist in academia. For example, a qualitative case study revealed that, when making hiring decisions, faculty committees in the humanities, social sciences, and natural sciences considered the relationship status of female but not male candidates (Rivera, 2017). The committees assumed that female candidates with partners who had senior positions in academia or other prestigious jobs would be unwilling to relocate. Strikingly, even though considering marital status when making a hiring decision was illegal and a violation of university policy, the committee openly discussed it in front of the researcher, who was present during these discussions. If faculty committees felt comfortable behaving in a biased way in front of a researcher, even stating they should not discuss candidates' relationship statuses, and then doing so (Rivera, 2017), one must question whether these behaviours at institutions may be amplified when there is no external person present. This demonstrates that women, unlike men, may need to fulfil additional requirements (a certain relationship status) to get the role. The faculty's approach may even imply beliefs that female lecturers' jobs were less important than those of their male partners. Nevertheless, this research was conducted in the U.S., where further cultural context must be considered. Specifically, two academics in a relationship may find it challenging to find two academic roles in proximity. Therefore, institutions worldwide developed formal dual-career

policies and programmes, also very commonly used in the U.S. (Tzanakou, 2017). The U.S. university at which this research was conducted lacked such policy. However, if committees were to consider the relationship status of candidates solely because of the lack of this policy, both men and women would likely be equally affected (Rivera, 2017).

One could also question whether gender bias could be less prominent in psychology departments, considering that because of the nature of their field, academics in Psychology may be more aware of possible biases compared to academics in other subjects. However, when 238 male and female academics in Psychology evaluated an identical curriculum vitae, which only differed by a male or female name, they were significantly more supportive of hiring a male versus female applicant, and more likely to indicate that male but not female applicant had sufficient research experience (Steinpreis et al., 1999). Therefore, despite the awareness of this issue, individuals may still behave in biased ways.

In terms of faculty response towards gender bias, British academic staff provided their perceptions regarding students' views and behaviours concerning lecturers' gender (Carson, 2001). All but two female lecturers commented on gender-related issues related to gender affecting their professional life. The female lecturers reported that if they tried to implement "firm" behaviour, students perceived them as "bossy". The female lecturers assumed students expected them to act more caring and provide more student contact compared to male lecturers. Most of the female lecturers suspected student evaluations of being gender biased against females (Carson, 2001). Consistently, the perceptions of teaching faculty regarding SETs hardly differed by lecturers' position or seniority but were strongly associated with their gender. Female lecturers reported feeling impacted by SETs more negatively than male lecturers and feeling angry or unhappy when reading evaluations more often than their male colleagues (Kogan et al., 2010).

However, because Carson's (2001) research is almost 20 years old, it could be argued that it represents somewhat outdated views and female lecturers are less likely to experience sexism nowadays. Unfortunately, abundant recent research suggests that sexist practices still occur or that gender equality is not viewed as a priority (e.g., Bocher et al., 2020; Bono et al., 2019; Bourabain, 2021; Morris et al., 2022). A project designed to collect and draw sexist experiences in academia yielded 89 accepted testimonies, out of which women wrote 97.8% and men 2.2% (Bocher et al., 2020). In a narrative piece, three female lecturers in geography departments reported sexist experiences, such as increased teaching duties but decreased proposed salary after informing a department chair of their

pregnancy (Bono et al., 2019). In another study, during in-depth interviews of 50 female early-career researchers (e.g., PhD students, postdoctoral researchers) in Belgium, many shared their experienced incidents of sexism, such as institutions or colleagues dismissing sexual remarks, harassment, or, appallingly, sexual assault, directed at them, female students or their colleagues. They also reported being discouraged to pursue a PhD because of maternity, and being undervalued, e.g., individuals did not use their academic titles when corresponding but did so for their male colleagues (Bourabain, 2021). Similarly, the academic staff in the U.K. reported during in-depth interviews that White cis straight men tended to dominate even in departments with a high proportion of women and more prestigious positions were usually occupied by men (Morris et al., 2022). These reports of sexism are further supported by empirical evidence. For example, the analysis of the anonymous comments that over 670 academics obtained during SETs revealed that female academics and academics from marginalised groups received the most frequent as well as the most derogatory abusive comments (Heffernan, 2023).

Therefore, even more recent studies show evidence of sexism. However, findings can be influenced by further contextual factors, such as academic subject (*see 2.4.6*) or culture. For example, two recently conducted experimental studies in Denmark revealed no evidence of gender bias in student evaluations of lecturers (Binderkrantz et al., 2022). These encouraging findings could be explained by cultural context, because Scandinavian countries such as Denmark are generally performing well in promoting gender equality (Husu, 2019). However, another study conducted in Iceland, a country associated with being at the forefront of gender equality, revealed some students may hold gender-related expectations of their lecturers when completing SETs (Sigurdardottir et al., 2023). It must be acknowledged that although gender-related problems persist, gender equality has improved over the last 30 years, with increasing numbers of women attending universities and working in academia (Stepan-Norris & Kerrissey, 2016). Furthermore, nowadays there are increasingly more initiatives aimed at improving gender equality or promoting women's education (e.g., programs designed to encourage female participation in STEM fields). All conducted studies depend on numerous contextual factors, such as specific institutions, culture, or individual differences between participants. Therefore, contradictions may be expected, and no findings can be completely generalised.

In sum, evidence suggests sexism and gender bias may still be a relevant problem in many educational settings (high schools, universities) and across different departments (*see also 2.2.6*), perpetuated not only by students but also by faculty themselves. Some

proposed a comprehensive list of possible solutions to combat gender bias on both individual level, such as ensuring one is familiarised with best hiring practices if in charge of hiring, but also an institutional level, such as increasing diversity of editorial panels or offering on-site childcare (e.g., Llorens et al., 2021). Importantly, the authors highlight that these are merely suggestions worth consideration that may not work in isolation and permanent solutions require both cultural and structural changes.

2.4.3 General overview of different biases in SETs: empirical findings

One of the most crucial problems with SETs concerns biases and stereotypes that students can be influenced by during the evaluation of their lecturers. This may consequently affect their SET ratings.

Students tended to rate female lecturers lower than male lecturers (e.g., Arbuckle & Williams, 2003; Boring, 2017; Fan et al., 2019; Mengel et al., 2017). However, it was unknown whether these differences in the ratings could be attributed to bias or simply to differences in teaching effectiveness between genders (MacNell et al., 2015). This was explored in a U.S. study, in which researchers manipulated the presentation of the lecturers' gender in an online course. Each lecturer taught two groups, one under their own identity and one under a false identity (of the opposite gender). The lecturers delivered sessions through a learning management system, and participants only had contact with lecturers through emails or written comments. Participants were significantly more likely to award higher ratings when they perceived their lecturer to be male (MacNell et al., 2015).

In a similar study informed by Macnell et al. (2015), researchers assigned participants to either an alleged male or a female teaching assistant, but, in fact, the same female teaching assistant completed the work for both groups (Khazan et al., 2019). Although participants awarded both assistants positive evaluations and the means of female and male assistants did not significantly differ, data showed some disturbing trends. Participants provided five times as many negative comments when female teaching assistant taught under her own identity compared to a false (male) identity, although it was the same person (Khazan et al., 2019). However, the sample size was low in both studies, ranging from eight to thirty-six student participants per condition. Results inferred from small sample sizes may be susceptible to inflated discovery rate, smaller power and high sensitivity to outliers (Ioannidis, 2005) and lead to incorrect conclusions (Uttl & Violo, 2021).

Further studies showed participants rated female lecturers lower than male lecturers for identical behaviours (Arbuckle & Williams, 2003; MacNell et al., 2015) or using identical teaching materials (Mengel et al., 2019; Mitchell & Martin, 2018). Specifically, when teaching identical online political science courses, male lecturers received either the same or higher scores compared to female lecturers on all 23 item scales. Male lecturers obtained higher scores than female lecturers even on the items unrelated to lecturers' behaviours, such as relevance and usefulness of course materials, meaning these scores, therefore, could not be attributed to differences in teaching abilities (Mitchell & Martin, 2018). Thus, because male and female lecturers were evaluated on the same actions, the differences in their SET scores may have been influenced by the perceived gender of the lecturer and thus biased.

I adopt the view from social psychology that claims that prejudice may be associated with different forms of discriminatory behaviour (e.g., Fiske, 2000). Students can, therefore, discriminate against their lecturers due to characteristics unrelated to gender. For example, students may rate lecturers who are members of ethnic minorities more harshly than White lecturers. After reading identical C.V.s that only differed in the ethnicity of the lecturers, participants rated Black professors as significantly less competent than Asian professors with an average difference between mean scores of 0.20. Participants also perceived ethnic minority lecturers to have lower interpersonal skills compared to White lecturers, as shown by difference of 0.28 between mean scores of White and Black lecturers and 0.27 between White and Asian lecturers (Bavishi et al., 2010). Similarly, in the online courses, where the only student-professor interaction consisted of the welcome video and the content was otherwise identical, students rated minority lecturers less favourably compared to White lecturers (Chávez & Mitchell, 2020). However, Chavez and Mitchell (2020) used a sample size of only 48 student participants, drawn from the same institution. Furthermore, an initial t-test did not yield a significant difference between student evaluations of ethnic minority lecturers and White lecturers. Authors proceeded with a regression analysis that controlled for grades but did not specify whether this was planned a priori. Researchers also chose $p < 0.1$ as their significance level without justifying this choice.

Students may also be biased against lecturers because of lecturers' sexual orientation. When male homosexual lecturers delivered a strong lecture, participants rated them more negatively than those lecturers whose sexual orientation was unspecified (even though this was not the case when the delivered lecture was weak) with a difference of

0.60 between mean scores. This suggests students may under certain conditions rate homosexual lecturers more strictly than heterosexual lecturers (Ewing et al., 2003).

Students evaluated those with a non-English speaking background less positively across nearly all faculties, as revealed by approximately 523,000 surveys collected at an Australian university over seven years. Male lecturers with English as their native language were most likely to obtain the highest ratings in all faculties, apart from Engineering (Fan et al., 2019). These findings imply that SET ratings may be *biased* and discriminate towards certain groups (Mitchell & Martin, 2018).

It is crucial to investigate this matter because SETs often form a significant factor in career-altering decisions for lecturers (Basow & Martin, 2012). Specifically, SET scores may form an important factor in decisions concerning their promotion, retention and salary (Benton & Cashin, 2014; Emery et al., 2003). However, if SETs discriminate against certain groups, using them in H.R. decisions may be illegal (Hornstein, 2017). Therefore, institutions must ensure that SETs are valid and that lecturers are treated equally, such as in terms of gender.

2.4.4 Theoretical framework: social psychological theories

Prejudice and discrimination have been widely investigated in psychology. To address the potential influence of gender stereotypes on student judgements in SETs, I apply theoretical frameworks that are consistent with mainstream social cognitive psychology, and predominantly developed by Susan Fiske. Specifically, I discuss the *stereotype content model* (Fiske et al., 2002), *ambivalent sexism* (Glick & Fiske, 1996), a *role congruity theory* (Eagly & Karau, 2002), the *status incongruity hypothesis* (Rudman et al., 2012), and *the cognitive miser theory* (Fiske & Taylor, 1991; S. E. Taylor, 1981).

Bias can be defined as an attitude, belief, or prejudice for or against a specific category of individuals based on their characteristics (Molinari et al., 2019). Bias can be *explicit*, which refers to bias that an individual is aware of and may even embrace (Boysen, 2009) or *implicit*, which people may be aware of to some extent and which can occur when they create associations towards individuals with certain characteristics (Ashford et al., 2018). Importantly, although there may be no intent to cause harm, individuals may do so by their actions, as bias can affect one's behaviours, judgements, or evaluations (P. L. Carter et al., 2017).

When people produce behavioural ratings, which are ratings related to the behaviours of other individuals, such as work performance, their expectations can

influence their ratings more than actual experiences (Heilman, 2012). This means that if raters believe they are evaluating someone who is a good leader, they may provide higher ratings on related dimensions; regardless of what behaviours this rated individual displays (Baltes & Parker, 2000; Binning et al., 1986). However, biased expectations will lead to biased evaluations, because raters may think of behaviours of the target individuals that are consistent with raters' expectations and dismiss those behaviours which are not (e.g., Heilman, 2012; Uher & Visalberghi, 2016; *see also Chapter 3*). This is problematic, because students may form a certain (potentially biased) impression about their lecturers and then choose a certain behavioural description to justify their impression (Basow et al., 2006).

Stereotypes can be defined as beliefs that individuals from the same group share a set of consistent attributes or behaviours (e.g., Hilton & Von Hippel, 1996; Johnson et al., 2018) or the specific expectations from the members of the same group (Ellemers, 2018). *Gender stereotypes*, therefore, refer to generalised beliefs about (or expectations of) men or women (Heilman, 2012). Gender stereotypes may influence the ways in which we perceive and judge the performances of others, leading to different judgement of identical performances between different genders (Ellemers, 2018). An infinite number of different characteristics may be a part of a stereotype, but the two crucial characteristics for predicting prejudice and discrimination are warmth and competence, associated with the stereotype content model.

The stereotype content model argues gender stereotypes may be represented by two dimensions: *competence*, stereotypically perceived as a masculine trait, and *warmth*, stereotypically perceived as a feminine trait (Fiske et al., 2002). Specifically, men are more frequently perceived as „competent" or „agentic" and associated with active, competitive, confident behaviour, whereas women are usually perceived as „warm" or „communal" and associated with nurturing, gentle or emotional behaviour (Kite et al., 2015). People assess competence to judge the abilities of other individuals and warmth to assess others' intentions (Fiske et al., 2002). Thus, because men are usually associated with competence and women with warmth, people may assume that men have stronger abilities, whereas women have better intentions compared to men (Ellemers, 2018). Importantly, the stereotype content model has four quadrants, and positive stereotypes in one dimension are not necessarily compatible with the other dimension. Therefore, attitudes towards individuals may be mixed, such as that they may be evaluated high on warmth but low on competence (Fiske et al., 2002; Menegatti et al., 2017). For example,

women in more traditional roles that may be perceived as subordinate such as housewives or secretaries would be likely seen as high in warmth, but low in competence, whereas the opposite would apply to women who hold more senior career roles (Fiske, 2012).

The theory consistent with the stereotype content model is ambivalent sexism (Glick & Fiske, 1996). *Sexism* can be defined as „discriminatory and prejudicial beliefs and practices directed against one of the sexes, usually women" (American Psychological Organisation, 2020). *Ambivalent sexism* theory claims that sexism has two components: benevolent sexism and hostile sexism. These two components exist because of hetero-normative society (in terms of social structures and beliefs about gender and sex) in which men and women live together. *Benevolent sexism* refers to benevolence towards women who conform to gender stereotypes, whereas *hostile sexism* to the judgement of those women who do not (Glick & Fiske, 1996). Benevolent sexism, therefore, promotes positive characteristics of women whose behaviours are consistent with gender roles and stereotypes and may be endorsed by both men and women (Fiske & Taylor, 2013). However, benevolent sexism is damaging because it describes women as warm but incompetent (Dardenne et al., 2007). For instance, the belief that women must be protected may be well-intentioned, but it positions women as weak (Hideg & Ferris, 2016). Benevolent sexism may initially appear helpful; however, it subtly encourages and reinforces gender roles and promotes gender stereotypes (Lamarche et al., 2020). Furthermore, dealing with benevolent sexism can be difficult, because due to its subtlety and frequent positive tones, people may fail to notice that it is a form of prejudice (Barreto & Ellemers, 2005).

Another theory about the influence of gender stereotypes is a r*ole congruity theory* that suggests female leaders are perceived as violating gender standards when exhibiting behaviour usually associated with men (Eagly & Karau, 2002). This theory heavily considers *social roles*, which include two kinds of *social norms*. Social norms are both *descriptive*, which refers to beliefs about what members of the group *do* as well as *prescriptive*, which refers to beliefs about what members of the group *should do* (J. C. Turner, 1991). The role congruity theory is partially developed from the *social role theory* (Eagly, 1987) that claims that people accumulate beliefs about social roles and norms by observations of people in these roles. These observations may eventually become stereotypical beliefs, such as *traditional gender role beliefs* (Kite et al., 2015). Importantly, people may perceive different behaviours as inherent to women or men, even though these behaviours may depend on environmental factors and context. For example,

individuals may observe women engage in child-minding activities more frequently than men. This may lead to a belief that women are associated with a warmth that enables them to occupy these roles and demonstrate related behaviours (Koenig & Eagly, 2014). Therefore, according to the role congruity theory, people may expect a woman to conform to the social norms of her gender role associated with warm, empathetic behaviour, but this may hinder her ability to conform to the norms of the leadership role associated with assertive behaviour (Eagly & Karau, 2002). Specifically, expectations of a woman's gender role contradict the expectations of a leadership role. The role congruity theory proposes that two types of prejudice may arise towards female leaders. Female leaders may be evaluated less positively a) due to leadership characteristics being commonly associated with men b) when engaging in leadership actions, because this type of behaviour is perceived as less desirable for women. Thus, female (unlike male) leaders can be seen more negatively as potential leaders and be evaluated harshly for the behaviour linked to leadership (Fiske & Taylor, 2013).

The role congruity theory shares some similarities with the *status incongruity hypothesis*, which states that women in prestigious positions may experience backlash because of perceived incongruence between their ascribed status as a woman (perceived as low) and achieved high status of the position (Rudman et al., 2012). Therefore, these women may be seen as threatening gender hierarchy, especially if they exhibit dominant or agentic behaviours. This does not apply to male leaders, whose ascribed status as a man and achieved status as a leader (both perceived as high) are congruent. This puts women in a difficult situation because they must exhibit agentic behaviours to justify their high position, but then risk that others may perceive them as violating status quo. In terms of academia, this could apply to female lecturers, especially in more prestigious (e.g., highly paid) departments (Fisher et al., 2019).

Another theory that explains how stereotypes may influence individuals' judgements is the *cognitive miser* theory, according to which individuals apply heuristics and mental shortcuts in order to avoid the depletion of resources (Fiske & Taylor, 1991; Fiske & Taylor, 2017; Taylor, 1981). To facilitate mental work, individuals rely on categorical thinking that may be based on stereotypes (Macrae & Bodenhausen, 2001). Specifically, individuals may find it easier to categorise encountered persons based on stereotypical information about the member of the category (e.g., a woman) rather than consider all available information about these persons (Corcoran & Mussweiler, 2010; Macrae et al., 1994). In sum, the cognitive miser model describes biases as inherent

aspects of cognition but, unlike the status incongruity hypothesis, does not acknowledge the role of motivation (Fiske & Taylor, 2017).

Applying these theories to an academic context, students may expect their female lecturers to display nurturing rather than assertive behaviour (Andersen & Miller, 1997; K. J. Anderson & Smith, 2005), but then perceive them as less competent (Dardenne et al., 2007). These expectations may be even more pronounced when students evaluate lecturers of a different gender than their own.

These theories informed my PhD research in several ways. I explored whether students consider specific teaching behaviours or attitudes as more salient for lecturers of a certain gender, which could mean students are influenced by gender stereotypes. I also investigated whether students rate male and female lecturers differently, either in their overall ratings or the ways in which students weight teaching behaviours in their ratings of lecturers. I also explored whether any differences in ratings emerge when controlling for any potential differences in teaching behaviours between male and female lecturers (i.e., by using standardised scenarios, *see Chapter 4*, 6). Finally, I analysed my findings through the lens of the above-discussed theories.

2.4.5 Gender-related research about SETs

The recent studies revealed that students frequently rated female lecturers lower than male (Boring, 2017; Fan et al., 2019; Martin, 2016; Mengel et al., 2019), especially when students were men (Boring, 2017; Mengel et al., 2019). However, earlier research showed female lecturers received either higher evaluations than male lecturers (Tatro, 1995) or there was no statistically significant gender difference (Cashin, 1995; Centra & Gaubatz, 2000; Feldman, 1993). Evidence suggests no differences in teaching effectiveness (as assessed by students' exam scores) between genders (Boring, 2017; Mengel et al., 2019). This contradictory evidence demonstrating both gender biases in SETs and lack thereof could be explained by the theories introduced above (Eagly & Karau, 2002; Fiske et al., 2002; Glick & Fiske, 1996).

Student expectations of their lecturers can be based on gender-role beliefs, such as expecting female lecturers to display nurturing rather than assertive behaviour (Andersen & Miller, 1997; K. J. Anderson & Smith, 2005). Consistently with ambivalent sexism theory, female, but not male, lecturers perceived as friendly received higher ratings from participants (Kierstead et al., 1988).

An analysis of data from 4,300 students at a French university contained over 20,000 observations spread out across five years (Boring, 2017). It was mandatory for students to provide these ratings. Students' higher ratings of male lecturers were associated with traditional male stereotypes, such as leadership skills, whereas higher ratings of female lecturers were related to traditional female stereotypes, such as the usefulness of feedback. Male students rated their overall satisfaction significantly higher when rating a male lecturer than when their lecturer was a woman. Overall, male lecturers were 20% more likely to be rated as "excellent" than female lecturers. These results could mean that the male lecturers at this university were simply more effective teachers than females. However, a lecturer's gender did not influence student performance on the exam, which all the students took at the end of the year. Even though the students learned the same amount of information from both lecturers, they rated male lecturers as more knowledgeable than female ones (Boring, 2017). These findings are consistent with the stereotype content model. Despite these findings, it could be questioned whether exams accurately assess student learning. However, evidence shows gender bias and stereotypes even in areas unrelated to student learning or performance (e.g., different ratings for the same materials or behaviours).

For example, an analysis of nearly 20,000 student evaluations at a Dutch university, predominantly in Business (54%) and Economics (28%) departments, revealed a significant gender bias against female lecturers. This was particularly the case if female lecturers were junior staff, working in mathematical departments, or if the students rating them were males (Mengel et al., 2019). Male students were more likely to evaluate course materials as worse if the lecturer was a woman, even though these materials were identical to those used by male lecturers. Further analysis showed that lecturers' gender had no significant effect on grades awarded or hours students spent studying (Mengel et al., 2019).

In accordance with the stereotype content model and the role congruity theory, students may expect female lecturers to offer more interpersonal support and hold them to higher standards regarding this support than male lecturers (S.K. Bennett, 1982). Male lecturers were described by participants more often than female lecturers as providing excellent course content or in terms of knowledge, whereas female lecturers in terms of individual student-lecturer interactions (Basow et al., 2006). Participants (particularly men) were more likely to describe their "worst" female lecturers as lacking interpersonal

skills. But they praised their best female lecturers for possessing interpersonal skills, especially approachability.

An analysis of data from two political science departments at two different U.S. universities showed interaction effects between lecturers' gender and the course size. The larger the course, the more likely it was for female lecturers to obtain a lower rating compared to male lecturers (Martin, 2016). These findings could be explained by the role congruity theory. Specifically, female lecturers in smaller but not large classes may have more extensive interactions with students and more opportunities to demonstrate warm or empathetic behaviour (Martin, 2016). Potentially, students may have associated large classes with a higher status of lecturers. These explanations show support for the stereotype content model and the status incongruity hypothesis.

However, although students seem to value "warmth" in their female lecturers, they may then perceive them as less competent (Heilman, 2012). Participants often referred to female lecturers as "teachers" and male lecturers as "professors", even when all lecturers had the same qualifications (J. A. Miller & Chamberlin, 2000).

The analysis of the text from an online forum mostly populated by Economics students showed that students frequently described female lecturers in terms of interpersonal characteristics or physical attributes, whereas they used academic and professional terms to describe their male lecturers (A.H. Wu, 2017). Open-ended comments in course evaluations over the span of eight years and across five STEM disciplines revealed students were more likely to refer to male lecturers by their professional titles and female lecturers by their first names (Terkik et al., 2016). Further analysis of qualitative data from two distinct sources, the RateMyProfessor website and the official course evaluations from students, revealed similar findings. Specifically, students commented more often on the lecturers' appearance and personality when the lecturer was female. They were also more likely to refer to female lecturers as "teachers" and male lecturers as "professors"(Mitchell & Martin, 2018). The focus on the interpersonal characteristics of female lecturers but the professionalism or credentials of male lecturers suggests that students valued competence in male lecturers but warmth in female lecturers, which is consistent with the stereotype content model.

These findings suggest that the role of the lecturer may be perceived by students as traditionally masculine or associated with high status. This means that students may evaluate their lecturers on criteria related to masculine stereotypes, especially when

evaluating lecturers' overall competence (Boring, 2017), which may disadvantage female lecturers. This provides support for the status incongruity hypothesis.

Thus, students seem to indeed expect their lecturers to behave in line with traditional stereotypes, expecting their female lecturers to be nurturing and their male lecturers to be knowledgeable and to provide excellent course content. However, all lecturers likely acquired years of education and qualifications before teaching at universities and possess in-depth knowledge about their subject. Therefore, it seems female lecturers are required to be knowledgeable, provide useful content and, in addition, offer interpersonal support and possess interpersonal skills in order to achieve good SET ratings.

Furthermore, female lecturers reported receiving more requests for special favours from students compared to male lecturers (El-Alayli et al., 2018). Students who felt entitled to success regardless of their performance, such as believing they should pass the class just because of good attendance, were more likely to request and expect more special favours from their female compared to their male lecturers. Female lecturers also reported they had dedicated more time to teaching, advising students and participating in other campus services compared to male lecturers, who reported having spent more time doing research (O'Meara et al., 2017).

Even institutions may engage in discriminatory behaviours. For example, female lecturers have received more work requests than male lecturers, and these requests were mostly related to non-research related activities (O'Meara et al.,2017). Conducting research and publishing is frequently a pathway to promotion (West et al., 2013). Therefore, spending time on "administrative" tasks may result in fewer opportunities for female lecturers to engage in research-related activities, which may be detrimental to their career advancement.

In sum, if women in leadership positions (such as lecturers) want to be perceived as effective, they must demonstrate a balance between sensitivity and strength, but for male leaders, it may be sufficient to only express strength (Johnson et al., 2008). In addition, female lecturers seem to spend more time on teaching and communal tasks compared to male lecturers. This suggests that women may need to work harder and cope with more demands than male lecturers to achieve the same SET scores (Basow, 1995; Carson, 2001; El-Alayli et al., 2018).

If student evaluations are gender-biased against female lecturers, the alternative solution could be to change the format in which lecturers are evaluated. Students who

received SETs which contained language intended to reduce bias rated female lecturers significantly higher than students filling out completing SETs without this anti-bias language intervention (Peterson et al., 2019). Their language intervention consisted of a statement that SETs are frequently affected by unintentional biases, and women and minority lecturers are rated lower than White men, even when there are no differences in teaching quality. Students were asked to repress the stereotypes and focus on the content of the course rather than on unrelated matters. However, diversity or anti-bias training may produce mixed results and even be counter-productive (Dobbin & Kalev, 2018). For example, previous research showed that participants continue to rely on gender stereotypes even after attempting to suppress them (Nelson et al., 1996). This effort may actually make stereotypes more cognitively accessible and subsequently activate and reinforce them (Dobbin & Kalev, 2018; Galinsky & Moskowitz, 2000; Macrae et al., 1994).

2.4.6 Discipline-specific differences in SETs

Students from different disciplines may rate their lecturers differently because of discipline-specific cultures. For instance, students in science courses may have very different expectations from their lecturers compared to students in other courses. Mathematics students may expect their lecturers to help them learn difficult information, and social science students may expect from their lecturers to be entertaining (Stroebe, 2016). Thus, students of different subjects may interpret a specific item, such as "lecturer made the subject interesting" differently. Some students could interpret this statement as "the material was challenging", whereas other students as "lecturer made jokes during lectures". Furthermore, students from certain departments may be more inclined to provide high or lower SET ratings. For instance, in a U.S. study, engineering participants provided the lowest ratings on each factor as compared to participants from humanities, social sciences, and natural sciences (Basow & Silberg, 1987).

In terms of potential gender bias, the gender of a female lecturer may be particularly salient in a "gender-atypical field", which refers to a field usually dominated by men; in departments where women are underrepresented or if she teaches women's studies (Basow, 1995). Participants rated identical scientific abstracts as having lower quality if they thought the authors were women compared to men, with the difference between estimated mean scores of 0.07, but especially if the topic contained themes traditionally perceived as masculine, such as political communication or computer-

mediated communication, with the difference between estimated mean scores being 0.26 (Knobloch-Westerwick et al., 2013). However, this significant difference may have been caused by a minority of participants rating in this direction rather than represent a general trend. In another study, only physics, but not biology professors, evaluated identical C.V.s with male names more favourably than those with female names in terms of hireability and competence (Eaton et al., 2020). The authors offer three explanations for this finding: a) more masculine culture in physics because of having a higher gender ratio imbalance compared to biology, b) 90% of participants from the physics department were men, who could have held in-group bias[5] towards women, c) a perception that physics requires a high level of mathematical ability, which may be stereotypically perceived as masculine (Eaton et al., 2020).

Similarly, when participants evaluated identical videos of the physics lecturers differing only by gender, male participants evaluated male lecturers significantly more favourably compared to female lecturers. Female participants evaluated female lecturers higher than male lecturers in terms of their interpersonal skills, but not scientific abilities (Graves et al., 2017). Female lecturers tended to receive higher ratings from female students and lower ratings from male students, particularly in Business and Engineering (Basow & Martin, 2012; Centra & Gaubatz, 2000). These findings provide further support that students may show in-group bias towards lecturers of the other gender.

Gender bias may be more prevalent in higher-paying departments, such as Business or Economics and weaker in lower-paying departments, such as English (Fisher et al., 2019), which is consistent with the role congruity theory and the status incongruity hypothesis. Specifically, students may perceive certain departments as masculine (e.g., physics, engineering), which may be reinforced by an imbalanced gender ratio. Students may therefore consider female lecturers in these departments to be violating gender standards because they teach subjects usually associated with men. Similarly, students may associate lecturers in better-paying departments with a higher status. Students may then perceive female lecturers in these departments as displaying incongruence between their status as a woman (that students may see as low) and the status of a lecturer in this department (seen as high). In sum, rating behaviours of students may differ based on departments and be influenced by the departmental culture or even a gender ratio balance.

[5] A tendency to give preferential treatment to the members of one's own group.

2.4.7 Importance of qualitative studies in SET research: stereotypes in language use

Language may frequently be the means through which stereotypes are spread because the stereotypes are codified in many languages (Menegatti & Rubini, 2017). For instance, there is no equivalent masculine form for the word "Miss", which may imply that the marital status is more salient for women compared to men. Furthermore, in English, "he" can describe a "man" or generically "a person", whereas "she" is usually only used to describe a woman (Menegatti & Rubini, 2017). This is problematic because it may perpetuate *androcentrism*, which is a specific form of bias centred around men, and privileging their experiences over those of women (Bailey et al., 2019). The authors emphasise androcentrism does not necessarily mean that men are seen as superior to women, but rather that men are perceived as prototypical humans, whereas women have to be classified as gender specific. Another example of androcentrism may be more frequent representations of men in terms of characters or roles occupied by both men and women (Bailey et al., 2019).

Qualitative research is useful for revealing biases not visible in quantitative research (Laube et al., 2007). This type of research on how students approach evaluating their lecturers on SET items is scarce, especially in the U.K. There seems to be only one U.K. study investigating student participants' interpretations of SET items (R. Bennett & Kane, 2014), another exploring the ways in which students make their rating decisions (Ashby et al., 2011), and one study investigating both (Robertson, 2004). None focused on potential biases or students' perceptions of specific behaviours and inferred attitudes of their lecturers.

In another study, male participants did not rate female lecturers significantly differently than their male lecturers. However, when asked to elaborate on differences between female and male lecturers with open-ended questions, they described their lecturer as "rigid" and framed this quality as positive for male lecturers but negative for female lecturers. Specifically, participants seemed to expect their female but not male lecturers to be flexible (Bachen et al., 1999). This implies that students may apply gender-specific schemata when evaluating their lecturers (Laube et al., 2007).

Qualitative research could, therefore, reveal hidden biases and stereotypes by investigating student comments and, specifically, words they used to describe their lecturers. According to assumptions of humans as a cognitive miser, individuals categorise others based on their social groups in order to process information quickly and effortlessly (e.g., Corcoran & Mussweiler, 2010). People may perceive the world through

the categories stored in their memories (Fiske & Taylor, 1991) and recall phenomena through the categories through which they have originally perceived these phenomena (Bem, 1981). The words that individuals choose to use may sometimes reflect their stereotypical beliefs (Eitzen & Zinn, 2016; Sprague & Massoni, 2005). For example, in recommendation letters, women are more likely to be described with communal terms such as warm, compared to men, who are more likely to be described in agentic terms, such as competitive (Madera et al., 2009). Furthermore, when writing recommendations, people use for men more often stronger adjectives, such as "brilliant" or "outstanding" than for women (Schmader et al., 2007). This provides further support for the stereotype content model because women are praised for being warm and men for being competent.

Similarly, an analysis of a large military dataset with over 4000 participants from military showed no significant differences between men and women in objective performance measures, such as military-grade point average or military ranking (D.G. Smith et al., 2018). However, women were more likely to be associated with negative attributes compared to men in subjective measures (attributes assigned by peers during evaluations). The most frequently used negative term to describe a woman was "inept", whereas the most negative attribute assigned to a man was "arrogant." People praised the women the most for being "compassionate" and men for being "analytical". Furthermore, men were described more frequently in terms of their abilities, with words such as "competent", "logical", "athletic" and women in terms of their interpersonal skills, such as "selfish", "temperamental" or "energetic" (D. G. Smith et al., 2018). This is consistent with the stereotype content model, because men were praised more frequently for their competence, whereas women for their warmth.

In the academic context, when asked to reflect on their 'worst' and 'best' lecturer, adjectives such as "giving", "compassionate", "sensitive" or "attractive" were used by participants only for their female lecturers. Although participants used the word "caring" to describe both male and female lecturers, female lecturers were described more frequently as "caring". Male lecturers were more commonly described as "funny", and the word "spontaneous" was only used in relation to male lecturers (Sprague & Massoni, 2005). Similarly, in the evaluations of medical faculty, the words "humour" and "master" were associated with men and "warm" and "empathetic" with women (Heath et al., 2019). But words can have different meanings depending on the context (Hagtvet & Wold, 2003; Uher, 2013). Specifically, it is unclear whether "caring" was intended to mean "nurturing" or simply "taking an interest". It is also unclear whether these words referred

to behaviours (observable actions of the lecturers) or their attitudes that participants assumed to underlie lecturers' behaviours; or if the meaning of the word even depended on the lecturers' gender (Sprague & Massoni, 2005). In sum, through the examination of the language that people use, a qualitative research approach may help to reveal hidden biases that would otherwise remain undetected.

2.5 General summary of this chapter

To conclude, information that students enter into SETs may contravene verifiable evidence. Furthermore, numerous characteristics unrelated to teaching quality may influence SETs. Another problem is that gender stereotypes may influence individuals from an early age, because these stereotypes are frequently present in the academic textbooks or even guides for the lecturers. This may result in students having gendered expectations of their lecturers. Gender stereotypes may also impact SETs evaluations. Female lecturers must conform to gender stereotypes (such as nurturing behaviour) but also adhere to expectations of their role as a lecturer (assertive behaviour) to be perceived as both warm and competent. However, women may be penalised for displaying behaviours usually associated with men. This means female lecturers must carefully balance two distinct roles to obtain high ratings.

Not only students but also academic staff themselves may hold gender stereotypes, and even familiarity with the concept of gender bias may not result in changed behaviour. Even academics in Psychology field, most of which should be familiar with the research on biases, seem to be also susceptible to gender bias, as shown by higher ratings of an identical curriculum vitae with a male (vs female) name. Gender bias may be more prevalent in some departments, such as in fields with a lower ratio of women to men. Furthermore, students may be influenced by other biases (e.g., bias against ethnic minority lecturers). In sum, using SETs for promotion decisions might be discriminatory. Importantly, qualitative research may reveal hidden biases not visible in quantitative research.

In this chapter, I introduced relevant social psychological theories that may explain the influence of gender stereotypes on student evaluations of their lecturers and discussed relevant empirical findings. The next chapter explores the methodological and methodical limitations of the standardised rating scales used for these evaluations.

Chapter 3: Methodological and methodical foundations of SETs

In Chapter 2, I discussed gender biases and stereotypes in SET research and the underlying theoretical framework and considered relevant theoretical positions that could be applied to sexism as it occurs in SET context. In the current chapter, I explore whether the manifestation of gender biases may be enabled or even promoted by the methodology and methods[6] of SETs. This concerns problems of human-based data generation that relies on students' memories and decisions, but also SET's structure, wording, and format. First, I outline a relevant theoretical framework (the TPS-Paradigm and key terms), several concepts of which I then apply to describe different types of data generation methods. Raters[7], as opposed to observers, face distinct requirements during data generation. To highlight these requirements, I contrast behavioural observations with assessments both commonly considered to enable measurement. The subsequent section focuses on essential measurement-theoretical foundations and the two traceability principles, data generation traceability and numerical traceability. These principles must be implemented to meet two crucial criteria for a process to be considered measurement: 1) justified attribution of the results to the measurands (particular quantities to be measured) and 2) the public interpretability of the generated numerical values regarding their quantitative meaning. Afterwards, I use these measurement-theoretical principles to scrutinise the processes established for developing rating scales and problems that arise from these practices, such as when theoretically defining or operationalising constructs.

Subsequently, I outline a major source for limitations in implementing the two basic measurement principles in rating scales, specifically, variations in raters' interpretations of item and answer scale categories. I also critically analyse the numerical recoding of verbal answer categories, widely used to create scores as quantitative data. Afterwards, I discuss the common quality criteria of rating scales, specifically their reliability and validity, and explore whether these criteria fulfil the measurement-theoretical principles introduced. I illustrate these concepts and problems with examples and findings from SETs. Examining potential shortcomings of rating scales enables me to

[6] *Methodology* involves theoretical and philosophical assumptions of a worldview applied to inform research. In contrast, *methods* are particular techniques, procedures or tools applied to obtain specific data (e.g., Hatch & Yanow, 2008; Kezar & Talburt, 2004; Uher, 2020; Wiggins, 2011).

[7] A rater, in this context, is defined as a person using a rating scale.

determine whether the standardised format of the SETs may promote rather than reduce influences of gender biases on student evaluations of their lecturers.

3.1 Theoretical framework: TPS-Paradigm

The Transdisciplinary Philosophy-of-Science Paradigm (TPS-Paradigm) for Research on Individuals (Uher, 2015a, 2018a, 2018b, 2019, 2020) is a novel paradigm grounded in concepts integrated and further developed from across several sciences. This paradigm provides detailed conceptual (philosophical, metatheoretical, methodological) frameworks for studying individuals and phenomena related to individuals. These complex frameworks also enable exploring the foundations of measurement and quantification across sciences (Uher, 2020, 2021b, 2021c, 2022a). The TPS-Paradigm outlines and describes clear and precise philosophy-of-science concepts for exploring the foundations of underlying scientific systems. It integrates concepts and frameworks from psychology, social sciences, metrology, life sciences, physical sciences, and philosophy of science. These foundations provide scientific principles that enable critical analyses of research practices from a holistic viewpoint.

There currently seems to be no other recent research paradigm that builds upon and methodologies and concepts from various disciplines, expands these concepts and develops the new ones. The TPS-Paradigm clearly states the assumptions on which it is based and methodologies and metatheories derived from these premises, which is rarely done in psychology (Omi, 2012; Uher, 2015c). Importantly, this paradigm also provides complex frameworks explaining concepts of measurement, quantification and data generation across different sciences and in psychometrics (Uher, 2023). For that reason, the TPS-Paradigm provides the most suitable conceptual framework for analysing the development and use of rating scales. It is, therefore, appropriate for scrutinising the use of SETs from the methodological point of view.

One of the basic philosophical assumptions underlying the TPS-Paradigm that is most relevant to my research is that *all science is made by humans.* Therefore, science is inseparable from the perspectives and beliefs of the people who make it, and these human perspectives involve particular risks of biases (e.g., gender biases). These biases can manifest, for example, through the use of rating scales, and influence research outcomes. Hence, consistently with interpretivism, I emphasise that the concept of reality depends on individuals (Lincoln et al., 2018; Poucher et al., 2020), who perceive, interpret and

mentally construct the world. In contrast, rating scales build on positivism, the assumption that there is objective truth independent of observers (*see also Chapter 4*).

3.1.1 Different types of phenomena studied in SETs

When completing SETs, students are asked to evaluate the lecturers' behaviours, attitudes and intentions (as inferred by students) using standardised rating scales. Therefore, to appropriately analyse SET ratings and students' formation of these ratings, the different types of phenomena studied in SETs must first be defined.

Behaviours are defined in the TPS-paradigm as "external changes or activities of living organisms that are functionally mediated by other external phenomena in the present moment" (Uher, 2016, p. 490). Behaviour denotes physical phenomena (such as movements) that are transient and bound to individuals' bodies but occur externally to these bodies and are thus publicly accessible (Uher, 2015a, 2018b). For instance, students can observe their lecturers' teaching behaviours (e.g., lecturers walking around in classrooms or asking questions).

In the TPS paradigm, behaviours are differentiated from the *psyche*, which is defined as "the entirety of the phenomena of the immediate experiential reality both conscious and non-conscious of living organisms" (Uher, 2016, p. 478). All psychical phenomena occur entirely internal to an individual. Therefore, they are only accessible by individuals themselves and never by other persons, thus only privately (Uher, 2015a, 2018b). Furthermore, psychical phenomena can be accessed only in the present moment and are always strictly bound to an individual (Uher, 2015a). The critical distinction between behaviours and the psyche is that other people *can* directly observe an individual's behaviours. In contrast, psychical phenomena (e.g., attitudes, thoughts) are accessible only to the individual experiencing them, and others can only infer them from observable (publicly accessible) phenomena (e.g., inferring attitude from behaviour or verbal expression). This process requires interpretation. That is, students may only *infer* their lecturers' attitudes or intentions from lecturers' verbal or non-verbal behaviours by interpreting them (e.g., a lecturer's lack of enthusiasm inferred from regular late arrivals to lectures or short, vague answers to students' questions).

Some of the most frequently studied psychological phenomena are constructs. *Constructs* can be metatheoretically defined as "abstract concepts describing complex constellations of phenomena that cannot be directly perceived at any moment in their entirety but that are only theoretically constructed as entities" (Uher, 2018b, p. 11).

Constructs are, therefore, abstract and socially constructed (Rosenbaum & Valsiner, 2011; Uher, 2018a); they refer to concrete phenomena only on a conceptual level and thus cannot be directly observed or measured in themselves (Uher, 2022a). However, this abstraction inevitably involves subjective interpretations of people, which may be influenced by different biases. Furthermore, humans use abstraction to emphasise some aspects of a construct but deemphasise others (Uher, 2021b; Whitehead, 1929). For instance, gesturing, walking, or standing up emphasise different features of movement but gesturing can additionally also imply communication. Importantly, because constructs only exist on an abstract level, it is essential to distinguish between constructs and their *referents*. Referents are the real-world phenomena that are represented by constructs on an abstract level (Uher, 2022a). These can be different types of phenomena (e.g., in terms of their accessibility) to which constructs can refer. Referents can be specific observable phenomena (e.g., behaviour) but also other concepts (e.g., sub-construct; Uher, 2021a). Constructs may also be blended and refer to phenomena of different qualities (e.g., attitude and behaviours; Uher, 2021c). A frequent error that researchers make is *construct-referent conflation*, in which they mistake constructs for the actual phenomena to which they refer (Rosenbaum & Valsiner, 2011; Slaney & Garcia, 2015; Uher, 2020, 2021c). However, constructs are only conceptual tools construed to make complex study phenomena accessible for empirical study (e.g., Uher, 2021a). For example, a construct explored in SETs is 'teaching quality' (or 'teaching effectiveness'), which may refer to different concepts of teaching behaviours, for instance, support that lecturers provide in assignments, but also specific behaviours, such as clear answers to students' questions.

Ratings are language-based methods. Many phenomena (e.g., constructs such as 'happiness' or some types of psychical phenomena such as thoughts) can only be studied through language (Uher, 2021b). Language is a sign system and therefore plays a crucial role in data generation (e.g., in standardised rating scales) and needs special consideration. The TPS-paradigm classifies written or spoken language and other sign systems as *semiotic* representations (Barthes, 1967). Semiotic representations comprise three different constituents: a) a physical constituent used as a *signifier* (e.g., the printed word "support" on a survey screen), to which people assign b) particular *meanings* (e.g., "help"), and both of which refer to c) particular *referents*, thus the actual objects or phenomena of interest (e.g., lecturers' teaching behaviours, help provided by lecturers; Uher, 2018b). Signifiers and referents are conceptually linked through the meanings that people assign them. The link between signifiers, referents and meanings establishes a

tripartite composite, which forms a sign (Uher, 2021b). Importantly, meanings are always only assigned to the signifiers by individuals rather than inherent to these signifiers in themselves (Uher, 2015a, 2018a). This highlights that the construction of the meaning is always fundamental to any language-based method. Signifiers are mostly arbitrary, i.e., they typically bear no resemblance to the objects or phenomena they denote. For instance, the written word "support" on a SET form cannot look like actual support this word refers to because 'support' is a construct and, as such, exists only abstractly in people's minds. However, this construct also has several referents that exist in the real-world. These may be specific behaviours, for example, lecturers providing detailed feedback. Therefore, students must not only construct a specific meaning for this word but also consider particular concrete and observable referents of 'support', and this process may be affected by their subjective beliefs, expectations, or biases (*see 3.2.4*). That is, language and other semiotic systems are composite phenomena and thus involve more complexities than the other types of phenomena (such as behaviours, attitudes or constructs).

One important type of semiotic representations are *data*, which are the sign systems used by scientists to represent the properties of study phenomena in a persistent and easily perceivable manner, enabling the subsequent data analysis. To properly understand the role that raters have in encoding the information about study phenomena into data, researchers must clearly define the above-mentioned three constituents that any sign system comprises. For example, in SETs, the physical constituents of language would be the statements printed on SET form (e.g., "The lecturer has helped me to learn"), to which students must assign particular meanings in the given context. Both signifiers and meanings refer to referents, specifically, actual lecturers' behaviours that students may have seen (e.g., lecturers explaining topics). However, because students may assign various meanings to the same SET statement, the particular referents that students have in mind may differ depending on their specific interpretations ("helped me to learn" can mean lecturers answering questions in-depth to one student and lecturers' manner of presentation to another student).

3.1.2 Methods of data generation

Data fulfil an important representational function because they can be analysed instead of the actual study phenomena. However, researchers can use data to derive correct inferences about the study phenomena only if these data appropriately represent the properties of the study phenomena (Uher, 2018b). Therefore, people (e.g., participants

or researchers) who generate data must encode in the data relevant information about the study phenomena in appropriate ways.

Accessibility of study phenomena plays an important role in data generation because some phenomena are only partially accessible (e.g., thoughts are only available to the individuals experiencing them; Uher, 2019). These modes of accessibility under everyday conditions are determined by different metatheoretical properties of these phenomena (*see below*), which also affects the phenomena's accessibility to researchers (Uher, 2015c). Therefore, considering phenomena's metatheoretical properties also enables defining the research methods needed to investigate these phenomena under research conditions (Uher, 2018a). For example, we can use observational methods to investigate teaching behaviours (such as lecturers explaining the topic) but not lecturers' thoughts. Different metatheoretical properties of phenomena are: a) location internal/external to the studied individual's body and b) the phenomena's temporal extension (Uher, 2019). *External* phenomena are phenomena that occur outside of an individual's body and can therefore be publicly observable. In contrast, *internal physical phenomena* (e.g., bones or brain) can only be perceived and accessible using special methods, such as technical or invasive methods (e.g., surgery; Uher, 2018c). *Psychical phenomena* are also internal, but only perceivable in oneself, and not accessible to others. Psychical phenomena and behaviours differ regarding their internal versus external accessibility (exclusively private versus also public accessibility). Researchers must therefore apply different methods to capture and study them (Uher, 2018b).

These differences in the accessibility of phenomena can be used to scrutinise methodological concepts in psychology. For instance, the concepts of introspection and extrospection are frequently differentiated from each other, even though individuals can introspect (perceive their inner phenomena) and extrospect (perceive external phenomena) simultaneously. Thus, they cannot be clearly differentiated as methods in any given investigation (Uher, 2015c). To remedy these conceptual problems, the TPS-Paradigm introduces methodological concepts to clearly distinguish between methods that can and cannot be used to study specific phenomena. For instance, psychical phenomena can be perceived only internally, and methods used to explore these phenomena must reflect that. Therefore, the TPS-paradigm distinguishes between *introquestive* and *extroquestive* methods. *Introquestive methods* are applied to explore phenomena only accessible and perceivable from within an individual itself and by nobody. *Extroquestive methods* are used to study phenomena that are perceivable as from outside of oneself and thus

accessible to other individuals as well (Uher, 2019). Two methods that researchers frequently use to generate data about behaviours are *behavioural observations* and *standardised rating scales*. Behavioural observations are extroquestive methods because behaviours occur externally to observed individuals' bodies and are thus publicly accessible. In contrast, standardised rating scales are *introquestive* methods, because rating procedures are memory-based and, therefore, involve raters' thoughts and other internal processes that are accessible only to raters and who generate the data on the basis of these perceptions. Rating methods involve introquestive data generation in retrospect, which enables raters to generate data even when study phenomena is absent (Uher, 2018b). For example, students can evaluate their lecturers on rating scales by reconstructing memorised events about these lecturers even in their absence.

Phenomena can also differ regarding their temporal extension. For instance, phenomena can occur in *a brief moment* (e.g., a nod) or be *longer-lasting* (e.g., beliefs; Uher, 2018c). These differences in temporal extension affect the accessibility of study phenomena, thus distinct types of methods must be applied to study transient versus temporally extended phenomena. *Nunc-ipsum methods* refer to methods that enable real-time recording of events (used in, e.g., behavioural observations (Uher, 2015d). Nunc-ipsum methods are therefore used for exploring transient phenomena, which must be present in the moments of data generation, and can be either brief (e.g., nod) or longer-lasting (e.g., walking). These methods can also be used to study psychical phenomena (Uher, 2018b). For example, individuals can be asked to record specific thoughts (e.g., every instance they think about exercise) at the same time as they experience them (thinking aloud technique). A potential disadvantage is that many individuals may simply not be used to this method and trying to think aloud may affect their thought process (Priede & Farrall, 2011). *Methods of long-term memory-based introquestion* (e.g., assessment methods), by contrast, can be used to explore phenomena that are temporally extended. For instance, people can be asked to record their beliefs, which they can reconstruct from their memories at any given time (Uher, 2019).

When information from a certain kind of study phenomenon is represented in another one, this is defined in the TPS-paradigm as *conversion* (Uher, 2016b). For instance, in data generation, people may encode information about specific study phenomena into data. However, conversions of information may be distorted or result in loss of information if study phenomena differ in their metatheoretical properties,

To highlight the different demands placed on the people generating data in behavioural observations versus with standardised rating scales, I now briefly describe the process for each method and outline what is expected from observers and raters. In *behavioural observations*, observers must decide how to separate, categorise and encode the behavioural events at the moment during which they occur, thus using nunc-ipsum methods (Uher, 2018b). Therefore, observers must know all the behavioural acts under study and the data schemes that observers should use to encode occurrences of these behavioural acts systematically. For instance, observers may use binary units to record the occurrence or absence of a specific behaviour (e.g., a gesture). To ensure that observers can meet these demands during observing, observers commonly receive instructions and training (Uher, 2018b).

In *standardised*[8] *rating scales*, researchers commonly assume that raters recall relevant events from memory, weight all the pieces of evidence recalled and then make a balanced decision on the basis of this evidence. For personality ratings, for example, this means they must compare the evaluated person's behaviours with behaviours of other individuals and over time, "average" the instances considered and make a judgement regarding the considered behaviours (Uher, 2018b). This highlights several demands placed on raters during the rating process. Specifically, raters must a) read and interpret the item statements and construct their meanings in the context of the given enquiry, b) read and interpret the answer scale categories, c) retrieve information about relevant past observations from memory, d) make decisions on how to separate, categorise and encode information retrieved from memory about study phenomena and its properties, e) quantify occurrences of specific behaviours and underlying beliefs, and, f) encode this judgement thus generated in the scale provided (Uher, 2018b). Thus, the demands placed on raters differ fundamentally from the demands placed on observers (*see also 3.2.4*).

Rating scales, therefore, involve so-*called introquestive data generation in retrospect* (Uher, 2019). Unlike in nunc-ipsum methods, which involve data generation in real-time, raters must reconstruct their ideas and beliefs from their memories. This means that the phenomena that raters consider (e.g., specific teaching behaviours) are no longer present in the moments of rating generation. Ratings, therefore, classify as *long-term memory-based introquestive methods* (Uher, 2018b). Furthermore, behaviours can only be

[8] When a rating scale is standardised, researchers interpret individual scores obtained from rating scale based on the scores from a normative sample (Bagby et al., 2005).

encoded at the moment in which they occur, but raters are required to consider past behaviours and events that are no longer observable during rating. Raters may then consider behaviours they think have occurred or opinions of other people with whom they may have talked about the evaluated person (e.g., other students) rather than behaviours that have actually taken place (Uher, 2018b).

For instance, students are asked to judge their lecturers' behaviours over the academic term (e.g., students' experiences with their lecturer's support). But at the time of completing SETs, students cannot directly experience these behaviours anymore because they occurred in the past. Instead, students must recall relevant information from their memories. However, this information may also be influenced by ideas and beliefs that students developed about their evaluated lecturers' behaviours (e.g., through the discussions with other students), which may be influenced by stereotypical biases. For example, abundant research suggests that people frequently remember stereotype-consistent better than inconsistent information or even mistakenly recall stereotype-consistent information that never happened (e.g., Dodson et al., 2008; Lenton et al., 2001; Sherman et al., 2003). Furthermore, some SET items (e.g., "The lecturer is enthusiastic about what they are teaching") require that students implicitly judge their lecturers' behaviours as compared to those of other lecturers. Specifically, students must quantify what is specific to their individual lecturers' behaviours. But because behaviours are dynamic and transient, individual specificity can never be directly perceived at any given moment. To determine individual specificity and explore whether behaviours vary among individuals and in ways somewhat stable over time, both differential and temporal patterns must be considered (Uher, 2015a). However, these patterns can never be directly perceived at any given moment. Therefore, the generated data can refer only to abstract ideas about individual specificity (Uher, 2014, 2018b, 2018c). As outlined above, these data may reflect only students' own ideas and beliefs that they have about their lecturers' individual-specific behaviours rather than behaviours that actually occurred – with all the biases that may thereby be involved.

3.1.3 Measurement-theoretical foundations and the two traceability principles

An important kind of data generation involves *quantitative* data, which are primarily used in physical sciences but also frequently in psychology and social sciences (Uher, 2022a). Importantly, not all quantitative data can be classified as measurement data (Abran et al., 2012). *Measurement* is a structured process in which numerical values are

allocated to the attributes of objects or events in clearly defined and traceable ways (Mari et al., 2016; Uher, 2019). The measurement process must involve an empirical interaction with the *measurand*, which is the particular quantity of the target property to be measured in given study phenomena (e.g., duration of lecturer's speech; Uher, 2020).

It is important to distinguish between measurement and assessments. Assessment is a broader term and involves collecting information for pragmatic purpose. This fundamentally differs from measurement as defined in metrology, the science of measurement. In contrast, measurement is a highly structured process based on the international standards of measurement, and involves applying detailed measurement procedures with consideration of admissible mathematical operations (Abran et al., 2012; Uher, 2023). I critically explore quantification practices in psychology through the viewpoint of measurement, which differs from the mainstream approaches commonly used in psychometrics.

SETs are frequently considered to be "measurements", and SET scores are treated as quantitative values, but variations in raters' use of rating scales may interfere with the two crucial criteria of measurement. These criteria are considered essential for measurement in all sciences regardless of the specific procedures used, and therefore, can also be applied in psychology. The two crucial criteria of measurement are: a) justified attribution of the results to the measurands, and b) public interpretability of the results and their quantitative meaning (Uher, 2020, 2022a). For justified attribution of the results to the measurands, the result must contain information specifically about the object measured. Furthermore, measurement must be reproducible, which refers to both the results and the actual process itself (Mari et al., 2016; Uher, 2020). This requires establishing unbroken and traceable connections between measurand up to the results. Public interpretability of the results and their meaning requires an unbroken and systematic connection of the numerical results to known and conventionally agreed quantity standards. The numerical values assigned to the research object must be interpretable in the same way by both people who generate and people who use them. Results must also be independent of the opinions of the people who operate a measurement process (Uher, 2020). Specifically, the results must represent the same information with regard to the measurand in all contexts. This helps to ensure that the meaning of results does not depend on the people who generate or use the data and enables the public interpretability of the results (Uher, 2022a). These criteria underlie frameworks from metrology, the science of measurement, and must be fulfilled for a

process to be considered measurement. To further investigate whether SETs may contribute to the manifestation of biases, I apply the concepts from the TPS-paradigm to explore whether SETs fulfil these crucial methodological criteria and whether the outcomes of SETs may be considered measurement data.

Over the years, researchers developed different theories of measurement. Here I focus on *representational theory of measurement* and on *metrological frameworks* of measurement. I introduce this theory and these frameworks in order to later apply the concepts derived from them to scrutinise psychological scale development as used for rating generation.

3.1.3.1 Representational theory of measurement

Representational theory of measurement (Krantz et al., 1971; Suppes et al., 1989) is a theory developed in social sciences (Mari et al., 2017). This theory defines measurement as a procedure in which structures observed in empirical study phenomena, *called the empirical relational system,* can be mapped onto symbolic (usually numerical) structures, called the *symbolic relational system,* in such a way that the numerical structures created (i.e., the data) appropriately represent the empirical structures observed (Tal, 2015; Uher, 2018b). The empirical relational system involves the study phenomena, their qualities, and quantitative properties, whereas the symbolic relational system is the data system that scientists developed to encode and analyse information about these study phenomena. The mappings between empirical and symbolic relational systems are called *assignment relations* (Krantz et al., 1971). Therefore, this theory defines measurement procedures as mappings of the empirical relational system onto a symbolic system, according to certain relations (see *Figure 1).*

However, this mapping is necessary but not sufficient for a procedure to be considered measurement (Mari, 2000). Essentially, representational theory considers measurement to be equal to scale construction or a numeral assignment and neglects other elements necessary for measurement, such as accuracy or error (e.g., Tal, 2013). Importantly, to constitute measurement, the same properties always need to be encoded with the same symbols so that results always represent the same information with regard to the properties studied (Uher, 2018b). Therefore, the conditions upon which an empirical relational system can be mapped onto a symbolic must be clearly defined and fixed. However, representational measurement theory fails to outline concepts for implementing these conditions (Mari et al., 2017; Uher, 2018b). It, however, outlines the

basic requirements for any data collection. Specifically, this theory states that study phenomena and their properties (the empirical relational system) must be appropriately represented by and encoded in the signs used as data in a study (the symbolic relational system). Therefore, this theory is also relevant to data generation using rating scales (Uher, 2021a, *see also 3.2.4*).

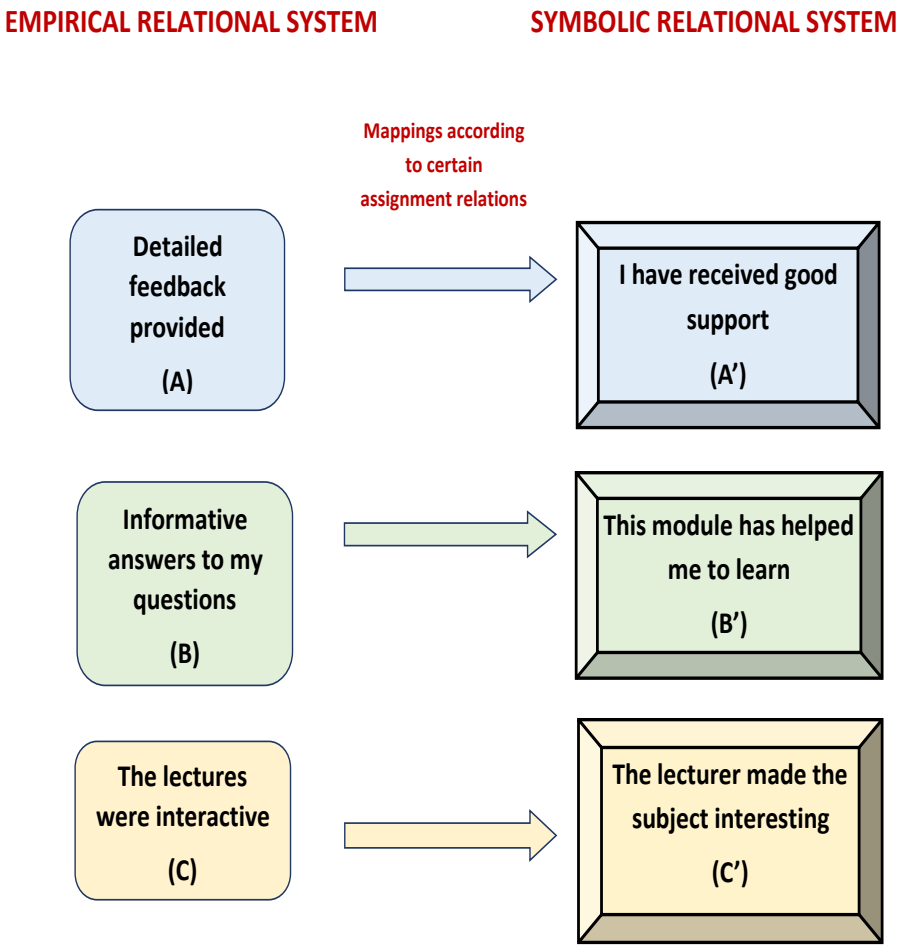

Figure 1

An Example of Mapping Relations Between the Empirical Relational System (e.g., Students' Experiences) and the Symbolic Relational System (e.g., SET Items)

3.1.3.2 Metrology: Basic methodological principles underlying physical measurement

Metrology is the science of measurement, originally developed for physics and engineering. Metrological principles are applied in many sciences but not commonly in psychology. But currently, the term "measurement" seems to have a distinct meaning in

different fields (Uher, 2021b). However, similarly to other scientists (e.g., physicists), psychologists generate and analyse numerical values to infer quantitative information about study phenomena. But to build knowledge of the real-world phenomena through examining findings from different disciplines and yield meaningful inferences about study phenomena from quantitative data, *basic principles of scientific measurement and consistent principles* should apply across different sciences (Uher, 2018b, 2021a). This is not possible if researchers define a crucial scientific activity such as measurement differently in various sciences. Therefore, merely changing the definition of 'measurement' does not establish comparability across different sciences, but only undermines it (Uher, 2023). In the TPS-paradigm, two key principles of scientific measurement that implement the two key criteria outlined above and underlie metrological frameworks on abstract, methodological levels are identified. These principles are *data generation traceability* and *numerical traceability* (Uher, 2020, 2021c). They highlight core ideas that are also applicable to other disciplines, including social sciences and psychology.

To properly understand the measurement principles discussed next, it is necessary to first distinguish between *qualities* and *quantities* of study phenomena. Qualities are properties that differ in kind, whereas quantities are divisible properties of qualities (Hartmann, 1964, as cited in Uher, 2022a). That means, quantities are always of some quality (Uher, 2018b, 2022a). Therefore, to enable measurement, researchers must always define both a) the quality under study (qualitative properties), b) the specific quantitative properties that are to be measured in these qualitative properties (Uher, 2022a). For instance, the quality under study could be duration of lecturer's speech, and the number of minutes it lasted would denote specific quantitative properties to be measured.

3.1.3.2.1 Data generation traceability

Importantly, study phenomena cannot be measured in themselves. Instead, we can measure only some of their properties (Uher, 2022a) because each phenomenon contains different properties that can be measured. For instance, we cannot measure the lecturer's "walking" as a behaviour in itself, but we could measure their walking speed or distance.

The term "data generation traceability" is not explicitly used in metrology, however, it underlies the metrological measurement processes and all measuring instruments on abstract levels (Uher, 2020). *Data generation traceability* requires that an unbroken, traceable connection chain is established between the measurands and the

generated outcomes (Uher, 2018b). Specifically, measurement results must be connected by this chain of comparisons to the measurand (the entity to be measured, e.g., duration of a lecturer's speech at certain occasion). This chain of connections starts with the specific property that is to be measured, called the input property (measurand), and its first interaction with a mediating property systematically connected to it, that researchers may introduce if measured properties cannot be accurately perceived by humans (e.g., exact duration of a lecturer's speech). Importantly, these steps establish proportional (quantitative) relations between quantities of these properties, from measurand up to result. The connection chain ends with the numerical value that is assigned to the measurand (Uher, 2019, 2020, 2022a).

An unbroken, traceable chain of connections enables tracing the results back to study phenomena (Uher, 2020). This fulfils the requirement of data generation traceability. Specifically, it clearly defines data generation procedures, including rules on how to assign results to study phenomena in ways that ensure these results represent valid information about these phenomena. Lack of data generation traceability, therefore, leads to questionable validity of the results.

In contrast, implementing data generation traceability ensures that the results are assigned to measurands in fully transparent ways and are therefore justifiably attributable to measurands, meeting the first crucial criterion of measurement.

3.1.3.2.2 Numerical traceability

The second basic scientific principle of measurement, numerical traceability (called metrological traceability in metrology), entails establishing a shared quantitative meaning of the numerical values assigned to the measurand (Uher, 2022a). Specifically, numerical traceability requires that the generated numerical values (e.g., "7") are clearly linked to known standards and meanings in transparent and defined ways (e.g., "7" of what and how much of that is "7"). A systematic connection must therefore be established not only between a measurand and the outcome (data generation traceability), but also between the numerical value (e.g., 1.70) assigned to the measurand (e.g., a person's body height) and a known/defined *reference quantity* of the property measured (e.g., standard metre; Uher, 2022a). These are concrete defined entities of a property, which are then used to define *measurement units* (e.g., metre, inch; Uher, 2022a). A measurement unit is a "real scalar individual quantity, defined and adopted by the convention, with which any other quantity of the same kind can be compared to express the ratio of the two quantities

as a number" (VIM, 2007). Therefore, measurement units that measure the same property must represent the same quality (e.g., 1 metre or 1 centimetre as measurement units of the same property length) and are used to measure (divisible) quantitative properties of this quality, e.g., different instances of length (Uher, 2018a). For quantity values to be understood with regard to their quantitative meaning, reference units must first be established (Uher, 2022a). For instance, "7" cannot be universally understood without the measurement unit because it could refer to different properties, such as 7 kg, 7 m or 7 h, all of which have different meanings. Similarly, the length of "7" cannot be interpreted on its own because it could mean 7 mm, 7 cm or 7 m, which are different measurement units of length. Measurement units are always designed purposefully, and this must be done before executing a measurement process. Scientists therefore developed the International System of Units (SI), in which they clearly defined and established measurement units and which is constantly updated (e.g., redefined by Newell and Tiesinga in 2019).

In metrology, numerical traceability is implemented practically through comparisons of results with *primary references*. These references must first be defined, ensured to be stable and internationally accepted (Uher, 2022a). A primary reference could be a measurement standard or a definition of a measurement unit (e.g., through an object or system that defines a relationship to the target quantity) and needs to be universally understood, specifically, to symbolise the same quantitative information regardless of time and context (Uher, 2020, 2021c). For instance, a primary standard reference can be an international prototype metre. The metre was formerly defined as one-tenth millionth of the distance between the North Pole and the equator passing through Paris, originally available in the form of a metal bar (Swindells, 1975). However, in order to dematerialise the (definition of) metre, metrologists redefined it as a specific length of time travelled by light (e.g., Dearden, 2014; Uher, 2022a). Therefore, metrologists regularly review and update primary references to guarantee measurement precision. *Calibration chains*, are unbroken documented chains of connections with stated uncertainties from primary references (e.g., international prototype meter) to a) secondary references (e.g., national standards), and from these to b) all working references used in measurement or everyday life in order to measure a specific property (e.g., thermometers; Uher, 2020, 2022a; VIM, 2004). This helps to ensure that the measurement results are precise and accurate. For instance, weighing scales in the laboratories are constantly calibrated (tested for accuracy) to ensure that they indicate a correct weight. This can be verified by comparisons to national standards (national measurement standard

laboratories), which are themselves compared to international standards (international measurement standard laboratories). In sum, metrologists invest enormous effort into ensuring the accuracy of their primary references and the quantitative meaning of the numerical measurement results.

Implementing these two traceability principles in data generation processes allows fulfilling the two criteria crucial for measurement: a) justified attribution of the generated results to the measurands, and b) public interpretability of the generated results and their meaning (Uher, 2020, 2022a). The traceability principles are also relevant to psychology and social sciences because psychologists frequently use numerical values to gain quantitative information about their study phenomena. Importantly, in the TPS-paradigm, measurement is distinguished from *numeralisation,* which can be defined as creation of numerical values without any reference to measurands, reference quantities or properties being studied (Uher, 2022b).

In sum, these concepts from the TPS-Paradigm can thus be applied to explore conceptual and methodological underpinnings of rating scales to investigate, for example, whether human-based data generation reliant on raters' memories and decisions, and the particular structure, wording, and format of standardised rating scales enables meeting criteria of measurement. These concepts will also be used to examine whether the use of rating scales may open up rather than reduce possibilities for raters' potential gender biases to manifest in their ratings due to the lack of specification of the elements in these scales.

3.2 Common practices of rating scale development critically analysed

In this section, I apply the just outlined methodological concepts of measurement of the TPS-paradigm's theoretical framework to critically explore the processes of scale development commonly used in psychometrics. *Psychometrics* is the branch of psychology related to the quantification of "mental attributes, behaviour, performance, and the like, as well as the design, analysis, and improvement of the tests, questionnaires, and other instruments" (APA, 2021). Rating scales are widely used in psychometrics because they provide a convenient, cheap and quick way to generate large amounts of numerical data from numerous respondents (e.g., Rose, 2005). Psychologists generally assume that a process involving the use of rating scales constitutes measurement. I outline the main steps of scale development and then apply the two basic criteria for measurement

and the methodological principles of measurement introduced above to scrutinise this common assumption.

3.2.1 Developers theoretically define constructs

To design tools that enable researchers to empirically investigate given constructs (e.g., 'teaching quality'), developers must first theoretically define these constructs. This involves reviews of existing literature and interviewing subject experts (e.g., other researchers). Afterwards, researchers must clearly define what study phenomena a construct should conceptually capture and establish a clear differentiation from other constructs (MacKenzie et al., 2011; Spector, 1992). This process should result in a clear, concise scientific *theoretical* definition of the construct that is more precise than informal everyday definitions, such as those found in a dictionary (Crump et al., 2018).

However, many definitions of constructs in psychology are ambiguous, overlapping, vague or even completely absent (Phillips, 1932; Uher, 2021a; Zagaria et al., 2020). This also applies to SET constructs. SETs should evaluate 'teaching quality' or 'teaching effectiveness'. For example, Teaching Excellence Framework (TEF) defines 'teaching quality' as "teaching practices, which provide an appropriate level of contact, stimulation and challenge, encourage student effort and engagement, and which are effective in developing the knowledge, skills, attributes and word readiness to students" (TEF, 2016, p. 12). Others define 'teaching quality' as "strong instruction that enables a wide range of students to learn" (Darling-Hammond, 2009, p. 5) or "what teachers do to promote student learning inside the classroom" (Tok, 2010, p. 4142). 'Teaching effectiveness', in some authors' views, refers to how successful are lecturers in helping their students to acquire knowledge (Boring, 2017; Shadreck & Isaac, 2012).

Others argue that 'teaching quality' can be divided into good teaching and effective teaching (Berliner, 2005). This definition suggests that 'teaching effectiveness' may be a sub-category to 'teaching quality'. Despite this, researchers frequently describe SETs as student evaluations of either 'teaching quality' or 'teaching effectiveness' and use them either interchangeably instead of defining 'teaching effectiveness' as a sub-construct of 'teaching quality', or even without specifying their meanings. Consequently, ambiguity in or even lack of these definitions may lead to different choices in the next steps of scale development, i.e., operationalising constructs, and therefore to the selection of distinctive sets of items (V. A. Miller et al., 2009). Researchers may also erroneously assume that

operational definition may substitute the theoretical one (Hibberd, 2019; Uher, 2020, *see below*).

3.2.2 Developers operationalise constructs

Constructs exist only on an abstract level. Therefore, to enable empirical investigation, developers must establish a nomological network consisting of a theoretical framework, an empirical framework and systematic connections between them (Cronbach & Meehl, 1955). In the theoretical framework, researchers define a certain construct and determine its sub-constructs, establishing its *theoretical definition*. However, because constructs are conceptual in nature, and their referents never occur all at once, they must then be *operationalised* in concrete, accessible, and potentially measurable empirical indicators that are specified in the empirical framework (Uher, 2018b, 2020, 2021b, 2023). These are either single and concrete components, often called *proxies*, or multiple concrete components called *construct indicators*. For instance, when evaluating teaching quality, a proxy could be the number of students who passed the lecturer's class, and several indicators could be chosen which represent a part of construct's referents (Uher, 2022b) that developers deem to be representative of the construct (e.g., a lecturer's achieved education level or teaching awards). Verbal items in rating scales can also serve as construct indicators. Importantly, connections between these construct indicators and a given construct depend solely on researchers' decisions (Uher, 2021b; 2018b). This is called in the social sciences *measurement by fiat* – by decree (Cicourel, 1964).

In construct operationalisation, links between constructs and their indicators are always decided by the researchers by decree instead of being empirically proven. Therefore, the relation between construct and its indicators is always only assumed (Uher, 2020). For implementing data generation traceability, results must be justifiably attributable to a measurand. Constructs do not exist as the real entities in the world and can only refer to measurands. Specifically, constructs in themselves as conceptual entities cannot be measured, only quantity of some of their empirical indicators, (e.g., a number of questions that lecturers answered). However, the abstract language used in rating scales obscures what exactly constitutes measurands (quantities to be measured). From the viewpoint of measurement principles, this process fails to establish documented unbroken empirical connection chains between possible quantitative properties of constructs (e.g., a construct 'teaching quality') and those of their indicators (e.g., scores generated for SET

items, representing researchers' attempts to quantify teaching behaviours). Construct operationalisation, therefore, cannot classify as measurement (Uher, 2020, 2021b).

Because constructs are conceptual entities, they frequently refer to distinct kinds of phenomena of different quality. For example, a construct 'teaching support' could be operationalised by construct indicators such as 'lecturers' feedback', but also 'lecturers' replies to questions' or inferred 'caring attitude', thus different phenomena with different qualitative properties. But in measurement, only measurands featuring the same quality may be compared to one another in terms of their quantities (divisible properties). However, we cannot identify any divisible properties in blended concepts, such as the construct 'teaching support', which involve phenomena not only of different qualities but also quantities, with many of these phenomena not even occurring at the same time. This makes the direct comparison of quantitative properties that are summarised in the blended concepts impossible (Uher, 2021b). Specifically, these quantifications refer to multiple entities of different quantities and qualities occurring at different times and therefore cannot be summarised. Using an example above, we cannot identify quantities in the abstract entity such as the construct 'teaching support' by itself, merely in its indicators (e.g., a number of answered questions). Furthermore, even though lecturers' scores for construct indicator "feedback" may provide information about 'teaching support', the link between this construct and its indicator was artificially chosen by researchers, instead of reflecting an actual empirical connection between them.

Therefore, quantifications derived from summarising scores of qualitatively different indicators are artificial. These abstract indicators (e.g., verbal items in rating scales) may refer to various phenomena of different quality. These artificial quantifications may provide some information about individual differences between raters, but it is impossible to identify divisible properties of any quality in these abstract items. This indicates that quantities in constructs cannot be established. In sum, construct operationalisation fails to meet the crucial criteria of measurement (justified attribution of results to measurands, publicly interpretable meaning).

3.2.3 Structure and format of rating scales

Rating scales generally consist of a set of standardised *item statements*, which may describe, for example, certain attitudes or behaviours (e.g., "The lecturer made the subject interesting"). Raters judge these statements by choosing multi-stage *answer scale categories*, which can be labelled numerically ("1, "2", "3") or verbally (e.g., from

"strongly disagree" to "strongly agree" or from "never" to "very often"). Raters may also have an option to mark an item statement as "non-applicable" if the item statement refers to something that the rater may have not experienced.

Stevens (1946) defined measurement as an assignment of numerals to objects or events according to certain rules. This definition is widespread in mainstream psychology, but incompatible with many measurement theories underlying metrological processes. Stevens also distinguished between different types of data, which determine what empirical operations may be applied. The commonly used categories are *nominal*, *ordinal*, *interval*, and *ratio* (Stevens, 1946). Nominal data represent membership in certain categories (e.g., male/female), whereas ordinal data represent categories that are ranked by a particular order (e.g., education level). In contrast, interval data represent a continuous scale with equal units but without an absolute zero (e.g., intelligence test). Finally, ratio data also have equal spaces between values but always have an absolute zero point below which they cannot fall (e.g., income).

SETs commonly consist of Likert scales, on which raters indicate their level of agreement with statements (Sinha & Sundaram, 1962). Likert scales are, therefore, commonly assumed to represent *ordinal scales.* In order to use parametric statistics, researchers frequently treat Likert scales as interval data. But unlike interval data, a difference between categories in ordinal data can be neither specified nor calculated (McCullough & Radson, 2011). For example, a presumed equal difference between answer categories is arbitrary and may vary between raters when they interpret rating scales.

The scores derived from chosen answer categories may not indicate what participants considered, specifically, scores representing answer categories on higher ends of the scale do not mean that participants who chose them considered more evidence (higher quantity of the evaluated quality) than those who selected answer categories on lower ends. For instance, evidence shows that some raters who chose higher ends of the scale quoted "insufficient evidence" as a reason for their choice (Uher, 2017, 2018b). Similarly, it remains unclear whether different answer categories represent different quantities of the same quality or rather completely different qualities altogether (Uher, 2018b, 2022a). For example, picking "strongly disagree" instead of "strongly agree" may indicate completely different quality rather than just lower level of agreement. This precludes identifying divisible properties in these blended concepts (Uher, 2022a, 2022b).

Importantly, the structure and format of rating scales depend solely on the researcher's decisions. For example, researchers choose the number of answer scale categories, an ascending versus descending format, or wording of the corresponding labels. The chosen format of the scale therefore defines the obtained data structure. For instance, an item statement with three answer categories can produce only three possible values. Similarly, the number of possible entries may be decided by the number of item statements. For example, raters may often generate only one value per each item statement and the number of item statements may therefore decide how many values are produced. Data are therefore defined by the scale's characteristics instead of properties of study phenomena. This differs from behavioural observations, in which observers generate data based on study phenomena and can obtain multiple data for the same variables if study phenomena occur repeatedly (Uher, 2018b).

3.2.4 Raters complete the scales

Scales are given to raters who evaluate phenomena described with item statements and generate data about them directly instead of, e.g., using technical instruments. Raters, therefore, interpret the item statements and answer scale categories, consider relevant study phenomena (e.g., thoughts, emotions, behaviours), make their judgement and choose a corresponding answer scale category. This means, on a conceptual level, that raters interact with both the properties of study phenomena and the method used (Uher, 2018b). This type of data generation inevitably involves raters' interpretations of item statements and answer categories but also their perceptions of the study phenomena.

In the lens of the representational theory of measurement, rating scales both describe empirical relational system (e.g., teaching behaviours) and form elements of the symbolic relational system (e.g., item statements). Rating scales, therefore, represent both the symbolic and the empirical relational system, which leads to their conflation (Uher, 2018b). But unlike in behavioural observations, raters typically receive no instructions about how to interpret the study phenomena and rating scales or encode their judgements. Raters must, therefore, intuitively develop their own interpretations and definitions of item statements and answer scale categories and afterwards construct their meanings in the context of the given enquiry. Raters must also decide what serves as the empirical relational system (e.g., what teaching behaviours to consider), determine specific meanings of a symbolic system (e.g., specific item statements) and conclude how to establish assignment relations between these two systems (Uher, 2018b, 2021a). These

decisions typically remain unexplored, which leads to the various problems as discussed in the following section.

3.2.4.1 Variations in raters' item interpretations

3.2.4.1.1 Raters tend to interpret item statements differently

Researchers frequently assume that if they standardise rating scales, raters will interpret item statements identically (Block, 1998; McCullough & Radson, 2011; Uher, 2018b). However, the lack of definition and the abstract wording of the rating items may entail that individuals develop a range of meanings for the same standardised statement, therefore called "a field of meaning" (Rosenbaum & Valsiner, 2011; Uher & Visalberghi, 2016). Numerous empirical studies support this idea (Arro, 2013; Lundmann & Villadsen, 2016; Uher & Visalberghi, 2016). For example, raters' interpretations of the item "tends to be lazy" about another individual used for the construct Conscientiousness in a popular personality inventory (Rammstedt & John, 2007) varied considerably between individuals, with 47.3% of respondents considering work-related activities, 44.6% inactive life and 36.6% appearance (Uher, 2018b). Similarly, when different participants completed a short version of the personality inventory, their answers revealed considerable variations between them in their interpretation of these items (Lundmann & Villadsen, 2016). Furthermore, participants considered on average only two meanings, which suggests that raters considered only a part of the "field of meaning" (Uher, 2017, 2018b), instead of the whole range of meanings that could be developed for this item statement – different raters therefore considered different referents (Uher, 2021b).

This also applies to SETs. Students who rated item "the lecturer making classes interesting" considered widely different meanings, e.g., how organised was the lecturer but also their perceptions that lessons went by fast (Block, 1998). Interpersonal differences may also affect raters' items interpretations. For instance, U.K. students considered the importance of the items differently depending on their level of engagement. Highly engaged students emphasised "challenge" when interpreting an item evaluating their satisfaction with the course, whereas less engaged students placed importance on the "course being fun" (R. Bennett & Kane, 2014). SETs usually do not involve exploring how students completing them interpreted the item statements and only

some ask about students' academic engagement (e.g., hours spent studying). It is, therefore, unclear how students generate SETs' results and whether these results even relate to the intended study phenomena (e.g., teaching behaviours) or rather to unrelated aspects (e.g., weather, students' individual characteristics). For example, when answering the same item statement, students can consider completely different behaviours of their lecturer (*see Figure 2*). This entails that students encode information about different study phenomena (e.g., 'great and clear feedback', 'weekly tutorials', 'help on assignments') into the same item statement, which precludes data generation traceability.

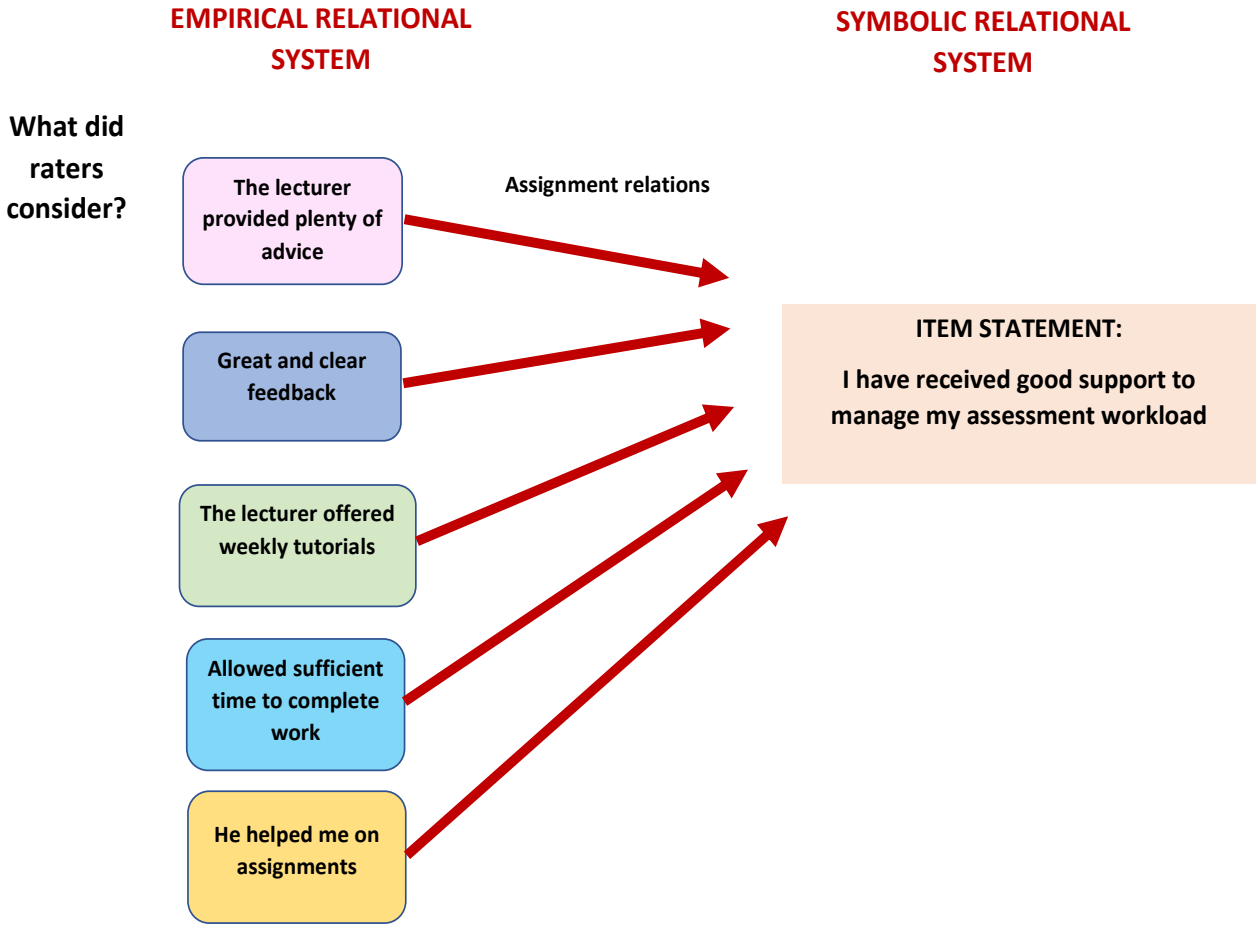

Figure 2

An Example of Different Study Phenomena that Raters Considered for the Same Item Statement

Note. This figure shows a wide range of lecturers' behaviours that student participants reported in my Study 1 when explaining their interpretations of the same item statement.

3.2.4.1.2 Interpretations of raters may differ from researchers' interpretations

Given these variations in item interpretations, raters can also consider different information than intended by researchers (Uher & Visalberghi, 2016). Raters may consider other behaviours of an evaluated person or different situations than intended in the item. For example, a five-method study involving behavioural observations and rating items revealed only 54.1 to 70.4% overlap between raters' and researchers' interpretations (Uher & Visalberghi, 2016).

This also applies to SETs. Scale developers intended to use the item "Information on the module was (not) available" to explore whether the lecturers provided module handbooks. However, students also commented on library resources or difficulty finding a research database (Robertson, 2004). But without asking what students considered in their ratings, this information would be lost and the result misinterpreted.

3.2.4.1.3 Vague and abstract statements may lead to biases

Item statements are frequently vague in order to enable individuals to apply them to many different behaviours or inferred attitudes (Uher, 2018b). This vagueness may, however, result in even more pronounced variations in interpretations. Individuals can also more easily alter their interpretation of evidence in ways to make it consistent with their expectations (Fiske & Taylor, 1991; Heilman, 2012). This means that individuals can be affected by different stereotypes and biases, such as gender stereotypes. Abstract wording of the item statements as well as lack of definition of what study phenomena (e.g., behaviours) should raters consider when rating these items may further increase rather than reduce the manifestation of biases (Uher et al., 2013). The effects of stereotypical biases on raters seem to be so strong that they even affect ratings of individual-specific behaviours in closely related non-human species, as previous methodological studies based on the TPS-Paradigm showed. Specifically, these empirical studies showed that associations with sociodemographic factors (e.g., age, sex) contained in ratings but not in behavioural observations suggest that ratings may have been influenced by stereotypes and attribution bias related to raters' sociocultural beliefs about particular groups of human individuals (Uher et al., 2013; Uher & Visalberghi, 2016).

Similarly, SET items as well are typically worded in general and abstract ways. People frequently tend to evaluate evidence based on their prior beliefs or opinions (*myside bias*; Stanovich et al., 2013). Because interpretations of SET items are left to students' subjective decisions, students may be influenced by beliefs they acquired over time about lecturers, which may contribute to the manifestation of biases.

In sum, variations in raters' interpretations of item statements preclude establishing an unbroken chain of connections between measurands (e.g., different lecturers' levels of 'teaching support') and results (SET scores). Data generation traceability is not established. Therefore, it cannot be ascertained that the results can be assigned to measurands in fully transparent ways and represent valid information about study phenomena.

3.2.4.2 Raters' interpretation and use of answer scale categories

In Likert scales, raters are asked to indicate their level of agreement with item statements. This may involve their judgement of other people's behaviours described through item statements with the use of different answer categories (e.g., "agree"). Therefore, raters need to interpret these answer scale categories and decide which one to choose to indicate their judgement. This process can be affected by several problems, such as variations in how raters interpret and use answer scale categories or conflating mid-categories with non-applicable categories.

3.2.4.2.1 Raters tend to interpret and use answer scale categories differently

Raters are commonly assumed to interpret the answer scale categories in the same ways (McCullough & Radson, 2011; Uher, 2018b). However, when participants rated a protagonist presented on video on an item "is outgoing, sociable" on a five-point agreement scale and provided the reason for choosing a specific answer category, their interpretations of these categories varied considerably between raters (Uher, 2018b). For example, participants were asked to watch a video clip featuring a person and then evaluate them. Fifteen participants (over 19% of the sample) provided "did not see enough evidence" as a reason for choosing "disagree". Two other participants also chose "disagree" because they "found behaviour not genuine", and three further participants reported they "found the behaviour was due to situation not the target person" (Uher, 2017, 2018b). This shows that raters interpreted the same answer scale category in distinct ways.

Similarly, students may choose different answer categories even to encode the same information. For instance, as Figure 3 below illustrates, students may consider the same behaviour as a reason for choosing different answer categories when rating their lecturers on the identical statement. Vice versa, students may also choose the same answer category but consider different information (*see Figure 4*).

3.2.4.2.2 Students may use "mid-scale" and "non-applicable" points interchangeably

Furthermore, some students may tick a scale mid-point (e.g., "neither agree nor disagree") instead of "non-applicable". Students frequently provided scores for non-applicable item statements even if the "non-applicable" answer category was available (Ashby et al., 2011; Robertson, 2004). Numerous students rated the "Feedback I received on my work" item statement even though their module did not involve any assessment, and despite instructions clearly stating to skip this question if the evaluated module contained no feedback (Robertson, 2004). Similarly, when U.K. students rated the National Student Survey items, one of the item statements assessed students' satisfaction with library resources. Some of the reasons for ticking "neither agree nor disagree" were "I don't really use the library. I rely on coursebooks" and "This particular course only requires the course materials" (Ashby et al., 2011).

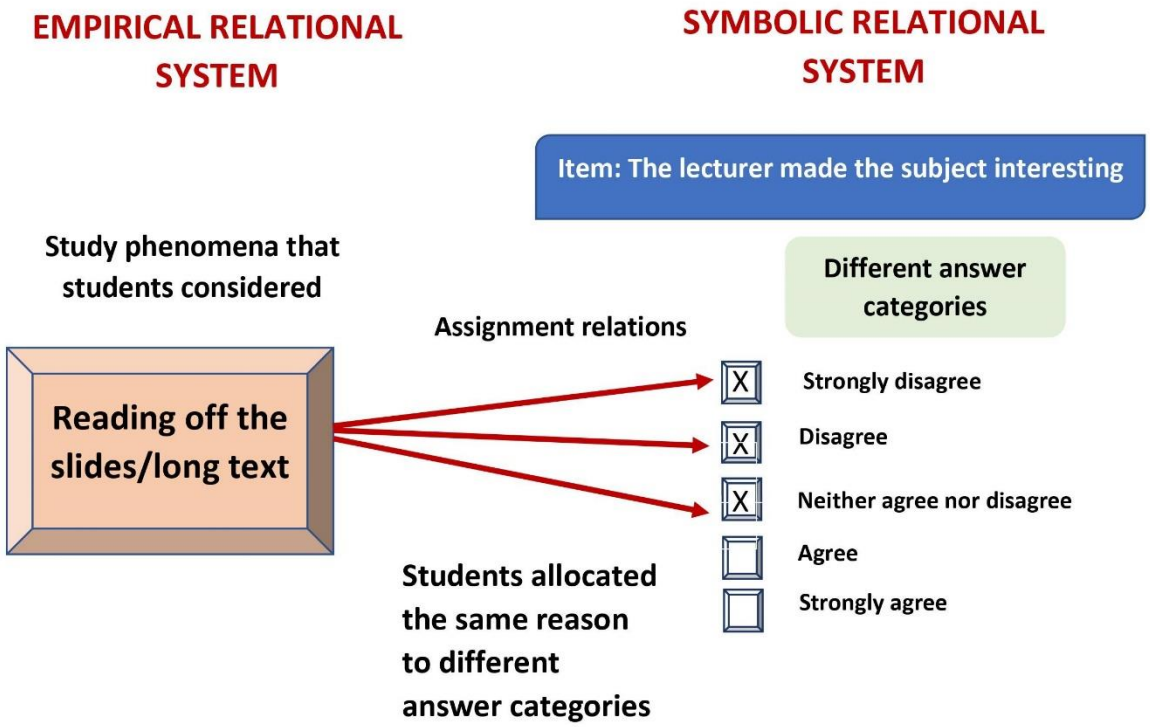

Figure 3

The Same Study Phenomena that Raters Allocated to Different Answer Categories

Note. Student participants in my Study 1 reported they considered "reading off the slides or using long text" as a reason for choosing different answer categories.

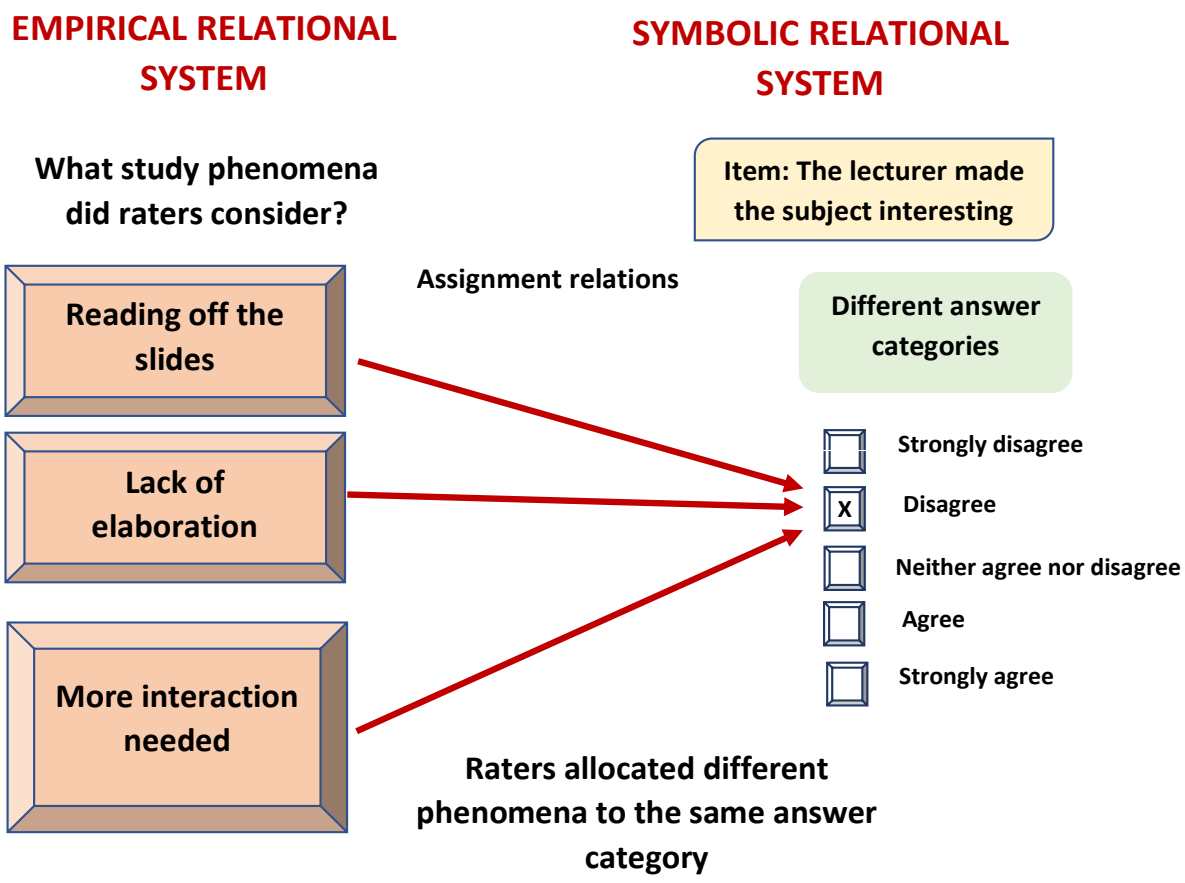

Figure 4

Different Study Phenomena that Raters Considered as Justifications for Choosing the Same Answer Category

Note. Student participants in my Study 1 who chose the same category "disagree" considered different study phenomena.

From the point of view of the two measurement principles, variations in how raters use and interpret answer scale categories and lack of measurement units mean that the results likely do not always represent the same information with regard to the measurands. Instead, these results are influenced by subjective perspectives of raters and quantitative meaning of the numerical values cannot be interpreted unambiguously across different contexts, thus precluding numerical traceability (*see below*). Therefore, public interpretability of the results cannot be established.

3.2.4.3 Numerical recoding of answer categories and analysis of scores

3.2.4.3.1 Definitions of numerals and numbers

Numerals differ from *numbers.* Numerals are signifiers that do not always indicate a quantity but can also indicate categories or labels (e.g., street number, order, phone numbers). In contrast, numbers represent a certain amount or quantity. Qualities are properties that differ in kind, whereas quantities are divisible properties of the same quality (Uher, 2022a; *see also 3.1.3.2*). For example, people colloquially use terms such as the "street number 23". However, "23" is in this example only a numeral because "23" serves as a label but does not refer to a particular quantity of something. Researchers may also decide to use a numeral (e.g., "1") to represent a nominal category (e.g., female students). Importantly, numerals can symbolise numbers only if people decide to attribute to them the meaning of numbers (Uher, 2021a).

3.2.4.3.2 Problems arising from numerical recoding of verbal categories from a measurement point of view

In rating scales, to create numerical data, researchers recode the verbal answer categories, by assigning numerical values to these categories. For instance, each time raters choose a category "strongly disagree", researchers may recode it always into "1". Psychologists often assume that this recoding of answer scale categories enables measurement (Uher, 2021a). However, this widespread assumption is erroneous.

First, this assumption is based on *numeral-number conflation* (Uher, 2022a). Specifically, researchers assume that recoding answer scale categories into numerals gives these scores quantitative meanings. That is, researchers then treat these recoded numerals as numbers with mathematical properties (Uher, 2022a). However, these numerals are merely signifiers without any quantitative meaning and do not represent any specific quantity that participants may have had in mind. Because meanings are never inherent to signifiers, but only attributed to them, numerals depend on raters' interpretations and use of answer categories, as well as researchers' decisions (e.g., about data format).

Second, answer categories recoded into numerals lack a specified measurement unit, as well as universal shared meaning, because their meanings have not been determined through conventions (Uher, 2022a). For instance, "disagree" may be converted into numeral "2", but what exactly "2" means and what quality and quantity it

refers to remains unspecified. This differs from measurement values such as 2 kg or 2 cm, which refer to different qualities (mass, length) and quantities (e.g., 2 mm, 2 cm, 2 m) that may be immediately deduced from these measurement units (Uher, 2021b). Because numerical values without measurement unit essentially lack quantitative meaning, psychologists attempt to create quantitative meanings by differential analyses, specifically, comparing different cases, and considering relative differences between them (Uher, 2022a; *see also 3.2.5*;). However, the scores thus obtained still fail to establish a quantitative meaning for the single measurands, because statistical results depend on the specific sample instead of being based on a known reference quantity.

Consequently, the generated scores cannot be linked to any known reference standards or universally represent the same quantity across different times and contexts. These differences in scores' quantitative meanings therefore prevent making meaningful inferences from comparisons of results and preclude numerical traceability.

3.2.5 Psychometric analyses of 'quality criteria' of rating scales

The analysis of psychometric key quality criteria of rating scales, reliability, and validity, forms an essential part of scale development. Psychometricians must deem both to be sufficient in order to release scales into the public for application. I first briefly define and discuss different types of a) reliability and b) validity of rating scales, provide specific examples from SETs, and then critically discuss reliability and validity in terms of the scientific measurement principles introduced in this chapter.

3.2.5.1 Reliability of rating scales

Generally, reliability is defined as a degree to which the scale produces consistent results. In SETs, analyses usually focus on *inter-rater reliability* and *internal reliability*, which I explore in this section.

3.2.5.1.1 Inter-rater reliability

Inter-rater reliability indicates a level of agreement in judgements between different raters of the same phenomenon. For example, in SETs, inter-rater reliability specifies a degree of agreement between students evaluating their lecturers. Research in SETs produced mixed findings, with some researchers reporting high values considered acceptable by mainstream standards (Polancos et al., 2013), whereas others found much lower levels (Clayson, 2018; Morley, 2012, 2014).

Applying the measurement concepts outlined above to inter-rater reliability highlights, however, that even a high inter-rater agreement fails to reflect whether raters have encoded the same properties of the same study phenomena in the same way into the data (Uher & Visalberghi, 2016). For instance, raters may consider the same phenomena but encode them with different answer categories (*see Figure 3*). Similarly, raters could have used the same answer category (e.g., "agree") but considered distinctive information *(see Figure 4*). Therefore, a clear assignment relation between the empirical relational system (study phenomena) and symbolic relational system (data encoded by raters) cannot be determined. But this assignment relations must always be known to establish documented connection chains from results to measurands. Thus, the results cannot be traced back to study phenomena, which means that data generation traceability cannot be implemented (Uher, 2018b).

In sum, a high inter-rater agreement merely reflects that raters selected identical (or similar) answer scale categories, but assignment relations between empirical and symbolical relational system that raters used remain unclear, resulting in lack of data generation traceability Therefore, this process fails to establish either type of traceability and thus cannot establish justified attribution of results and known meanings required for public interpretability of the results as necessary for measurement.

3.2.5.1.2 Internal reliability

Internal reliability is the consistency of results generated with different items operationalising the same construct. High internal reliability reflects that all the items intended to operationalise the same construct are positively associated with each other and therefore enable the evaluation of the *same* construct (Maltby et al., 2010).

Now I apply the above-introduced measurement concepts to consider internal reliability. To achieve high internal reliability, during the item selection phase of questionnaire development, researchers eliminate items that are associated poorly with the other items on the scale. But the excluded statements may actually capture important further properties of the study phenomena (Uher, 2015b). Wording items slightly differently just to meet particular statistical criteria does not guarantee that we can adequately capture the evaluated behaviours. As a result, researchers risk adapting their data to fit their statistical assumptions rather than considering the actual study phenomena, such as teaching behaviours, thus at the expense of creating data that may reflect these phenomena less accurately. Internal reliability, therefore, only shows how scores of a

specific item relate to scores of other items but does not establish a connection between variations in the actual study phenomena and variations in obtained results. Therefore, the process is *result-dependent*, aligning data generation to statistically optimal results rather than to the actual study phenomena and their properties (Uher, 2021b). This indicates that unbroken traceable chains of connections between measurand and results cannot be established, and the data generated by students cannot be traced back to the lecturers' behaviours, thus precluding data generation traceability.

3.2.5.2 Validity: Do rating scales assess what psychometricians claim they assess?

Reliability is considered a precursor to validity, but high reliability cannot guarantee high validity because an instrument may simply provide consistent data but about a different than the intended construct (Clayson, 2018; Moore, 1990). Validity is commonly defined as the extent to which the scale captures the construct that participants are supposed to evaluate (Maltby et al., 2010). Three types of validity are usually considered in SETs: *content validity*, *construct validity*, and *criterion-related* validity.

3.2.5.2.1 Content validity

Content validity is the degree to which a scale captures all facets of the empirically investigated construct (e.g., Maltby et al., 2017). For example, in the evaluation of 'teaching quality', scale content should represent the sub-domains of 'teaching quality' (e.g., explanation skills) that researchers selected during scale development. Specifically, people may have different ideas about what this construct involves. Therefore, items on the scale should capture various behaviours (such as lecturers answering questions) or inferred attitudes (such as caring) related to 'teaching quality'.

However, SET scales frequently consist of item statements that do not sufficiently capture aspects of 'teaching quality' that students consider important. In a mixed-methods study, students provided what they considered to be characteristics of effective teaching and their definitions. Their answers, however, differed from SET statements used at their university (Onwuegbuzie et al., 2007). Similarly, content analysis of SET items used by over 400 institutions revealed that over 54% of SET item statements were vague, unclear or subjective, and 58% of SET answer scale categories were flawed in terms of being negatively or positively skewed, ambiguous, unclear or not clearly corresponding with the item statement (Tagomori & Bishop, 1994). Thus, the content validity of SETs may often be low.

3.2.5.2.2 Construct validity

Construct validity concerns the degree to which a clear relationship exists between the construct at a theoretical level and the methods used to assess this construct, i.e., between theoretical and operational construct definition (Maltby et al., 2010) . Construct validity has two sub-categories: *convergent* and *discriminant validity* (*see also below*; Campbell & Fiske, 1959). Convergent validity requires empirical associations of scores of the examined construct with scores of conceptually similar constructs. In contrast, discriminant validity is a degree to which scores evaluating conceptually different constructs diverge, which should indicate that constructs are unique and empirically different. These sub-categories are often analysed with a *multitrait-multimethod matrix* (Campbell & Fiske, 1959), which involves correlations amongst two or more methods developed to empirically investigate two or more constructs (Coaley, 2012).

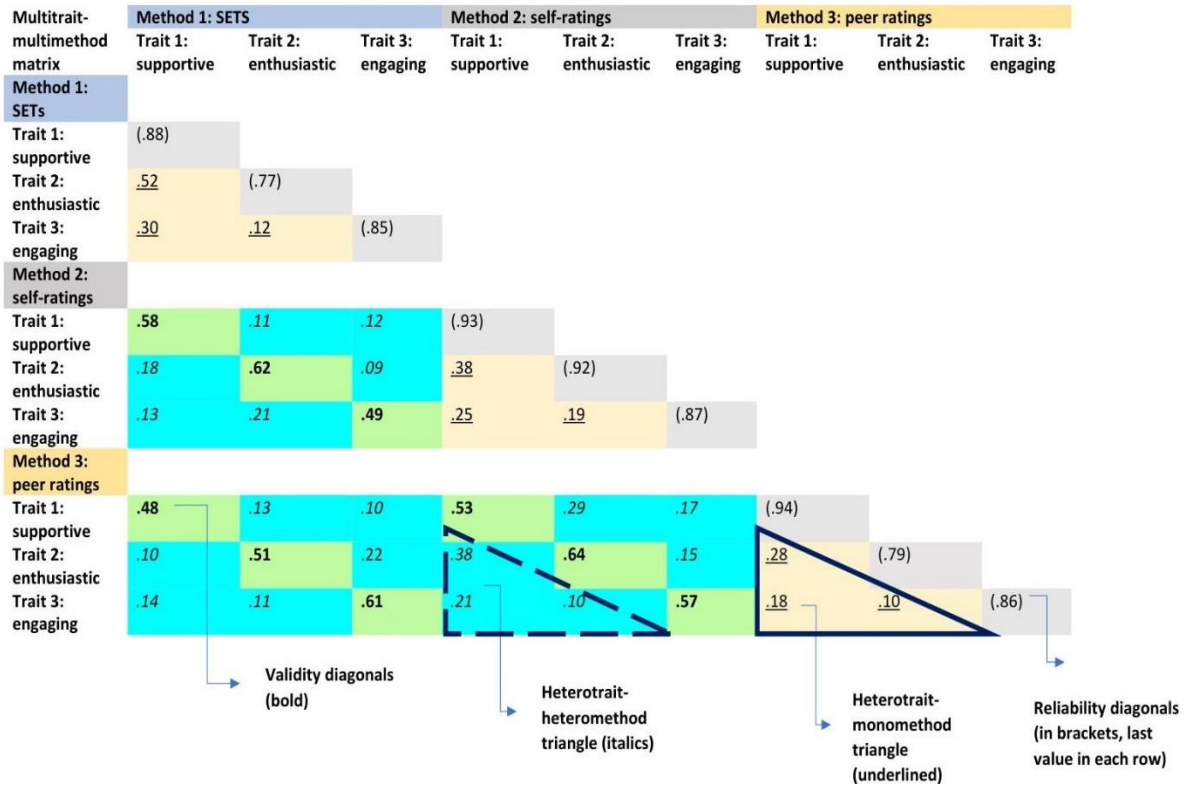

Figure 5

A Hypothetical Example of a Multitrait-multimethod Matrix for the Validation of 'Teaching Quality'.

Note. In this hypothetical example, reliability diagonals are in parentheses. In bold are validity diagonals, which show the correlations between the scores of the same traits (constructs) assessed

by different methods. In italics are values that form heterotrait-heteromethod triangles, specifically, correlations between the scores of the different traits assessed by different methods. Underlined are values that form heterotrait-monomethod triangles, therefore, correlations between scores of the different traits assessed by the same method. The reliability diagonal and the adjacent heterotrait-monomethod triangle make a monomethod block. The validity diagonal and the two hetereotrait-heteromethod triangles (surrounding diagonal) together form a heteromethod block. The presented matrix involves three traits (constructs) related to 'teaching quality' ('supportive', 'enthusiastic', 'engaging'), which are investigated with three different methods (student ratings, self-ratings, peer ratings). The values indicate correlations between traits and methods, all of which must meet specific requirements to establish convergent and discriminant validity (see below for details).

This matrix displays correlations between the scores obtained from multiple methods and various people when evaluating a certain construct (e.g., 'teaching quality'). This may include SETs (students' ratings of lecturers), lecturers' self-ratings (self-assessments) or peer assessments by other lecturers. *Convergent validity* is the degree to which responses obtained by a certain method correlate with responses gained by different methods intended to investigate conceptually similar constructs. To establish convergent validity, validity diagonals, which show the correlations between the scores of the same traits obtained from different methods, must be sufficiently high and significantly different from 0 (Campbell & Fiske, 1959). To examine how these criteria apply to SETs, I use a hypothetical example of a multitrait-multimethod matrix (*see Figure 5*). In the presented example, the values in validity diagonals are all over $r = .40$ and different from zero. Therefore, by mainstream psychology standards, this requirement is fulfilled, and convergent validity is demonstrated.

Discriminant validity can be established if responses obtained with a certain method do *not* correlate with scores from methods intended to investigate conceptually unrelated constructs. To establish discriminant validity, three further criteria must be met. First, a value of validity diagonal must always be higher than the values in the same row and column (in heterotrait-heteromethod triangles). This ensures that a validity value for a certain trait is higher than the correlations between different traits evaluated with the use of different methods (Campbell & Fiske, 1959). This requirement always applies in the hypothetical example (*see Figure 5*). For instance, a correlation between the trait "supportive" evaluated by method 1 (SETs) and method 2 (self-ratings) is $r = .58$ and higher than all the values in relevant heterotrait-heteromethod triangles. Therefore, scores

of the same traits (in this case, 'support') investigated with different methods (SETs vs self-ratings) show higher correlations than the scores of different traits assessed with different methods. Second, a value of validity diagonal for a certain trait should be higher than its values in heterotrait-monomethod triangle. The values of the same trait investigated with different methods should show higher correlations than the values obtained with the same methods that explore a different trait. The second requirement applies for most but not all values in the presented matrix. For instance, the correlation between scores for the trait 'enthusiastic' evaluated by method 1 and method 3 is $r = .51$. But the correlation between traits (constructs) 'supportive' and 'enthusiastic' is $r = .52$. Third, all triangles should show the same pattern of trait interrelationship (Campbell & Fiske, 1959). The third requirement also applies, as overall, the obtained values seem to follow a similar pattern. Therefore, the presented example supports convergent validity and provides some evidence for discriminant validity.

Generally, there is some support for convergent validity in SETs, for instance, evidence for the correlation of SET scores with alumni ratings or lecturers' self-ratings (Marsh et al., 1979). Specifically, responses on SET scales seem to frequently correlate with responses obtained by other conceptually similar methods that were developed to generate data about teaching quality.

I now explore convergent and discriminant validity from the perspective of the above-mentioned measurement principles. The current mainstream approach to examining convergent validity focuses on establishing correlations of the generated results with other results. However, even if results obtained from both methods are highly correlated, these responses may reflect different study phenomena and different properties of these phenomena generated under different conditions. For instance, SET scores provided by students (method A) can highly correlate with peer-ratings provided by fellow lecturers (method B). However, students versus lecturers, given their potentially varying beliefs, values and perspectives, may have interpreted item statements and answer categories in distinct ways and considered completely different teaching behaviours. Importantly, results are not systematically connected to the study phenomena in traceable ways (Uher, 2021b). Specifically, the process through which raters generated data and information raters considered remains unexplored, which precludes data generation traceability. This means that the justified attribution of results to measurands cannot be established. The results may not represent the same information with regard to the measurands in all contexts (e.g., an answer category recoded into numeral "2" represents different scale

categories based on the scale used) and furthermore depend on the subjective perspectives of raters, precluding numerical traceability.

Indeed, different evidence implies low discriminant validity of SETs. For example, SET scores correlated with lecturer's ascribed charisma (Naftulin et al., 1973), enthusiasm (Williams & Ceci, 1997), and likeability (Clayson & Sheffet, 2006), the type of subject they teach (Uttl & Smibert, 2017), or even the weather (Braga et al., 2014), thus, constructs unrelated to teaching quality (*see also Chapter 2*). This implies a frequent lack of evidence of discriminant validity of SETs. From the viewpoint of traceability concepts, it remains unclear whether the values reveal reliable information about the properties of teaching quality or rather completely unrelated constructs.

3.2.5.2.3 Criterion-related validity

Criterion-related validity requires that scores on one scale (predictor) correlate with scores on other sources (criteria), such as certain factors and outcomes which should be related to ratings conceptually. For instance, students who score well on exams during university interviews should also perform well during their studies.

SETs are supposed to enable researchers to evaluate teaching quality, which should correlate with student learning (e.g., as indicated by students' exam scores) chosen as criteria. Therefore, we could assume that students should learn the most from the instructors who score highly on SETs. However, students taught by the lecturers who scored in the middle range of SETs demonstrated the highest amount of learning, whereas students taught by the lecturers with the lowest or highest SET scores displayed the lowest amount of learning (Galbraith et al., 2012). A review of three meta-analyses (Clayson, 2009; P. A. Cohen, 1981; Feldman, 1989) similarly showed that being taught by lecturers with high SET scores was not significantly related to student learning, highlighting that previously reported significant correlations were likely only a result of small sample size effects (Uttl et al., 2017) rather than evidence of a correlation between high SET scores and enhanced student learning. Thus, unlike commonly assumed, SET scores did not seem to correlate with student learning linearly. SET scores therefore either do not appropriately reflect teaching quality or student learning is inappropriate as its criterion. But student learning forms an important outcome of teaching quality. Therefore, SETs may often have only low criterion validity.

3.2.5.2.4 Problems related to validity

Validity is established based on the correlations of scores (e.g., interview performance scores versus future learning scores), yet without considering how the data were actually generated and how they relate to the actual study phenomena (Uher, 2021b). Results obtained, therefore, derive from two or more *different* data generation processes (e.g., one for the criterion, one for the predictor). It follows that each data generation process involves different measurands in different study phenomena (Uher, 2021b). Given that the information these raters considered remains unknown, unbroken, documented chains of results to both a) the measurands b) the reference standards with universally understood quantitative meaning are missing. Therefore, despite high correlations, it is impossible to ascertain whether the results reflect accurate information about the study phenomena and are independent of raters' beliefs, precluding both types of traceability. In sum, even high validity only demonstrates that scores created for different individuals and regarding different study phenomena are related. However, it cannot guarantee that the scientific principles of measurement, which justify attribution of generated results to measurand and establish results' public interpretability, are fulfilled.

3.3 General summary of this chapter

The TPS-paradigm is a novel paradigm that provides detailed conceptual frameworks for studying individuals and phenomena related to individuals, which enable exploring the foundations of measurement and quantification across sciences. To explore whether SETs may contribute to the manifestation of biases, I applied these frameworks, which differ from mainstream ideas in psychology. This investigation revealed that data generation with rating scales such as SET fails to fulfil crucial methodological measurement criteria (justified attribution of the results to the measurand, public interpretability of the generated results' quantitative meaning) and does not qualify as measurement.

SETs involve studying different types of phenomena, such as behaviours, attitudes, constructs and semiotic representations. These phenomena may differ in their location to the studied individual's body (internal vs external) or temporal extension, which determines the accessibility of these phenomena to raters and researchers. Researchers, therefore, must apply appropriate methods suited to the specific type of study phenomena to ensure relevant information about these study phenomena is represented in the appropriate ways in data. A specific kind of data generation concerns quantitative

data, but not all quantifications classify as measurement in the sense of two basic measurement criteria (justified attribution of the results to the measurand, public interpretability of the generated results' quantitative meaning).

The TPS-paradigm, building on metrological theories, defines measurement as a structured process in which numerical values are assigned to the study phenomena in clearly defined and traceable ways, starting from an empirical interaction with the measurand (particular quantity to be measured in target property). This fundamentally differs from the definition of measurement in mainstream psychology, which defines it as "assignment of numerals to objects or events according to certain rules". Over the years, researchers developed different theories of measurement. One of them is representational theory of measurement, developed in the social sciences, which defines measurement procedures as mappings of an empirical relational system onto a symbolic system under certain assignment relations. This theory, however, fails to clearly establish concepts for implementing conditions and rules necessary for measurement. It differs from the principles of measurement underlying metrological frameworks: data generation traceability and numerical traceability. Data generation traceability requires an unbroken, traceable chain of connections that establishes proportional relations between the phenomena under research and the generated outcomes. Numerical traceability requires that the assigned numerical values are also connected to a known quantity standard in transparent and defined ways thereby establishing their quantitative meaning. Implementing these two principles helps to establish crucial criteria of measurement: justified attribution of results to the study phenomena and the public interpretability of the numerical values with regard to the quantity that they indicate.

Applying the TPS-paradigm to SETs highlighted several limitations of standardised rating scales, such as variations in students' item and scale category interpretations which preclude establishing data generation and numerical traceability. Specifically, abstractly worded item statements and answer scale categories mean that students must rely on their intuitive decisions about how to interpret these elements of rating scales. These interpretations commonly vary between raters. In consequence, the same answer scale categories fail to reflect the same information about study phenomena, as needed for justified attributions of results to measurand, thus measurement.

In scale development, theoretical definitions of constructs are frequently vague, overlapping, ambiguous or even missing. This is prominently reflected in SETs by the ambiguity in defining the constructs "teaching quality" and "teaching effectiveness".

Construct operationalisation does not constitute measurement because establishing connections between the theoretical and empirical framework of the construct always involves researchers' interpretive decisions and beliefs and cannot be empirically proven. The blended nature of many constructs also makes it impossible to compare their quantities and to establish unbroken connection chains from results to (potential) quantitative properties in the concrete phenomena.

The common practice of numerical recoding of answer scale categories does not enable researchers to explore how raters interpret and establish assignment relations between the symbolic (e.g., SET items) and the empirical relational system (e.g., lecturer's behaviours). Furthermore, researchers often conflate numerals (signifiers) with numbers, thereby ascribing numerals mathematical and empirical properties. However, numerals are merely labels (signifiers) that can also represent numbers but only if individuals decide to attribute them the meaning of numbers. Because a sign's meaning is not inherent to its signifier but depends on the individuals who are using this sign, the numerals also depend on people who generate data. But measurement results cannot be contingent on the people involved. Instead, results must always represent the same information about study phenomena. Without a connection between the values generated and a known reference quantity, public interpretability of the results' quantitative meaning cannot be determined. This precludes establishing numerical traceability and thus prevents meaningful quantitative comparisons.

SETs generally show acceptable internal reliability but mixed findings of inter-rater reliability by mainstream psychology standards. However, internal reliability only shows how scores of a specific item relate to the scores of other items but does not guarantee that the results provide valid information about study phenomena in terms of their qualitative and quantitative properties. For example, items that provide important information about study phenomena may be excluded to achieve high internal reliability. This practice focuses on obtaining results that are statistically desired rather than results that reflect actual properties of study phenomena. Similarly, even high inter-rater reliability only suggests that raters agreed about the outcome but fails to demonstrate how raters arrived at their judgements and what study phenomena they actually considered, precluding the establishment of a connection chain between the results and study phenomena.

SETs often have low discriminant, content, and criterion-related validity, even by mainstream psychology standards. SETs' convergent validity is generally deemed

acceptable. However, because of differences in students' interpretations of item statements and answer scale categories as well as distinct rating tendencies between students, students' responses may reflect different study phenomena and different properties of these phenomena despite a high correlation between scores obtained by different methods. It follows that even high validity does not implement the traceability principles needed to justify attribution of results to study phenomena and establish results' public interpretability regarding their quantitative meaning.

In sum, psychometric scale development is commonly aligned to particular statistical structures desired for the results rather than to the study phenomena. Neither data generation traceability nor numerical traceability can be established for the data generation processes involved for SETs. Rating scales, therefore, do not fulfil the necessary principles under which data generation could be considered measurement, even if they fulfil psychometric quality criteria. This has important consequences for interpretation of SET scores. These SET scores depend on student raters' subjective interpretations and decisions, which may be influenced by different biases. In sum, methodological and methodical limitations of rating scales discussed in this chapter may contribute to rather than reduce the manifestation of biases. The values generated with the use of rating scales lack established and transparent quantitative meanings as needed for quantitative investigations and therefore cannot be publicly interpretable nor justifiably attributed to the evaluated persons (e.g., teaching behaviours to lecturers).

Chapter 4: General methodology of this PhD research and its two empirical studies

In this PhD research, I aimed to investigate how students use and interpret rating scales and generate ratings and the ways in which students' underlying stereotypical beliefs may potentially affect their perceptions of their lecturers' teaching behaviours and inferred teaching attitudes. The first aim of my research was to explore in what ways students hold their lecturers to potential gendered expectations. Gender-stereotypical beliefs could result in gendered expectations, which may then influence SET evaluations. My further aim was to examine the methodological limitations of standardised rating scales such as SETs and investigate whether they may contribute to the manifestation of gender biases in ratings. These limitations include variations in raters' interpretations of two key components of SET scales, item statements and multistage answer categories, due to their abstract wording and lack of definition, which lead to a lack of data generation traceability and numerical traceability (*see Chapter 3*). If these two traceability principles fail to be implemented, two crucial criteria of measurement—attribution of results to measurands and public interpretability of numerical values—cannot be met.

In the current chapter, I discuss the overall methodology of my project and its two studies. Specifically, I justify the need for this research and demonstrate its originality by identifying a gap in the current literature. In the subsequent section, I outline my research questions and explain how I addressed them. I also describe the epistemological rationale and underlying philosophical assumptions of this PhD research and analyse possible limitations of my chosen stance. Next, I present a general design of this PhD research as well as a detailed overview of its methodological foundations and rationale, including their possible strengths and limitations. Last, I explain how I ensured research quality and addressed potential ethical concerns.

4.1 Rationale and research aims of this PhD research

The overall aim of this PhD research was to critically analyse rating scales as methods of data generation and assess their general suitability for evaluating teaching quality at universities in the U.K. I focused on two crucial problems. Problem A is that students may hold stereotypical beliefs about their lecturers, which may influence SET scores. This may lead to differences in lecturers' ratings unrelated to teaching quality.

Different types of bias (e.g., ethnicity bias, LGBT bias, age bias) can affect students. This PhD research explored bias related to gender stereotypes that may influence students' judgements of their lecturers' teaching behaviours and inferred attitudes. Problem B concerns the application of standardised rating scales and related problems that may stem from their methodological foundations. These problems, which remain largely unexplored, involve individual differences of the raters affecting their interpretation of items and answer scale categories (Lundmann & Villadsen, 2016; Rosenbaum & Valsiner, 2011; Uher & Visalberghi, 2016). This may result in a lack of data generation and numerical traceability, thus questioning the attributability of the generated results to the evaluated people (Uher, 2018a, 2020, 2021c). Problems A and B are intertwined because the methodological limitations of rating scales may lead to the manifestation of biases (e.g., in SETs). Specifically, the standardised format of the SETs (e.g., abstractly worded items, restricted answer format) requires contextualisation and interpretation by respondents, which may increase rather than reduce the influences of unintentional biases on the data generated (Reinsch et al., 2020).

Furthermore, students' opinions about teaching evaluations and the ways in which they approach SETs are largely unknown. It is unclear how much time students dedicate to this task, whether they complete them simultaneously when asked or whether they even find SETs useful in general. I, therefore, also aimed to explore students' perceptions, attitudes and approaches towards SETs in this research.

4.2. Originality of this research

Numerous institutions, including universities, use standardised rating scales to generate a large amount of data in a quick, cheap, and convenient way (Rose, 2005). However, whereas researchers generally invest efforts into enhancing methods of data analysis and establishing reliability and validity of rating scales, it is still unclear how exactly raters accomplish the task of generating data (e.g., Uher, 2018b) and thus how these data relate to the actual study phenomena. Furthermore, little is known about how individual characteristics of raters may affect the generated data, what precisely these data are meant to indicate and whether the produced outcomes may even reflect quantitative information or measurement results as commonly assumed.

Despite the vast body of literature investigating response biases and judgement formation in standardised rating scales, as well as gender bias and gender stereotypes in SETs (*see Chapter 2*), only a limited number of researchers (e.g., Reinsch et al., 2020;

Uher et al., 2013; Uher & Visalberghi, 2016) have proposed that these two problems may be intertwined, specifically, that the methodological limitations of rating scales may increase the manifestation of biases. Nobody seems to have investigated this problem *empirically* in the context of SETs at U.K. universities.

Furthermore, researchers commonly assume that raters interpret item statements and answer scale categories in standardised ways and rarely explore whether raters' interpretations align with researchers' own interpretations. Raters are seldom asked to provide their interpretations of item statements. However, if raters interpret specific questions differently, they are essentially all answering a different question (e.g., R. Bennett & Kane, 2014). Potential variations in raters' interpretations may also contribute to the manifestation of biases during a rating process (*see Chapter 3*).

In my PhD research, I aimed to find out how students generate data on the SET scales. This included exploring students' interpretations of item statements and answer scale categories to investigate whether students interpret these elements in standardised ways as commonly assumed. I also explored whether gendered expectations and gender stereotypes may have affected students' ratings of their lecturers. Furthermore, I examined a relationship between student participants' ratings on the SET items and information students considered when rating their lecturers on these specific items. Regarding the originality of this research, there only seems to be one U.K. study that explored module evaluations in terms of this relationship so far (Robertson, 2004). Unlike Robertson's study, however, my PhD addressed not only a relationship between students' ratings and interpretations of items but also how these interpretations may reflect gender stereotypes.

Moreover, researchers also rarely explore whether a different format of rating scales (e.g., scales starting with "strongly disagree" versus "strongly agree") affects rating processes. Evidence shows some support for bias towards the left side; that is, respondents favoured answer scale categories on the left part of the scale (Friedman & Amoo, 1999). However, what remains unknown is whether raters interpret or use answer categories differently depending on their format. I, therefore, examined how a different format (ascending versus descending order of answer scale categories) of scales affected students' responses in SETs. Researchers in the U.K. used both ascending and descending scales to research SETs (e.g., Brown et al., 2015) but did not specify whether a specific scale format influenced SET results. No previous study in the U.K. so far seemed to have examined whether students' interpretations and use of the answer categories varied depending on the scale direction.

This research involved the application of the TPS-paradigm, which provides, amongst others, integrated conceptual frameworks for analysing research methods (Uher, 2015, 2018a, 2018c, 2019, 2020). No other research had so far applied the TPS-Paradigm to examine the methodological limitations of SETs and how they may lead to the manifestation of gender biases in SETs. Research on gender biases in SETs is frequently either atheoretical (e.g., Fan et al., 2019) or applies theoretical frameworks focused solely on gender stereotypes, such as the stereotype content model (e.g., El-Alayli et al., 2018). Application of the TPS-paradigm enabled the investigation of how students evaluate their lecturers and the potential influence of gender stereotypes using a sound methodological and conceptual framework anchored in several disciplines. In sum, my research provided one of the first extensive examinations of students' interpretations and use of rating scales during completing SETs.

4.3 Research questions

I organised the research questions into four sections: the first section explored students' views on their lecturers' behaviours and attitudes to investigate what type of information students consider important when reflecting on or rating their lecturers' teaching. The next section focused on potential variations in students' interpretations of the key elements of rating scales. I followed by analysing whether gender stereotypes may influence how students judge their lecturers. The last section investigated students' general expectations of their lecturers and students' approaches towards SETs.

4.3.1 Students' views on lecturers' teaching behaviours and inferred attitudes

RQ1: Which of their lecturers' teaching behaviours do students generally consider important?

RQ2: Which inferred attitudes of their lecturers do students generally consider important?

RQ3: Which teaching behaviours and inferred teaching attitudes do students consider in their SET ratings?

4.3.2 Students' interpretations and uses of item statements and answer scale categories

RQ4: In what ways do students interpret the *item statements* of SET scales in general, and what meanings do students construct for these statements?

RQ5: In what ways do students interpret and use the *answer categories* of SET scales when completing teaching evaluations?

4.3.3 Gender differences between lecturers

RQ6: What are possible differences in the teaching behaviours and inferred teaching attitudes that students consider in SET ratings when judging *male* versus *female* lecturers?

RQ7: How do students weight specific teaching behaviours and presumed attitudes when generating their ratings of the lecturers and does it differ by the lecturers' gender?

RQ8: What are potential differences in the ways in which students use and interpret typical answer categories on SET Likert scales when judging *male* versus *female* lecturers?

4.3.4 Students' general attitudes towards SETs and expectations between faculties

RQ9: How do students regard SETs in general, and how do students approach their completion?

4.3.5 Over-arching question: the manifestation of gender biases

R10: Can students' underlying potential stereotypical gender beliefs and expectations of their lecturers influence how students evaluate their lecturers with rating scales?

4.4 Epistemological foundations of this research

Epistemology refers to the area of philosophy concerned with knowledge generation. Specifically, epistemology explores how people acquire knowledge and what type of knowledge can they acquire (Willig, 2008). An epistemological stance is defined by a set of assumptions that researchers hold about knowledge. This stance acts as a philosophy-of-science basis that underpins the research project (DSouza, 2017). Epistemological stance influences a researcher's methodological approach, the research methods chosen to investigate study phenomena, as well as a researcher's approach to data analysis and interpretation (Barbour, 2014; Carter & Little, 2007; Willig, 2008).

Positivist epistemological stance assumes a direct relationship between observed reality and an individual's understanding and perception of this reality (Willig, 2008). Standardised rating scales build on positivist epistemology (e.g., Uher, 2018a), which involves raters' generalised abstract judgements and therefore disregards potential variability in these judgements. It also encompasses deterministic input-output models that capture raters' reaction to items but discount raters' roles in the construction of meanings for these items (Uher, 2021d). However, the process through which people acquire knowledge, as well as interpret and perceive the social world, is influenced by

their beliefs and mental constructions (e.g., Hinton, 2017; Oakes & Turner, 1990; Uher, 2021d; Westra, 2019). This entails a risk of anthropo-centric, ego-centric and ethno-centric biases, which occurs when an individual's own role in the world (e.g., in terms of their own gender or socio-cultural background) affects how they perceive and conceive phenomena (Uher, 2015c). For instance, when students indicate their teaching evaluations on standardised rating scales, students' interpretations of their lecturers' behaviours and inferred attitudes may be influenced by students' own ideas and beliefs about gender and thus be susceptible to biases.

To explore those influences, the epistemological position I applied to investigate students' judgements about their lecturers' behaviours and attitudes in this research is *social constructionism*. This approach emphasises the active role of persons who construct and experience reality through the use of language. However, the focus is neither on the individual experience nor on the essential nature of reality, but rather on the processes through which people construct knowledge about their social world (Willig, 2008). Many researchers use terms 'social constructionism' and 'interpretivism' interchangeably. Others noted some differences between these positions, such as more defined approaches to reflexivity in social constructionism or differences in the role of language, seen as a 'tool' in interpretivism but taking a central role in 'social constructionism' (e.g., Y. Y. Chen et al., 2011). Specifically, in interpretivist approaches, researchers usually see the language as a means used to describe one's experience. In social constructionism, researchers focus on how individuals use language to construct their perceptions of reality. For example, the same experience or object can be described in either a positive or negative way (Burr, 2003; Y. Y. Chen et al., 2011; Willig, 2008). The core idea of social constructionism is that human experience is mediated by language, putting language into the central focus. I acknowledge these differences between interpretivism and social constructionism but perceive them as subtle. Both positions emphasise the importance of meanings and social context, as well as the active role of participants and researchers. Therefore, I conceptualise these positions as synonymous. In my research, I analysed participants' individual interpretations of statements, but my focus concerned exploring broader patterns in the ways in which participants constructed meanings in this social context.

Consistent with the social constructionism approach, I made the following assumptions in my PhD research:

a) Gender, gender roles and beliefs heavily depend on context. Specifically, gender roles and beliefs are constructed historically and socially instead of constituting the inherent traits of men or women (Baber & Tucker, 2006). Gender is therefore seen as a social process heavily shaped by social interaction (Marecek et al., 2004).

b) How individuals perceive certain phenomena depends on their pre-existing schema, making perception a constructive process (Bem, 1981). For instance, people derive meanings and a sense of reality from the socio-cultural categories available to them (Marecek et al., 2004). Moreover, individuals use these categories to memorise information and later recall memories of originally perceived phenomena through these socio-cultural categories (Bem, 1981). People's choice of words could therefore reflect gender stereotypes (Sprague & Massoni, 2005).

c) Social context differs by culture (Burr & Dick, 2017). Specifically, social concepts and categories do not hold universal meanings but are unique to each place and time (e.g., Marecek et al., 2004). People from different backgrounds may therefore interpret meanings in various ways. For example, in Study 1, I explored whether students interpret commonly used SET item statements in standardised ways, as frequently assumed. In Study 2, I examined potential variations in students' interpretations of answer scale categories.

Social constructionism is frequently described as "relativist", specifically, based on the assumption that reality is subjective. This position emphasises the diversity of people's interpretations of the world, which leads to the potentially controversial implication that definite truth about the nature of social reality does not exist (Burr & Dick, 2017). Specifically, this approach denies the existence of the "universal" or absolute truth about the social world. Instead, it implies that truth is affected by relations within society (e.g., within different groups) and their relationship with language. For instance, people can construct categories differently based on their socio-cultural background, making their knowledge about the world relative. Therefore, researchers adopting the social constructionist approach must accept that people are likely to adopt multiple perspectives on any given situation, event, or object (Burr & Dick, 2017). This forms one of the central ideas to my research.

In this PhD research, I applied a *moderate* (less relativist) version of social constructionism. A moderate version differs from a radical one, which focuses on how individuals construct meanings in particular social contexts (Willig, 2008), but can lead to an assumption that everything is a social construction and there is no objective knowledge about the world. This may lead to a paradox as this version of social constructionism undermines its own claims (Elder-Vass, 2012). In contrast, a moderate version aims to explore the connection between the construction of social reality and the socio-cultural context. This version can be seen as less relativist because it partially accepts the role of context. Specifically, this approach acknowledges that the pre-existing social reality may influence individuals and the ways in which they construct meanings about their reality (e.g., through language; Willig, 2008). Moderate version of social constructionism differs from another approach, *critical realism*, in several aspects. Critical realism claims that individuals are bounded by both physical and social world. Specifically, it assumes the existence of the real world but recognises a socially constructed world as well (S. P. Taylor, 2021). Critical realism acknowledges that multiple perspectives may exist and be influenced by social context, but also aims to explore hidden underlying mechanisms that may shape events but exist independently of human perception. According to this position, the real world exists, and individuals can interpret it in different ways but these interpretations may differ, with some of them being better connected to a reality beyond words than others (Robinson, 2022). Furthermore, it is highlighted that our interpretations of a real world are likely to be imperfect or partial (Robinson & Smith, 2010), for example, we may be limited by our cognitive abilities when trying to access or understand a reality (Uher, 2018a). Therefore, researchers adopting this approach usually seek to explore both the real and social world. In contrast, moderate social constructionism also accepts that the real world exists, but this acknowledgment occurs on a more superficial level, as there is no commitment to fully explore this world. Instead, social constructionism focuses on how individuals construct knowledge and meanings through the use of language (Fopp, 2007) and on their interaction with the social world.

In my PhD research, I acknowledge that participants experienced real events (e.g., teaching behaviours, lectures, provided materials) and critical realism approach could be used to explore these events. However, I emphasise that students' interpretations of these events may be influenced by subjective factors (e.g., students' previous experiences) but also social context, (e.g., interactions between lecturers and students may differ between cultures). I also assume that gender, one of the key elements studied in this research, is

socially constructed and, through my research questions, I scrutinise the ways in which participants construct meanings for item statements. In sum, I acknowledge the existence of a real world, but my research focuses on students and exploring their interpretations. Therefore, I chose moderate social constructionism as the most suitable approach.

4.5 General design of this PhD research: Two complementary studies

This PhD involved two empirical studies. Study 1 explored students' potential gendered expectations of their lecturers and variations in students' item interpretations. Study 2 examined how students interpret and use the answer categories of typical rating scales and how they form their overall judgements when evaluating their lecturers. Based on my epistemological stance, I chose to apply a *multi-method design* in this PhD research *(see 4.6.1)*. Study 1 and Study 2 each involved different methods of data collection and several methods of data analysis, both qualitative and quantitative, in order to simultaneously address the same research questions from different perspectives (Morse, 2010). Qualitative data consisted of participants' open-ended answers, and quantitative data involved participants' ratings of standardised SET items, as well as percentages and numbers of participants providing answers categorised into the same sub-theme or theme. In both studies, data generation methods involved online surveys with a) open-ended questions, b) a set of predetermined item statements on which participants rated their lecturers indicated in answer scale categories. Study 1 involved students' reflections on specific lecturers (real-life persons students had encountered in their studies), whereas Study 2 involved two scenarios about fictitious lecturers.

Table 1

Overall strategy for addressing research questions in the two empirical studies

Research questions	Information collected in Study 1	Information collected in Study 2
RQ1 – considered teaching behaviours	Participants' reflections on teaching behaviours they considered important	N/A
RQ2 – inferred teaching attitudes	Participants' reflections on teaching attitudes underlying teaching behaviours	N/A
RQ3 – teaching behaviours and attitudes outlined in reasons for participants' ratings	Teaching behaviours and teaching attitudes participants considered in ratings of their own lecturers	Teaching behaviours and attitudes participants considered in their ratings of fictitious lecturers

RQ4 – meanings of item statements	Participants' general interpretations and constructed meanings for selected SET item statements	N/A
RQ5 – interpretations and use of answer categories	N/A	Participants' specific interpretations and reasons for choosing SET answer scale categories + their ratings of fictitious lecturers
RQ6 – perceived gender differences between lecturers' teaching behaviours and attitudes in SETs	Teaching behaviours and inferred attitudes of specific male versus female lecturers that participants considered in their ratings	Reasons that participants considered in their ratings of fictitious lecturers + answers to an open-ended question
RQ7 – relations of information students considered and SET ratings + potential gender differences between lecturers in students' ratings	Teaching behaviours and attitudes participants considered and their influence on participants' ratings + participants' ratings of real-life lecturers (rating scores)	N/A
RQ8 – perceived gender differences in students' use of answer scale categories	N/A	Reasons that participants considered in their ratings of fictitious lecturers + their influence on participants' ratings of lecturers + potential differences in participants' ratings of male and female lecturers
RQ9 – students' approach to SETs	Participants' reflections on their approach to SETs, reported time invested and general opinions of SETs	N/A
RQ10 – stereotypical gender beliefs that may affect and influence the rating process	Participants' reflections on teaching behaviours and attitudes considered in their ratings + students' ratings of their specific lecturers	Reasons that participants' considered in their ratings of fictitious lecturers and their influence on participants' ratings

Because SETs generally involve standardised ratings, I used rating scales in both studies to examine the rating process. Specifically, in Study 1, I explored information that students had in mind when rating a lecturer from their university studies. In Study 2, I examined whether students' ratings of their lecturers differed by lecturers' gender and whether the different organisation of the answer categories in rating scales (starting with either "strongly disagree" or "strongly agree") impacted rating scores that student participants gave their lecturers.

The overall design of this research was qualitatively driven, which means that a qualitative approach formed the core of this project. In sum, I analysed the methodology of rating scales as a quantitative method of data generation with primarily qualitative

methods. Specifically, qualitative data enabled the analysis of potential shortcomings of a standard quantitative method. Furthermore, qualitative data enabled a better understanding of what information students consider when evaluating lecturers and supporting quantitative data provided a clearer picture of how exactly students use rating scales. The sample consisted of the students (or people who completed their studies less than a year ago) at different universities in the U.K. from a wide range of disciplines. Table 1 above represents a matrix to describe Study 1 and Study 2 in terms of what research questions were answered with which collected information.

4.6 Methodological foundations and rationale

4.6.1 The rationale for the present multi-method design for data analysis

The purpose of both studies was to "obtain different but complementary data on the same topic" (Morse, 1991, p.122). This was done using a multi-method approach, in which qualitative and quantitative methods are both concurrently applied (Gelo et al., 2008). The multi-method approach involves using different data types or methodologies to draw on the different methods' strengths when exploring research questions (Domanski, 2004; Gelo et al., 2008). Researchers frequently use the terms 'multi-method design' and 'mixed-method design' interchangeably. In my PhD research, multi-method design refers to the use of multiple methods, which I chose from the start without yet knowing the outcome of the other research methods. Because my aim was to use qualitative methods to critically scrutinise the use of quantitative method, I applied these methods concurrently rather than consecutively and without using the outcomes of one to inform the design of the other as typical for mixed methods.

I applied predominantly qualitative methods to explore how raters generate data, use and interpret item statements and answer scale categories on rating scales (a standard quantitative method). This approach allowed for the investigation of rating behaviours, but the integration of different methods also provided unique insights and perspectives regarding the researched phenomena (Hammond, 2005), such as the teaching behaviours and attitudes students considered important but also students' potential gender biases. For instance, these biases could then be explored with both quantitative (rating scores) and qualitative methods (open-ended questions). Qualitative data were crucial because the type of information (e.g., students' reflections on their lecturers' behaviours) that students consider when rating their lecturers and students' interpretations of item statements and

answer scale categories had to be studied through open-ended questions. This approach allowed me to explore potential variations in students' answers and investigate how students interpret and use rating scales. Importantly, collecting both qualitative and quantitative data permitted me to analyse the connections between participants' answers and rating scores, therefore providing essential insights into how students generate their ratings.

I also obtained students' interpretations of item statements and answer scale categories on SET rating scales which students used to evaluate their lecturers (open-ended answers formed qualitative data) and student ratings on specific SET items (rating scores produced "quantitative" data). I analysed students' open-ended answers with thematic analysis in which I quantified occurrences of the themes to determine their frequency. Students' ratings (scores) could be analysed quantitatively. The multi-method approach in my research enabled obtaining students' quotes about specific teaching behaviours and teaching attitudes students considered important when rating their lecturers on standardised scales.

Each research method has inherent flaws (e.g., McGrath, 1981; Scandura & Williams, 2000). A multi-method approach can partially solve this problem because the strengths of one method could counterbalance the weaknesses of another method (S. F. Turner et al., 2017). However, potential limitations of the multi-method approach may arise from requirements that researchers have skills and expertise in applying and using both chosen methods. Because my multi-method design involved both qualitative and quantitative data, preparing for conducting this research and analysing multiple data sets has been likely more time-consuming than research that would be either qualitative or quantitative (DeMarrais & Lapan, 2003). However, broadly researching and reflecting on the topic can help develop the researcher's independent critical thinking, strengthen the study's methodology (Kahlke, 2014) and gain important insights otherwise not achievable. On the one hand, if applied methods yield different findings (e.g., rating scales versus open-ended questions), a researcher must find a way to understand and explain these differences (Jick, 1979). On the other hand, the application of several methods in study design can help generate a better understanding of the study phenomena (S. F. Turner et al., 2017). For instance, this type of design may reveal characteristics of phenomena not yet incorporated in theory (Jick, 1979).

This is highly relevant to my research because how people generate rating data and how students use and interpret SETs remains largely unknown. In my research, qualitative

information revealed what teaching behaviours and inferred attitudes students considered, as well as students' interpretations of answer scale categories and item statements. Quantitative information enabled an understanding of how this information influenced ratings, as well as exploring how widespread were particular ideas in my sample.

4.6.2 Quality assurance of research

Quality assurance refers to various approaches intended to develop and maintain standards of science and higher education (Dill, 2009). Because I evaluated the quality of the standardised rating scales with predominantly qualitative methods, I considered the most relevant quality criteria of both a) *survey research in psychology* (Protogerou & Hagger, 2020) and b) *transparent qualitative research*, which included the rationale for the chosen qualitative data analysis method, transparent description of the coding process (Aguinis & Solarino, 2019), credibility and reflexivity. Several of these quality criteria applied to both qualitative and quantitative elements of my research and are therefore presented below.

4.6.2.1 Quality criteria applying to both quantitative and qualitative research

4.6.2.1.1 Position of a researcher along the insider-outsider continuum

I conceptualised "insider-outsider" status as a degree to which I felt like a member of the group based on my shared status or experience with participants (Gair, 2012), as well as my relationship towards them (Aguinis & Solarino, 2019). I adopt the view of Eppley (2006) in rejecting a binary approach to insider-outsider positioning and perceive it rather as a continuum. As a PhD student at the U.K. university, I placed myself closer to the "insider" position on the insider-outsider continuum in terms of my student status (Aguinis & Solarino, 2019). Specifically, my own subjective experiences as a student may have affected how I perceived and analysed the generated data, which could differ from the perspective of persons who do not identify as students. Similarly, I would generally be closer to "insider" in terms of gender because most participants identified as women. I am also "outsider" in some aspects, such as being an E.U. student, whereas majority of student participants were U.K. students. However, because this research involved also participants identifying as men or non-binary, as well as International and E.U. students,

my position in this continuum differs based on the demographics of specific participants and can therefore be perceived only in this generalised context.

In both studies, I had no direct contact with participants, who remained anonymous throughout the surveys. I provided my email on a consent form to give participants an option to contact me to ask any questions before starting and after finishing the survey. No participant contacted me with questions before completing the survey, but several contacted me afterwards, with positive feedback or requesting a future update about my findings.

4.6.2.1.2 Sampling approach and procedures

In both Study 1 and Study 2, participants were students from across U.K. universities (or people who completed their studies less than a year ago) of different ages and genders. I chose this target group because of its representativeness, as SETs at universities are always completed by students.

Study 1

In the first stage of recruitment, I used opportunity sampling, in which participants are selected due to their availability, by advertising this study on websites designed for that purpose (e.g., SurveyCircle, Sona systems)[9], as well as during one lecture at the University of Greenwich. This type of sampling may, however, potentially limit diversity in the sample. For instance, most participants recruited during the first stage of Study 1 had psychology academic backgrounds. In the second stage of recruitment, and to enable recruiting participants from various disciplines to explore different viewpoints and reduce potential sampling bias, I used purposive sampling. This type of sampling refers to maximum variation, specifically, collecting a wider range of participants (e.g., from different academic backgrounds) to capture different viewpoints about the phenomena. My purposive sampling strategy involved contacting the Directors of Research of disciplines outside psychology at the University of Greenwich, one of which forwarded my advertisement to his students.

[9] Sona system is a research participation scheme also used by the University of Greenwich. Student participants recruited through Sona system receive Sona points (credits) for their participation.

Study 2

The aim in Study 2 was to recruit a large number of participants to improve the representativeness of findings. To achieve this, I used opportunity sampling combined with snowball sampling, in which a study participant recruits other participants. Snowball sampling, which involved inviting participants to advertise the study in their study groups, enabled reaching a high number of students, and capture diversity as shown by participants' various academic backgrounds.

4.6.2.1.3 Sample size and data saturation

Qualitative researchers generally tend to recruit a small size of participants due to the time-consuming nature of data collection and analysis (e.g., Willig, 2008) and because they focus on the richness of data rather than large sample size typical for quantitative research (e.g., Migiro & Magangi, 2011). However, in some types of qualitative research, such as brief text research (e.g., qualitative surveys), the richness of data stems from the diversity of participants' responses rather than their depth (Robinson, 2022). Furthermore, qualitative researchers may intend to reach some degree of representativeness if the explored phenomena apply to more people than the study participants (Willig, 2008). I aimed to produce some general knowledge on how students generate ratings and use and interpret rating scales, and, therefore, achieve a certain level of representativeness in my research. Because I explored a standard quantitative method with predominantly qualitative methods, my student participants provided numerous but relatively brief answers when explaining their interpretations of different item statements and answer categories. I, therefore, needed to recruit a sample size larger than would be typical for a predominantly qualitative project.

The sample size is closely related to the *saturation point*, which is reached when the information provided by participants becomes repetitive and fails to provide any novel insights (Bowen, 2008). In both studies, I read generated data already during the data collection process. I continuously considered the collected information and reflected on whether participants' answers still appeared to provide new insights. I ended the data collection when the answers seemed to become repetitive and unlikely to yield novel information about phenomena.

A general recommendation for sample size in qualitative surveys conducted for a PhD project and analysed by thematic analysis is 50+ participants for a part of the project and 200+ for the whole project (Šula, 2018). In line with this recommendation, but also guided by saturation point, I recruited 181 participants for Study 1 and 336 participants for Study 2. This enabled me to obtain a diverse range of participants' answers and achieve a sufficient degree of representativeness.

4.6.2.1.4 Research setting

I set up both studies online. I conducted Study 2 during the COVID-19 pandemic, which hindered the use of in-person research settings I originally planned. The online setting might exclude those students who dislike or are reluctant to use online technology. However, numerous universities started implementing completing SETs online (Lowenthal et al., 2015), and during pandemic, institutions worldwide transitioned fully to online mode of teaching and related activities (Peimani & Kamalipour, 2021). Another potential limitation is that technical problems (e.g., loss of network connection) could also impact participation (Lefever et al., 2007). However, the advantages of online research settings were complete anonymity, which may have allowed participants to describe their experiences with their lecturers freely, as well as convenience in terms of time (Lefever et al., 2007). Furthermore, online research settings likely enabled recruiting a larger (and thus more representative) sample compared to in-person research settings (Robinson, 2022). The online setting also allowed recruiting participants from different universities in the U.K., and therefore different locations, which improved the sample's representativeness. If I had used an in-person setting, such recruitment would not have had been feasible due to time and travel constraints. The use of online surveys may have also enabled reducing my own potential biases that could affect participants because, unlike with interviews, participants had no direct interaction with me (Sargeant, 2012). Therefore, participants were less likely to be influenced by any of my leading research questions, my potential beliefs or biases about gender, gender roles, or gender stereotypes.

4.6.2.1.5 Research materials

Researchers should always report research materials (e.g., rating scales) in full, including item statements used in the study (Protogerou & Hagger, 2020). I initially researched what SET items are used in various U.K. universities. Ultimately, I derived the item statements from the SET questions used by the University of Greenwich, according

to two conditions, a) item statements that were repetitive and frequently used also by other universities, b) item statements that seemed directly related to lecturers' behaviours and thus potentially more open to interpretation. For instance, item "Feedback has helped me develop and improve my performance" would be chosen instead of "Feedback on my work has been timely (normally within 15 working days)" because the latter would likely generate more homogenous responses, whereas the former item statement is less specific and may result in varying interpretations. When necessary, I slightly adapted item statements to suit the context (e.g., "Their feedback has helped me develop and improve my performance" instead of "Feedback has helped me develop and improve my performance"), to avoid students simply commenting on the usefulness of feedback in general (*see also Chapter 6*). The specific item statements are enclosed with each study.

4.6.2.2 Criteria specific to quantitative research

Researchers usually assess the quality of quantitative research by examining its validity and reliability, such as that of applied research instruments (e.g., rating scales; Cameron, 2011; Frambach et al., 2013; Golafshani, 2003). Most psychologists assume that establishing validity and reliability would allow them to replicate and generalise their findings. These concepts are strongly rooted in a positivist paradigm (Cameron, 2011; Carminati, 2018).

4.6.2.2.1 Validity and reliability

I derived these established rating scales from the U.K. universities. Importantly, I did not apply these scales to gather information about teaching quality, instead; I explored how students interpret and use rating scales, to examine any potential problems with their validity and reliability with the use of predominantly qualitative methods (*see also Chapter 3*). I provided a full overview of the used item statements in Chapters 5 (Study 1) and 6 (Study 2).

4.6.2.2.2 Methods of quantitative data analysis

My research contained two types of quantitative data: a) participants' ratings of their lecturers, b) themes and sub-themes of teaching behaviours, inferred attitudes, and interpretations of items, which frequency in the sample I quantified. Initially, I obtained participants' ratings with a shortened version of standardised rating scales generally used by the U.K. universities (*see Research materials*). This approach enabled me to critically

analyse and empirically investigate potential methodological or methodical limitations of rating scales.

To explore how participants weighted the information they considered in their ratings of lecturers, I ran bivariate correlation analyses. Multiple tests may lead to a higher risk of Type I error, which occurs when the null hypothesis is rejected despite being true (i.e., false positive). Researchers proposed different correction methods to control for this risk, of which the most commonly used is Bonferroni-procedure. However, the Bonferroni method is conservative and may lead to over-adjusting p-values and false negatives (Menyhart et al., 2021). In contrast, the Benjamini-Hochberg procedure is less stringent, it aims to limit false negatives, and is suited for exploratory research (Benjamini & Hochberg, 1995). This procedure consists of ranking of p-values in ascending order and calculating Benjamini-Hochberg critical value for each p-value with the use of formula: a rank of each p value is divided by the overall number of analyses and multiplied by False Discovery Rate. This error rate is chosen by researcher and may be higher than 5%. I applied False Discovery Rate of 10% in Study 1 because of its exploratory nature. In Study 2, I built upon the insights from Study 1 and therefore applied more stringent False Discovery Rate of 5% in order to minimise Type 1 error. Once the highest p-value is found to be lower than the calculated critical value, all the previous (lower) p-values are considered significant (Benjamini & Hochberg, 1995; Menyhart et al., 2021).

To further examine these limitations, I then quantified the prevalence of the identified themes and sub-themes to explore potential variations in students' interpretations of the item statements and answer scale categories studied. Finally, I calculated the occurrences of each theme and sub-theme in the sample (e.g., even if the participant mentioned a certain sub-theme several times, I only counted this occurrence once per item).

To achieve transparency and establish a clear link between the research questions, study phenomena, and data analyses, I justified each statistical technique I used and explained what knowledge I aimed to gain (*see Chapters 5, 6)*.

4.6.2.2.3 Measurement principles

Although two measurement principles are generally not used as quality criteria in quantitative psychology, I now consider my analysis from the viewpoint of these principles. Unlike participants' ratings, the quantified frequency of the themes fulfilled a requirement of data generation traceability. Specifically, I established an unbroken link of

connections between themes or sub-themes and produced frequencies. The quantitative results that I generated could be traced back to the particular properties (themes and sub-themes of considered teaching behaviours, attitudes, and interpretations of items) in clear and transparent ways, which established object-dependence. Similarly, numerical traceability was established, because the generated quantitative results were linked to conventional objective standards and were independent of people interpreting or using them, in that certain percentage of participants always referred to the same value in this study, thus implementing subject-independence. For instance, 50% of participants considering content related to the theme "availability and quality of content" when rating structure and organisation of the module, can be traced to all specific instances in which they mentioned content related to this theme and, therefore, a specific number.

4.6.3.3 Criteria specific to qualitative research

Qualitative research applies different procedures than those standardised in quantitative research (Willig, 2008). This can make it difficult to assess a study's quality, which may undermine research findings (Mays & Pope, 2000). Quality assurance can help establish the trustworthiness of qualitative research and diminish possible concerns about its usefulness and validity (Collingridge & Gantt, 2008). Quality assurance of qualitative research consists of two important elements*: the authenticity of the generated data* and *the trustworthiness of the analysis* (Sargeant, 2012). Ensuring both advances the rigour and overall quality of research. To address the trustworthiness of the data, I clearly described the analytical processes[10], the procedure for resolving differences in findings between team members (Study 1) and discussed how my biases may have influenced research (Sargeant, 2012). Specifically, I documented how my beliefs, biases or subjective opinions may have affected the ways in which I conducted this research (e.g., during data interpretation; *see also 4.6.3.3.3*). To establish the authenticity of data, which concerns data quality and sampling procedures (Sargeant, 2012), I ensured that the sampling approach contributed to answering the research questions and was selected in a way that would reduce sampling biases.

[10] I elaborate on analytical processes in Chapters 5 and 6.

4.6.3.3.1 Methods of qualitative data analysis

I used thematic analysis, which is a flexible method usually applied to detect patterns and meanings in qualitative data (Braun & Clarke, 2006). Thematic analysis is suitable for analysing large datasets with over 60 participants and allows analysis of any kind of data, including qualitative surveys (Vinet & Zhedanov, 2011). Both studies involved predominantly qualitative data: students' reflections on behaviours and attitudes as well as interpretations of item statements (Study 1) and answer scale categories (Study 2). Because Study 1 involved 181 participants and Study 2 included 336 participants, the thematic analysis was well suited for my datasets. Thematic analysis also has theoretical flexibility and is not tied to any particular epistemological position (Willig, 2008). This method, therefore, could be used in my research, in which I applied the social constructionism approach.

However, the relative accessibility of thematic analysis may sometimes lead to a lack of actual analysis when a researcher reports data extracts but does not interpret these data to answer the research questions (Terry et al., 2017). Furthermore, because of its theoretical flexibility, researchers must clarify their epistemological position to make thematic analysis meaningful (Willig, 2008). To avoid these pitfalls, I clearly established my epistemological approach before conducting my research and followed the criteria by Braun and Clarke (2006) on how to conduct thematic analysis well. Specifically, I ensured that the coding process was thorough, data have been actively analysed with each response being given sufficient attention, and I clearly and consistently reported the themes and connected my findings to the original data.

I applied standard thematic analysis, which involves six steps: familiarising with data; generating initial codes; searching for themes; reviewing potential themes; defining and naming themes; and producing the report (Braun & Clarke, 2006). In Study 1, I analysed the open-ended answers data with thematic analysis and quantified the occurrence of particular sub-themes and themes in my sample. In Study 2, I applied most principles from structured tabular thematic analysis (ST-TA), which positions itself on the spectrum between thematic analysis and approaches to analysing brief texts (Robinson, 2022). This method combines a qualitative approach with limited quantification and is highly suitable for the analysis of brief open-ended answers by many respondents and, therefore, fully appropriate for this study.

This approach shares similarities with standard thematic analysis by Braun and Clarke (2006). Both seek meaningful reoccurring patterns in data and allow flexibility because of their compatibility with deductive/inductive and manifest/latent analyses. The structured tabular thematic analysis (ST-TA) builds on assumptions that supporting qualitative findings with numbers (e.g., of theme frequencies) may enhance data analysis's overall quality and precision (Robinson, 2022). Therefore, unlike standard thematic analysis, it provides guidelines for calculating theme frequencies. These frequencies provide important information about the prevalence of themes and facilitate the ways in which readers interpret findings. Another difference concerns calculating inter-analyst agreement, which was criticised by Braun and Clarke (2019) as enforcing a positivist agenda on qualitative research. However, researchers may check inter-analyst agreement in order to enhance transparency, clarity and rigour of analysis rather than to establish objective facts. ST-TA could also be seen as similar to qualitative content analysis, which may be used to analyse brief texts. However, whereas content analysis emphasises frequencies, ST-TA focuses on establishing meaning and context of the themes, such as by illustrating them with quotes, with quantification playing a supportive rather than main role (Robinson, 2022).

A potential disadvantage of ST-TA concerns sampling. Due to its emphasis on breadth and diversity rather than depth of responses, the sample may frequently be larger than for in-depth qualitative studies. Researchers should, therefore, consider flexible sampling approaches that may differ from (solely) purposive sampling strategy frequently applied in interview studies with smaller samples (Robinson, 2022). I used opportunity sampling combined with snowball sampling. Participants who took part in my study that involved a considerable amount of writing may not represent typical university students, which presents a potential risk of sampling bias. For instance, the views of certain groups of students may not be captured. This may, however, follow patterns in real life, as different groups of students can be more or less inclined to complete SETs.

4.6.3.3.2 Thematic analysis: coding process

Study 1

I followed the procedure by Braun and Clarke (2006; *see Chapter 5 for a detailed description of this process*). I discussed the themes and sub-themes for each coded item

with a member of the supervisory team and revised themes based on this discussion. Furthermore, I calculated a degree of agreement between one of my supervisors and me for the coding of two items. Additionally, I identified frequencies of themes and sub-themes that I visualised in corresponding figures. To make this process understandable and transparent, I created coding frames that demonstrate how codes were assigned to participants' quotes and how these codes generated sub-themes and final themes.

Study 2

I was guided by a) the procedure for conducting thematic analysis (Braun & Clarke, 2006), and b) most principles described by Robinson (2022), specifically his process for analysing brief texts (ST-TA). Robinson's approach involves a) a priori theme development for hybrid and deductive thematic analysis, b) deep immersion in the data, c) generating initial codes and naming themes, d) tabulating themes against data segments, e) checking inter-analyst agreement, f) exploring theme frequencies, g) developing thematic maps and diagrams and h) producing a report.

In my study, I followed a process of analysing data that was similar to ST-TA with several alterations. At first, a) I considered my themes from Study 1 as orientating constructs that I used as a starting point. I later modified these constructs based on data from Study 2. Next, b) I deeply immersed myself in data through reading participants' answers and adding detailed notes, which I have done at least twice for each dataset. Afterwards, c) I generated initial codes and named themes, aiming to convey their meanings in clear ways. In relation to step d), I conducted the first part of my analysis in Word instead of Excel, and therefore, used a different system to attach data segments to themes. For step e), I checked the inter-analyst agreement through informal discussions with my supervisors. I used a more structured process consistent with the one described in Robinson (2022) to check the inter-analyst agreement in Study 1, but applied a less structured process in Study 2. Afterwards, I f) explored the theme frequencies as this may provide additional clarity and precision to the analysis (Robinson, 2022). Additionally, these frequencies, here understood as supportive to qualitative data, may also help readers better understand data derived from complex findings. In relation to step g), to visualise themes and present analytical patterns concisely, I created tables instead of diagrams, as this allowed to clearly demonstrate how themes related to the examined answer scale

categories. Last, h) I produced a report. Because brief texts formed a qualitative part of data, I discussed themes in line with the standard qualitative data write-up section.

4.6.3.3.3 Credibility and reflexivity

Credibility serves as an important criterion for ensuring the internal validity of qualitative research and increases with an accurate description of the context and interpretation process, as well as an application of triangulation (i.e., multi-method approach) and reflexivity (Hammarberg et al., 2016). To address credibility, I kept the records of participants' responses and a record of my own assumptions, decisions, and thoughts during the coding process. I applied a multi-method approach by using a predominantly qualitative method to study rating scales as a standard method of quantitative data generation and aimed to provide detailed descriptions of context.

To ensure *reflexivity*, I tried to maintain a continuous awareness that my personal stance may affect the research outcome and apply constant critical self-evaluation during this process (Berger, 2015). I engaged in reflexivity to acknowledge any of my potential subjective opinions or unintentional biases affecting the way in which I construct and use reality and language, as well as my interpretation of data (Berger, 2015; Finlay, 2002). Applying reflexivity helped to reduce the potential effects of my biases, experiences, or opinions on the research process, or, when unavoidable, interpret findings in light of these biases (e.g., Berger, 2015) and enhance the transparency of this process. I include the following *reflexivity statement* to describe my background and its potential impact on this research:

I am a cisgender White female postgraduate E.U. student, working on this research during my late twenties and early thirties. I grew up in Slovakia and moved to London after graduating from secondary school. My cultural background could have impacted this research because the higher education system in Slovakia differs from the one in the U.K. There is a high degree of formality in instructor-student interaction. Students in Slovakia tend to address their instructors formally or using their titles. Although students may informally evaluate their lecturers through the websites for this purpose, official student evaluations of teaching organised by universities are less frequent compared to the U.K. After completing a bachelor's degree in Psychology with Management at Goldsmiths University, I obtained a master's degree from City University in Organisational

Psychology. I wrote this dissertation for my PhD degree in Psychology at the University of Greenwich.

I am familiar with teaching, both at university and in sports settings. I worked as a qualified tennis coach for over ten years. This profession is rather male dominated. Only 23% of tennis coaches are women, and this number declines with increasing levels of qualifications. I obtained Level 3 (out of 5), which is considered relatively high and enables coaches to run individual or group sessions with players of any level without supervision. During my Level 3 coaching course, the instructor remarked on much better than a usual representation of women in our group (29%). As a woman, my occasional experiences with gender bias (usually microaggressions or benevolent sexism) could also have influenced my research. During my PhD studies, I also taught at the academic level, leading seminars for undergraduate Level 4 and Level 5 students. This enabled me to better understand the challenges that lecturers may face when teaching students.

4.7 Ethical considerations

To ensure that participants a) did not suffer any harm and b) maintained their well-being and dignity (e.g., Willig, 2008), I identified and addressed relevant ethical considerations, guided by the BPS Code of Ethics (British Psychological Society, 2014). Because the use of online surveys may result in additional ethical issues, I also focused on four general ethical issues associated with the use of online surveys, specifically: a) informed voluntary consent, b) use of incentives, c) privacy, anonymity and confidentiality, and d) ensuring data quality (Roberts & Allen, 2015). This section discusses general ethical issues pertaining to both studies. Issues specific to Study 1 or Study 2 are discussed in more detail in Chapter 5 (Study 1) and Chapter 6 (Study 2).

To gain *informed voluntary consent*, the survey started with a consent form and only participants who provided consent could proceed. All participants in both studies were informed that their participation was voluntary, and they could withdraw at any time without providing any reason. Participants could also withdraw up to two weeks after completing the study by contacting me with a code they generated during the survey, but nobody did so. Participants studying or planning to enrol to the University of Greenwich could feel uncomfortable about withdrawing from this study. Therefore, students were informed through the participant consent form that their future with the university would not be affected in any way if they decided to withdraw. Furthermore, participants stayed anonymous throughout the survey (*see also below*).

Regarding the *use of incentives*, participants from the University of Greenwich could use the Sona system to participate in Study 1 or Study 2 for credit. Following the recommendation by Roberts and Allen (2015), this was done through participant pools at the departmental level. To protect *students' privacy and anonymity*, I did not collect any I.P. addresses. The anonymised collected information was only seen by either my supervisory team or me. Participants from universities recruited outside of the University of Greenwich did not have to specify their university to protect their anonymity further (this was not always possible at the University of Greenwich if participants used Sona systems). Students could also choose to participate in other surveys advertised in Sona system instead.

In an online setting, respondents choosing not to answer all questions may result in missing data and reduce the overall *quality of data* (Roberts & Allen, 2015). However, because participants have the right not to answer all questions, applying forced questions may lead to ethical issues if participants feel uncomfortable providing answers. It can be unclear whether the participant aimed to skip the question intentionally or by accident. To reduce unintentionally missing answers, I chose to use alerts instead of forced responses for individual survey questions. These alerts notified participants they skipped answering a question, but participants could still proceed and omit questions they were not comfortable answering.[11]

4.8 General summary of this chapter

The overall aim of this PhD research was to investigate the suitability of SETs for their purpose. Specifically, I explored whether and how the methodological limitations of standardised rating scales may contribute to, rather than reduce, the manifestation of gender biases in students' ratings. To study this problem, I applied the conceptual frameworks from the TPS paradigm.

I investigated how raters interpret and use rating scales in the context of SETs, including connections between the student participants' ratings and the information they thereby considered as well as the potential influence of gender stereotypes. I also explored how students use and apply common answer categories of agreement scales and whether

[11] Originally, I planned Study 3, the archive study involving existing module evaluations from the University of Greenwich, which would have allowed the analysis of real official teaching evaluations from students. Numerous efforts including email correspondence and meetings with relevant members of the University were held in order to facilitate this originally planned Study 3, but ultimately the Planning and Statistics team (despite UREC approval) did not feel able to grant access to the student data.

the different directions of scales (starting with agreement vs disagreement, e.g., “strongly agree” versus “strongly disagree”) may affect SET scores and students’ interpretations of answer scale categories. I applied the epistemological stance of the social constructionism framework, assuming that gender, gender roles, and gender beliefs are socially constructed and dependent on context. This also allowed me to acknowledge the active role of a researcher and the important function of language, which people use to construct reality.

To provide unique insights about the study phenomena and enhance the overall quality and understanding of findings, I used a multi-method approach, in which qualitative and quantitative methods are complementary and applied simultaneously. Online surveys with standardised ratings and open-ended questions enabled the collection of both qualitative and quantitative data, as well as exploring the connection between them. I followed the relevant criteria to make my research in this PhD research transparent, enhance its trustworthiness and authenticity, and establish credibility and reflexivity. All efforts were made to maintain the ethical standards and address concerns specific to the online surveys and students’ judgements about real-life lecturers.

In sum, this chapter summarised the general methodology of my PhD research and its two empirical studies. Study 1, discussed in the next chapter, examined how student participants interpreted SET item statements, what specific information participants considered when rating their specific lecturers, how they weighted this information in their ratings of their lecturers and whether this differed by a lecturer’s gender.

Chapter 5: Study 1

In this chapter, I present the methodology and findings from my first empirical study. I explore how students interpret SET item statements and potential variations in these interpretations. The specific research aims of this study are: a) to examine what teaching behaviours and inferred attitudes students generally consider important (RQ1, RQ2), b) to investigate what specific information students consider when rating their specific lecturers, "worst" or "best", respectively, that they ever had during their university studies (RQ3), c) to explore how students interpret SET item statements and whether in different ways (RQ4), d) to examine whether students hold their specific lecturers to gender-related expectations and are influenced by gendered schemas when evaluating their lecturers (RQ6, RQ7), e) to explore how students weight specific information in their ratings of their lecturers and whether this weight on rating differs by their lecturers' gender (RQ7), f) to examine how students view and approach SETs in general (RQ9).

5.1 Introduction

Previous evidence shows that raters interpret item statements in different rather than (as commonly assumed) standardised ways (e.g., Arro, 2013; Lundmann & Villadsen, 2016; Rosenbaum & Valsiner, 2011; Uher, 2018). There is surprisingly limited research about how students interpret SET item statements and what information (e.g., teaching behaviours, inferred attitudes) students consider salient when rating their lecturers (R. Bennett & Kane, 2014; Benz & Blatt, 1996; Robertson, 2004). But potential variations in the ways in which students interpret item SET statements could increase rather than reduce the manifestation of biases, including gender biases. For example, due to the abstract wording and lack of definition of these items, students must rely on their subjective expectations, interpretations and decisions, which can be influenced by students' different beliefs (e.g., sociocultural beliefs; *see Chapter 3*).

Specifically, the meanings are not inherent to signifiers, but constructed by raters. Therefore, how raters interpret item statements will depend on raters' perspectives and beliefs. For example, evidence suggests that individuals differ in how they perceive even the same event (Hastorf & Cantril, 1954). Individuals also tend to interpret, evaluate and generate evidence in the ways that support their prior beliefs or expectations (Nickerson, 1998; Stanovich et al., 2013). This can include stereotypical beliefs or biases, as the

previous empirical research shows (Uher et al., 2013; Uher & Visalberghi, 2016). In fact, individuals may recall stereotype-consistent information better than information inconsistent with stereotypes, and even falsely recall stereotype-consistent events that never occurred (Dodson et al., 2008; Lenton et al., 2001; Sherman et al., 2003).

SET items are commonly ambiguous or vague, to allow raters to interpret them in regards to many different behaviours and consider a wide range of study phenomena. But ambiguous information may reinforce biases and stereotypes as also evidenced in forensic research (Curley et al., 2022; De la Fuente et al., 2003). Furthermore, individuals are even more likely to rely on their prior beliefs and biases during depletion of their cognitive resources (Curley et al., 2022; Petty & Cacioppo, 1986). Because ambiguous statements can make judgements more cognitively taxing, students may then process information in a superficial way, relying on stereotypes.

5.2 Method

5.2.1 Research design

I used a multi-method design, in which different methods are concurrently applied to answer my research questions. This study's general design consisted of online surveys with standardised rating scales on which student participants[12] rated their specific lecturers from real-life contexts (chosen as "worst" or "best" lecturers that participants ever had), and open-ended questions. This between-participants design had two independent variables with two levels, a) the lecturer's gender (man or woman), b) the type of the lecturer ("worst" or "best"). Seven dependent variables were the participants' ratings of their lecturers for each of the seven chosen SET items. This experimental design and the outcomes from open-ended questions enabled me to scrutinise how participants used and interpreted SET item statements and what information about their lecturers' teaching behaviours and inferred attitudes participants considered in their ratings of lecturers.

5.2.2 Participants

The sample consisted of 181 participants (138 women, 40 men, 2 participants who identified differently and one unspecified). All participants were 18 years or above, and either currently enrolled as students at a U.K. university or had finished their studies within the last twelve months. Most participants were from the U.K. (N = 129). The rest

[12] Unless otherwise specified, the term "participants" always refers to the (student) participants in my study in this chapter, whereas "students" to students in general.

were either International (*N* = 28) or E.U. students (*N* = 23). Due to data collected, it cannot be ascertained whether all U.K. and E.U. participants paid tuition fees for their degrees. Specifically, participants may have attended a Scottish university, which would mean Scottish and (in some cases E.U.) participants may have been exempt from tuition fees (*see also Chapter 2*). I did not ask participants whether they attended research-intensive or teaching-intensive type of institutions. Research-intensive universities encompass strong research culture and prioritise research performance over teaching. In contrast, teaching-intensive universities allocate more time and resources to teaching and supporting students in meeting academic standards, rather than cultivating research culture (Tomas & Jessop, 2019). Students from different types of institutions may perceive teaching quality differently and vary in what they expect from their lecturers. However, in order to determine what type of university participants attended, they would likely have to disclose their institution's name. I decided not to collect this information for ethical reasons as it could compromise anonymity of participants or evaluated lecturers. The distribution of participants' academic backgrounds is displayed in Tables 2 and 3.

I obtained all data (participants' answers to open-ended questions and their ratings of SET items) directly from participants, whom I recruited through the University of Greenwich, as well as other universities, and through survey websites and social media. I used opportunity sampling in the first stage of the recruitment, and purposive sampling in the second stage (*see Chapter 4 for the rationale*).[13]

[13] Participants from the University of Greenwich received points for participation through SONA system.

Table 2

Percentage of Participants' Study Academic Backgrounds (N = 180) [14]

	Female students		Male students		Non-binary students		Overall	
Subject	**(%)**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)**	**N**
Art/Design	3.3	6	2.8	5	0	0	6.0	11
Business and Law	17.2	31	7.2	13	0	0	24.0	44
Computer Science	1.1	2	0	0	0	0	1.0	2
Humanities	5.0	8	1.1	2	0	0	6.0	10
Natural Sciences	3.3	6	1.1	2	0	0	5.0	8
Psychology	28.3	51	3.9	7	0	0	32.0	58
Social Sciences	16.7	30	5.6	10	0	0	22.0	40
Subjects Allied to Medicine	2.2	4	0.6	1	0.6	1	3.0	6
Other	0	0	0	0	0.6	1	1.0	1

Note. The table depicts a number (N) and percentage (%) of female, male, non-binary, and overall participants to see the composition for each subject across the whole sample. One participant did not indicate their study subject background and was excluded from this analysis. The Art/Design cluster also involves Music. Business and Law include Marketing. Humanities involve Communications, Languages (including English literature), Historical, and Cultural Studies. Natural sciences are Mathematical, Physical and Biological Sciences. Social Sciences involve Education, Economics, Political Sciences and Social Studies.

[14] I was generally guided by HESA (the Higher Education Statistics Agency) subject codes, but I made a few minor revisions (such as using a label "Social Sciences" instead of "Social Studies). I also combined some related clusters (Mathematical, Physical and Biological Sciences into Natural Sciences, separate clusters Business and Law into one cluster Business and Law, and Communications, Languages including English literature, Historical, and Cultural Studies into the cluster Humanities). I also analysed Psychology separately from Social Sciences because it is commonly considered also Natural Science.

Table 3

Percentages of Academic Backgrounds for Participants Rating their "Best" Lecturers (N = 96) and "Worst" Lecturers (N = 84)

	"Best" lecturers				**"Worst" lecturers**			
	Female		**Male**		**Female**		**Male**	
Subject	**(%)**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)**	**N**
Art/Design	2.1	2	6.3	6	1.2	1	2.4	2
Business and Law	7.2	7	13.5	13	11.9	10	16.7	14
Computer Science	0	0	2.1	2	0	0	0	0
Humanities	3.1	3	4.1	4	2.4	2	1.2	1
Natural Sciences	0	0	2.1	2	6.0	5	1.2	1
Psychology	16.7	16	16.7	16	11.9	10	19.1	16
Social Sciences	5.2	5	15.6	15	13.1	11	10.7	9
Subjects Allied to Medicine	3.1	3	1.0	1	1.2	1	1.2	1
Other	0	0	1.0	1	0	0	0	0
Total number of lecturers	37.5	36	62.5	60	47.6	40	52.4	44

Note. The table depicts the number (N) and percentage of the female and male lecturers who participants in my study chose as their "best" and "worst" for each subject across the overall sample.

As seen in Table 3, participants overall chose similar percentages of "worst" female (47.6%) and male (52.4%) lecturers. However, 62.5% participants reported their "best" ever lecturers were men and only 37.5% women. This distribution differed between academic subjects, for example, participants with Psychology academic background chose the same number of "best" male and female lecturers, but more male than female lecturers as their "worst".

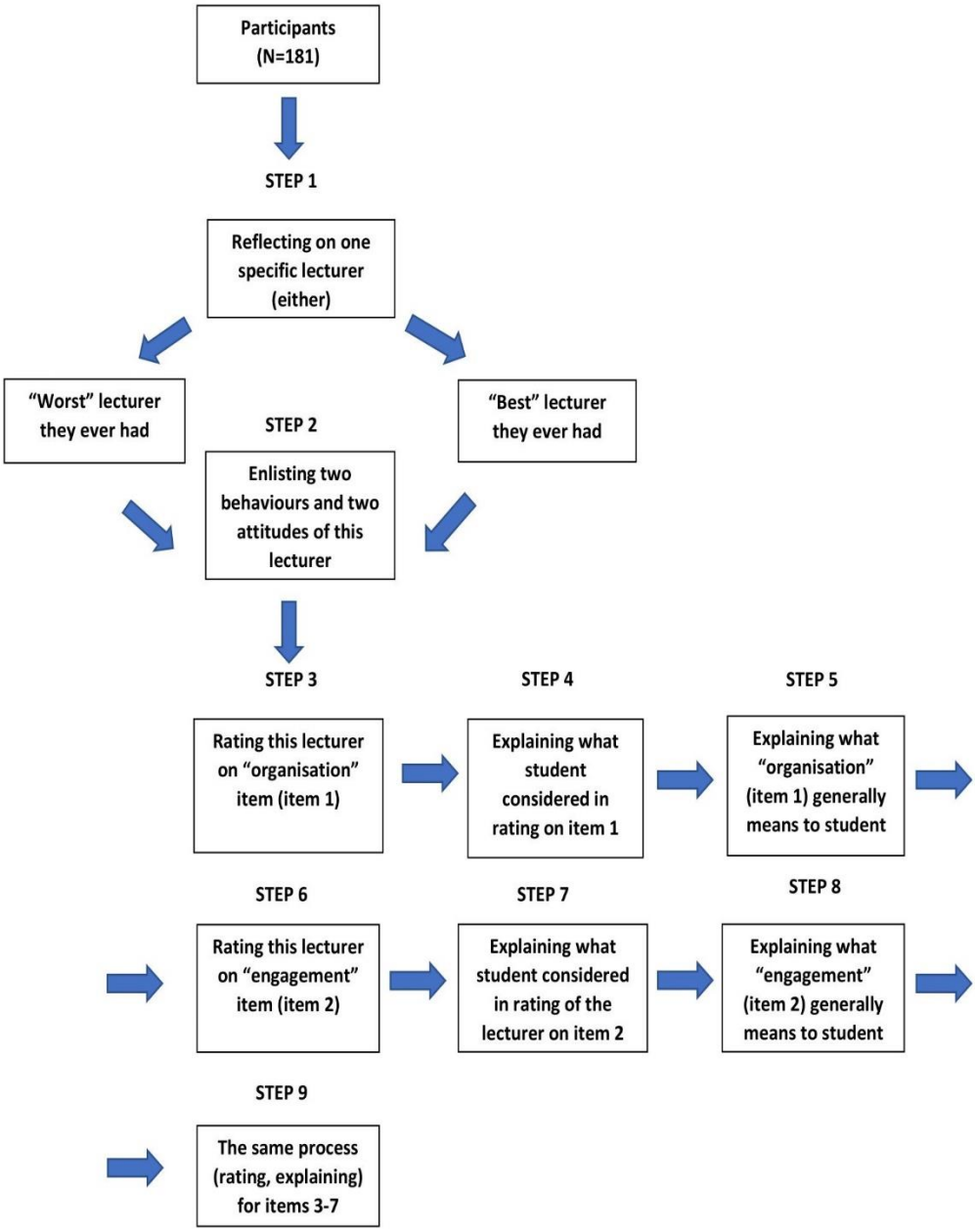

Figure 6

Procedure Followed by Participants (N =181)

Note. This figure depicts the procedure for participants in my study, who reflected on either their "worst" or "best" lecturers. Participants provided each two behaviours and inferred attitudes of their lecturers before providing any rating. These teaching behaviours and attitudes are therefore unrelated to any specific items. Afterwards, participants rated their "best" or "worst" lecturer on each item, explained what information they considered when rating their lecturers and how they interpret the meaning of each item.

5.2.3 Materials

I used the following seven SET items derived from the SET questions used by the University of Greenwich: [15][16]

Item 1: (henceforth called the "organisation" item): Their module was well-organised

Item 2: (the "engagement" item): The lecturer has made the subject interesting

Item 3: (the "effective teaching" item): The way their module was taught has helped me to learn

Item 4: (the "support with the assessments") item: I have received good support to manage my assessment workload

Item 5: (the "feedback" item): Feedback has helped me develop and improve my performance

Item 6: (the "challenge" item): This module has challenged me to do my best work

Item 7: (the "overall satisfaction" item): Overall, I was satisfied with the teaching on this module

5.2.4 Procedure

Initially, participants reflected on either their "best" or "worst" university lecturer whom they encountered during their university studies, describing two teaching behaviours and two inferred attitudes of this particular lecturer, prior to any rating. Participants then rated this specific lecturer on all seven SET item statements, justified each of their ratings, and provided their general interpretations of the seven items (Figure 6). Therefore, I obtained not just participants' ratings of these real lecturers on specific SET items, but also participants' comments on their lecturers' behaviours and teaching attitudes that they inferred from these behaviours, participants' general interpretations of item statements and the information that they considered when rating their lecturers (*see Chapter 4*).

[15] I only included a few selected SET items and not the full scale. The reliability for these seven items as computed by Cronbach's alpha was $\alpha = .86$ for "best" and $\alpha = .85$ for "worst" lecturers, which indicates a good reliability by mainstream standards. However, my predominant aim is to scrutinise students' interpretations of item statements as well as shortcomings of so-called quantitative methods with qualitative methods rather than focus on the differences between lecturers' ratings (*see also Chapter 4*).

[16] I slightly adapted a few items to suit the context (e.g., "Staff made the subject interesting" to "The lecturer made the subject interesting").

Nineteen percent of participants were (N = 34), in addition, were asked to provide an average time they thought they usually spend on completing SETs. These participants also described through open text how they generally approach SETs (e.g., spontaneously versus preparing in advance) and their general opinions on SETs. All participants provided demographic information about themselves and specified the considered lecturer's gender.

5.2.5 Data analysis

5.2.5.1 Key strategies for analysing qualitative data

I analysed the participants' open-ended responses with thematic analysis. I performed coding with Microsoft Word, following the procedure by Braun and Clarke (2006) and using a hybrid, predominantly deductive approach (Braun & Clarke, 2019). Two datasets ("worst" and "best" lecturers) were coded separately. In the first stage, I analysed data by each question, specifically, the lecturers' behaviours and the lecturers' attitudes that participants inferred from these behaviours (thus, participants' reflections on their lecturers *before* providing any ratings), and comments derived from the seven SET items (specific information that participants considered when rating on each item, and their general interpretations of items). Once I had identified the themes and sub-themes for a specific section (e.g., behaviour interpretations provided for "worst" lecturers), I used them to inform the coding process for the equivalent question for the "best" lecturers. The inductive element of my coding refers to my attempt to stay close to the data during the first stage of my coding and set my theoretical prescriptions aside (Braun & Clarke, 2006). I analysed data on a latent level, looking beyond what participants said and considering the underlying meaning of their words.

In order to ensure the transparency and trustworthiness of my analytical scheme and coding process, I also checked the inter-analyst[17] agreement (Robinson, 2022). I sent two small samples of data (responses from 2x 20 participants for two separate items) to a member of my supervisory team. I also shared relevant coding frames and my coding strategy with hypothetical examples. My supervisor performed coding prior to any discussion, and we then reviewed discrepancies in our categorising of the themes and sub-themes for each item and revised them on the basis of this discussion. I calculated a

[17] Although this process is frequently called "inter-rater" reliability in the literature, I agree with Robinson (2022) that this term may be confusing and conflate qualitative analysis with processes applied in psychometrics (*see also Chapter 3*). I therefore use his term "inter-analyst" instead.

degree of agreement using a "percentage agreement formula" (Miles & Huberman, 1994) between this supervisor and me for the "organisation" item for "worst" lecturers (82% after our final discussion) and the "overall satisfaction" item for "best" lecturers (93.8% after our final discussion). To make the coding process even more understandable and transparent, I created 32 coding frames that demonstrate how I assigned codes to participants' statements and generated the sub-themes and final themes *(see Table A6, Appendix, for an example of one of these coding frames*).

Furthermore, I identified frequencies of the themes and sub-themes that I visualised in 32 corresponding figures. I counted every participant's comment only once for each sub-theme, even if they mentioned contents related to it several times. However, a participant's comment could appear in several sub-themes of the same theme (e.g., if I categorised a participant's statement into both sub-themes "content was engaging" and "content was informative", that were a part of the main theme "content-related comments"). In each coding frame, I specified these occurrences (e.g., ten participants mentioned content categorised into *one sub-theme*, seven reported content coded into *two sub-themes* and four participants mentioned content categorised into *three sub-themes* of this main theme).

I analysed participants' comments about their opinions and approaches to SETs with inductive thematic analysis. I also calculated a mean of the time that participants reported they usually spend on completing SETs.

5.2.5.2 Key strategies for analysing the rating data

I tested for potential overall differences in lecturers' ratings on the seven SET items depending on lecturers' gender. One independent variable with two levels was the chosen lecturers' gender[18]. The seven dependent variables were the ratings that participants gave to their chosen lecturers on each of the seven SET items. To see whether lecturers' gender had an effect on their SET scores, I analysed the rating scores with One-Way MANOVAs (separate for "worst" and "best" lecturers). I calculated effect sizes represented by partial etas squared. For this analysis, I interpreted effect size using

[18] I considered examining the effect of a students' gender (and its interaction with a lecturers' gender) on these scores. However, due to the unequal distribution of the sample (only $N = 40$ of participants were men), I decided to only explore the effect of a lecturers' gender.

Cohen's guidelines (1988), that established conventional cut-off points, as ηp^2 = .01 as small, ηp^2 = .06 as medium, and ηp^2 = .14 as large[19].

After I identified the sub-themes and themes, I applied a binary scheme, indicating "1" if a participant's statement was coded into a particular sub-theme or theme and "0" if it was not. I then conducted chi-square analyses to compare what information participants considered more frequently when reflecting on their lecturers' teaching behaviours and attitudes (before providing any ratings). I also explored similar teaching behaviours across items, also with chi-square analyses, to study whether participants considered information related to certain themes with a similar frequency for both male and female lecturers while rating them.

Afterwards, I ran bivariate correlation analyses to explore how participants weighted the information considered (represented by the number of sub-themes) in their ratings of their lecturers and whether this differed by their lecturers' gender. I used the guidelines by Gignac and Szodorai (2016) and interpreted correlations of r= .10 as small, r= .20 as medium, and r= .30 as large[20]. I considered only the most frequent theme per each of the seven items (conducted separately for "worst" and "best" lecturers).

To correct for multiple tests but retain the exploratory nature of the analysis (*see Chapter 4*), I applied the Benjamini-Hochberg method with a 10% False Discovery Rate (Benjamini & Hochberg, 1995). Participants occasionally provided negative comments for their "best" or positive comments for their "worst" lecturers. I aimed to investigate how the frequency of positive or negative comments can materialise in rating scores (e.g., whether numerous positive comments are associated with a higher rating score). Therefore, in this analysis I considered only positive comments for "best" and negative comments for "worst" lecturers.

Finally, I analysed participants' reported opinions and approaches to SETs with inductive thematic analysis.

---

[19] However, these arbitrary points must be interpreted with caution and heavily depend on context (Cohen, 1988).

[20] These guidelines differ from the frequently used guidelines by Cohen (1988), who interprets r=.10 as small, r=.30 as medium, and r=.50 as large. The evaluation of Cohen's guidelines from empirical perspective revealed that they may be too restraining, because less than 3% of correlations showed correlation equal or above .50 (Gignac & SzodoraiI, 2016).

5.2.6 Special ethical considerations for this study [21]

This study was approved by Research Ethics Committee at the University of Greenwich and I was guided by the BPS Code of Ethics (British Psychological Society, 2014). The main ethical concern involved the anonymous collection of potentially critical or condemnatory information about real lecturers working at U.K. universities. To protect the lecturers' and the participants' identities, I instructed participants in the advertisement not to disclose any information that could reveal the identity of their lecturers or universities or themselves, to avoid mentioning any names and provide only relevant demographic information. In the information sheet, I advised participants that any identifiable information about the lecturers or themselves would be deleted immediately.

I offered all participants a chance to enter a prize draw for a £25 Amazon voucher. Participants entered their email addresses in a separate survey to further protect the anonymity of their answers, and I awarded six randomly chosen participants Amazon vouchers. This is in line with the recommendation that financial incentives for participation may be used but should not be excessive (Koocher et al., 2015). Due to the multiple open-ended questions, my study involved more than a typical amount of writing and the use of incentives, therefore, seemed appropriate.

5.3 Findings and discussion

5.3.1 RQ1: Which of their lecturers' teaching behaviours do students generally consider important?

5.3.1.1 Wide range of lecturers' teaching behaviours reported by participants

Participants' reflections on their lecturers' behaviours varied considerably. Specifically, when reflecting on the behaviours that they generally find important (before providing any rating for a specific lecturer and therefore not related to any item), participants as a group considered information that I coded for "best" lecturers into two themes and for "worst" lecturer into three. Each theme had between zero and five sub-themes (M = 3.60, SD = 2.07). In total, 97 participants used 1693 words to describe their "best" lecturers' behaviours and 84 participants used 2029 words to describe the behaviours of their "worst" lecturers. This may suggest that, on average, participants focused more on negative than positive aspects of teaching behaviours, even though more participants considered their "best" rather than "worst" lecturers.

[21] I described general ethical considerations that apply to both studies in Chapter 4.

I categorised teaching behaviours reported by participants into three main themes: 'Teaching related behaviours' (mentioned by 79.4% of participants commenting on their "best" and 60.7% on "worst" lecturers), 'Interaction with students' (44.3% for "best" and 45.2% for "worst") and 'Assignment-related concerns' (16.7% for "worst", but, surprisingly, 0% for "best"; *see Table 4*). The theme 'Teaching-related behaviours' included content related to teaching behaviours occurring during lectures (e.g., clear communication during teaching). 'Interaction with students' covered the ideas about the lecturers' approach and relationship to students (e.g., support and care or lack thereof), and the theme 'Assignments-related concerns' included any comments about problems with assignments (e.g., unclear guidelines). This wide range of reported behaviours implies that students may emphasise certain teaching behaviours (e.g., interaction with students) and disregard others (e.g., lecturers asking challenging questions, detailed feedback provided for assignments). It also indicates individual differences in what students expect of their lecturers.

Table 4

Comparison of Teaching Behaviours Reported for Participants' "Best" and "Worst" Lecturers (N= 181)

Themes and subthemes	**"Best" lecturers**	**Themes and** subthemes	**"Worst" lecturers**
% of participants mentioning content related to the themes/sub-themes			
Teaching-related behaviours	**79.4**	**Teaching-related behaviours**	**60.7**
Good engagement	50.5	Lack of engagement	26.19
Displayed passion/enthusiasm	36	Lack of passion/enthusiasm	9.5
Clear communication	22	Unclear communication	23.8
Good organisation	10.3	Lack of organisation	14.3
Displayed knowledge	9.3	Lack of knowledge	9.5
Interaction with students	**44.3**	**Interaction with students**	**45.2**
Provided support and care	25.8	Failed to show support and care	7.1

Friendly-approachable towards students	21.7	Mean-patronising-rude-unapproachable towards students	28.6
Clear-provided answers to students' questions	4	Insufficient answers to students' questions	14.3
Respectful communication with students	3	Disrespectful communication with students	4.8
Assignments	–	**Assignments**	**16.7**

Note. The table illustrates different behaviours that participants emphasised and mentioned and how this differed for lecturers they considered their "best" versus "worst". The percentages indicate how many participants mentioned content related to the particular theme or sub-theme for their "best" or "worst" lecturers. The main themes are in bold font, whereas sub-themes are indented. Participants' comments could be categorised into several sub-themes, the percentages may therefore not always add to 100%. "–" indicates no participants mentioned any content related to the particular theme.

5.3.1.2 Analysis of the main themes

5.3.1.2.1 Theme A: Teaching-related behaviours

Most participants provided statements about teaching behaviours occurring during the delivery of lectures, mostly related to engagement, clarity of communication, and lecturers' perceived levels of passion (*see Table 4*). The most frequent sub-theme involved comments about the 'Level of engagement'. This referred to participants expressing the need to be actively involved during the lecture, but also seeking to be entertained. For example, participants commented *"He engaged everyone in class (it was a small group) by asking for everyone's views, asking to read, and organised students to present interesting topics to the rest of the class - the exchange was now, not only between students and teacher but also between students"* (B15[22]) or *"They invited discussion rather than relying on delivering a prepared lecture from a powerpoint presentation"*

[22] Indicates a label for participant, labels starting with "B" means that participants reflected on their "best" lecturer, whereas those with "W" on their "worst" lecturer

[sic][23] (B63) as examples of lecturers making students feel involved. Other participants wrote, *"They made jokes throughout the lecture rather than just constantly feeding you information which helped to avoid losing interest over time."* (B53) or *"Involved the lecture theatre without making it too intense Included humour so that content was memorable"* (B8) as examples of lecturers providing entertainment. In contrast, participants reflecting on their "worst" lecturers frequently commented on extensive reading from the slides, e.g., *"They would specifically only be reading off the screen and would not add too much input into the topic we are discussing and therefore it was quite drab"* (W26), as an example of a low level of engagement. This suggests that many students may wish to actively participate in their learning, but also hints at a potential belief that lecturers should entertain students to keep them engaged throughout their lectures. For example, 40.2% of student participants in another study described their ideal lecturer as funny or entertaining (Strage, 2008). Similarly, academics perceive that students demand entertaining lectures, but this could be problematic if lecturers prioritise roles of entertainers versus educators, for example, to comply with students' demands in order to achieve higher SET scores (Wong & Chiu, 2019). These student expectations could be explained by increasing tuition fees and students applying a more consumerist approach. Treating students as consumers could also affect SET scores. Students may perceive the pursuit of their degree as a financial transaction, and shift responsibility for their learning to their lecturers (T. Bennett, 2021; Bunce et al., 2017). Specifically, students may focus on obtaining a degree rather than acquiring expertise in their academic subject. This can lead to SET scores being perceived as 'customer satisfaction' scores, which could affect teaching practices. For instance, academics reported having to make content more entertaining at the expense of academic rigour in order to appease students (T. Bennett, 2021). Similarly, to fulfil students' expectations and obtain high SET scores, lecturers may hesitate to provide critical feedback and appropriately challenge students; or feel obligated to simplify their lecture content. This may negatively impact both student learning and teaching quality (T. Bennett, 2021; Bunce et al., 2017; Emery et al., 2003).

[23] To maintain the authenticity of participants' responses, I did not correct any grammar or spelling mistakes.

5.3.1.2.2 Theme B: Interaction with students

The theme 'Interaction with students' predominantly involved participants' reports about behaviours related to their lecturers' approachability and support. Interaction is seen on a more personal level here; it is not merely engagement with class or interactive sessions, but instead portrays a relationship between participants and their lecturers. The sub-theme with the highest frequency of comments (across the overall sample evaluating both "best" and "worst" lecturers) covered content related to behaviours towards students, that I categorised as 'Mean-patronising-rude-unapproachable' (for the "worst" lecturers) or correspondingly 'Friendly-approachable' (for the "best" lecturers).

Participants disliked if lecturers acted in ways they perceived as arrogant or patronising, for instance, *"Constantly condescending students in the lecture."* (W12), *"Answered questions in a sarcastic manner"* (W9), *"Trivializing my concerns when I bring something up that I'm nervous about or struggling with"* (W13). Similarly, this sub-theme included comments about lecturers acting in rude or mean manner, such as *"Was rude to students which made them feel uncomfortable"* (W34), *"acting super defensive/hostile towards any question or comment"* (W5). One participant mentioned rude and sexist behaviours from the lecturer targeted specifically towards women, "*Sexist remarks eg using the word bitch, saying all women love shoes"* [sic] (W55).

In contrast, participants appreciated lecturers acting in a friendly, modest or approachable manner, e.g., *"They also tried to relate to the audience and make themselves seem less far away from the students in terms of status and experience"* (B53), *"did not act as if he/she was better than those who did not understand"* (B13) or *"Conversational, asked how everyone was…"* (B89). Participants also commended their lecturers for providing academic or interpersonal support, e.g., *"Provided academic support when I didn't know how to do it and emotional support when I had no strength"* (B77) or *"General interest, when talking independently wanted to know how you were getting on with work and if there was anything they could do to help"* (B89). However, some participants reported behaviours that could be seen as unfeasible, *"Providing personal contact number to text or call when there were any issues"* (B16), or even problematic in some contexts, e.g., *"Socialising"* (B95). In sum, participants seemed to dislike feeling belittled or disrespected and appreciated if their lecturers were understanding, friendly or supportive. This is in line with previous findings in which students described their ideal lecturers as approachable and supportive (Su & Wood, 2012). However, it may present a

potential problem for lecturers who may want to act in an approachable manner but also need to maintain their professional boundaries.

5.3.1.2.3 Theme C: Assignment-related concerns

The theme 'Assignment-related concerns' included participants' reports about their lecturers' behaviours related to poor quality of feedback, unclear communication about assignments or lack of support specifically with assignments. These involved statements such as *"The lecturer isn't very professional: she comes across as frustrated and fed up when grading papers and sometimes writes comments like "eh?" which aren't helpful"* (W27) or *"Not being supporting students with assignments"* (W60), or *"Criticism that wasn't constructive"* (W3). This implies that many students take assignments seriously and any problems with assignments can negatively affect their perception of their lecturers' behaviours. It also indicates that some students focus on assessments more than their overall learning experience, which is in line with findings by Arthur (2020), where students commented on negative aspects affecting their performance such as too many assessments or even external factors.

5.3.1.3 Differences in teaching behaviours considered for "best" vs. "worst" lecturers

Occasionally, only the "best" (but not "worst") lecturer descriptions involved certain teaching behaviours and vice versa. Similarly, participants sometimes mentioned for "best" lecturers specific behaviours more frequently than for "worst" lecturers.

5.3.1.3.1 Assignment-related concerns

Participants mentioned 'Assignment-related concerns' only for "worst" but not their "best" lecturers (*see Table 4*). This implies that students may notice any potential problems with assignments (e.g., in terms of lack of support or poor quality of feedback), but not particularly appreciate or find it worth mentioning if assignments run smoothly.

5.3.1.3.2 Engaging and supportive behaviour

Participants commented on engaging lecturers' behaviours and support provided by their "best" lecturers (50.5% of participants mentioned content related to engagement, and 25.8% to support) much more often compared to a lack thereof for the "worst" lecturers (26.2% commented on lack of engagement, 7.1% on lack of support). This suggests that students may especially value engaging behaviour during lectures or support

provided by their lecturers and observing these behaviours might lead to a more positive overall perception of their lecturers (*see the previous section for direct quotes*).

5.3.1.3.3 Lecturers' passionate and enthusiastic behaviours (or lack thereof)

Participants appreciated when their "best" lecturers displayed 'Passionate' or 'Enthusiastic' behaviours (related content mentioned by 36.1% of participants reflecting on their "best" lecturers). However, participants mentioned a lack of passion or enthusiasm much less frequently when thinking about their "worst" lecturer (only 9.5% of participants mentioned related behaviours for their "worst" lecturers). This may indicate that students highly appreciate if lecturers show passionate or enthusiastic behaviours, but do not particularly mind if these are lacking.

5.3.1.3.4 Unexpected findings

Student participants frequently interpreted from their lecturer's body language, behaviours as passionate or energetic behaviours, such as "moving around a theater a lot and presenting energetically" [sic] (B9), or "walked about the lecture room and were expressive with their body language" (B58). A considerable number of participants (36%) mentioned content related to 'passion-enthusiasm' as a behaviour of their "best" lecturers. That can be problematic, because students may perceive lecturers who cannot walk or present energetically (e.g., because of a physical disability) as displaying less passion or enthusiasm. This could lead to potential discrimination of these lecturers, because they could be disadvantaged during SETs.

5.3.2 RQ2: Which attitudes that students ascribe to their lecturers do students generally consider important?

Attitudes can be defined in different ways, and there is currently no clear consensus regarding a definition of attitude (Fiske & Taylor, 2017; Gaiseanu, 2020). For example, some researchers define attitudes as evaluative judgements (e.g., Crano & Prislin, 2006). Other definitions describe an attitude as a tripartite model, comprising affect (feelings), cognition (thoughts), and behaviours (e.g., Breckler, 1984). In my study, I approached attitudes similarly to Perloff (2010) who conceptualised attitudes as hypothetical constructs that can never be observed but only inferred from people's actions. To maintain this distinction between teaching behaviours and inferred attitudes, I, therefore, instructed participants to describe two attitudes of their lecturers that they thought were *underlying* their lecturers' behaviours.

5.3.2.1 Wide range of lecturers' attitudes reported by participants

Student participants mentioned attitudes (before providing any ratings) that I categorised into four main themes and twelve sub-themes across the main themes (for both "worst" and "best" lecturers). This indicates that participants overall considered a wide range of inferred attitudes. In total, 97 participants used 1076 words to describe attitudes of their "best" lecturers and 84 participants used 1163 words to describe "worst" lecturers' attitudes. Again, even though the number of participants reflecting on their "worst" lecturers was lower than those reflecting on their "best", they used more words to describe negative aspects of their lecturers' attitudes.

I identified the following themes: 'Interested-enthusiastic' attitude (88.7%) or 'Uninterested-unenthusiastic' (57.1%), 'Positive-relaxed-confident' attitude (27.8%) and corresponding 'Negative-frustrated-insecure' (11.9%), 'Student-centred' (9.3%) or 'Self-centred' (22.6%), and 'Potentially desirable' (11.9%) or 'Undesirable' (5.2%) attitudes (*see Table 5*). Overall, participants mentioned their "best" lecturers seemed to feel passionate about teaching or conveyed a positive, confident, modest or approachable attitude. In contrast, their "worst" lecturers were described as portraying a lack of enthusiasm, acting arrogantly, or being frustrated with students and teaching. Some participants also seemed to dislike their "worst" lecturers challenging them or having high expectations. Interestingly, participants mentioned content related to positive attitudes (e.g., willing to help) for lecturers chosen as "worst". Vice versa, participants mentioned content related to negative attitudes (e.g., strict) for their "best" lecturers.

Table 5

Participants' Reflections on their Lecturers' Inferred Attitudes Before Providing any Ratings

Themes and sub-themes	**"Best" lecturers**	**Themes and** sub-themes	**"Worst" lecturers**
% of participants mentioning content related to the themes or sub-themes			
Interested-enthusiastic	**88.7**	**Uninterested-unenthusiastic**	**57.1**
Caring-interested in students	49.5	Uncaring-uninterested in students	14.3
Enthusiastic in/about teaching	48.5	Unenthusiastic in/about teaching	40.5
Engaging-funny-interactive	16.5	N/A	0 (not mentioned)
Organised-structured	11.3	Disorganised-unstructured	8.3
N/A	-	Lazy	6
Positive-relaxed-confident	**27.8**	**Negative-frustrated-insecure**	**11.9**
Positive in general	15.5	Negative in general	3.6
Relaxed-accommodating	5.2	Frustrated-annoyed	4.8
Confident	4.1	Defensive	4.8
Happy	4.1	Anxious	7.1
Student-centred	**9.3**	**Self-centred**	**22.6**
Open-minded	5.2	Close-minded	9.5
Modest in general	4.1	Superior-arrogant	13.1
Potentially undesirable attitudes	**5.2**	**Potentially desirable attitudes**	**11.9**
Frustrated-nervous-tired	3	Challenging-demanding expectations	7.1
Strict	2	Willing to help	6

Note. The percentages depict how many participants mentioned attitudes related to the particular themes or sub-themes. The main themes are in bold, and the sub-themes are indented. N/A indicates that no content was mentioned for the corresponding theme (e.g., 'hard-working' as corresponding attitude to 'lazy').

5.3.2.1.1 'Interest-enthusiasm' (or lack of) as an important attitude

The overwhelming majority of participants provided content related to 'Interested' or 'Enthusiastic' attitudes (or lack thereof) that they ascribed to their "worst" and "best" lecturers. Specifically, participants appreciated if lecturers seemed to care about their learning, but also valued lecturers' general enthusiasm about teaching. Participants wrote, *"They had an interest in their subject matter which was apparent in the way they talked about it"* (B69) and [The lecturer] *"seemed to genuinely enjoy being there"* (B89). They also commented on lack of enthusiasm, e.g., "*it seemed like they were not interested in teaching, and saw it as a hassle*" (W22) or "*Couldn't be bothered Maybe didn't want to teach particular class or topic*" (W49). One participant also discussed specific lecturers' behaviours from which they inferred their lecturer's disinterested attitude, *"she went around with playing her cell phone"* (W35). Importantly, lecturers may need to use their phones for activities related to the lecture (e.g., mentimeter) but this may be perceived differently by students. Lecturers' attitudes related to engagement, organisation or structure (or lack of) were also categorised in this theme, with the obvious assumption that an engaging, organised or structured lecturer implies their interest or care in a subject or students' learning. For example, participants mentioned "*Willingness to share – was always willing to share new papers that they read, interesting news and their own experiences working on the field"* (B15) or "*They [cared about the teaching quality] and invested their time into lecture preparation"* (B10), or *"constantly keeping up to date. The importance of organisation and timekeeping"* (B22) as positive examples, or *"the lecturer may have poor time keeping skills causing them to rush to start lecturers."* (W32) as a negative example. The idea underlying all the listed attitudes is a lecturer's level of passion.

5.3.2.1.2 'Self-centred' versus 'Student-centred' attitude

Student participants frequently mentioned attitudes related to lecturers acting in the ways participants perceived as 'Self-centred', specifically, 'Superior, arrogant' or 'Close-minded'. For example, participants commented, *"teaching was his way of having a*

stage to tell students how great he is" (W19) or *"I think they felt that they were better than the students"* (W34), as examples of an 'Arrogant or condescending' attitude. Similarly, participants wrote, *"The lecturer seemed to disregard student opinions. The lecturer seemed to believe that his viewpoint was the only correct one."* (W6) or *[All about activities - they clearly believed that the students should spend the whole class interacting with each other], regardless of if this was what the students were interested in doing. They also believed they were right, and that students who did not agree with their approach were inherently wrong"* (W64), as examples of close-minded attitudes. This suggests that students may particularly dislike if lecturers act in ways students perceive as superior or dismiss students' views. In contrast, participants commended their "best" lecturers for the corresponding attitudes, such as being 'Student-centred', for example, 'Open-minded', e.g*., "valued the students opinions and perspectives"* [sic] (B74) or 'Modest', e.g., *"Want to learn from their students"* (B81) or *"They didn't feel the content was beneath them or boring"* (B69).

5.3.2.2 Differences in teaching attitudes considered for "best" vs. "worst" lecturers

Participants reflecting on their "best" and "worst" lecturers generally reported attitudes that could be categorised into the corresponding themes, e.g., 'Self-centred' attitude (for "worst" lecturers) as the correspondent of 'Student-centred'. For example, a participant wrote "*think they are more superior because they are in a higher position as a lecturer, therefore university students are below them*" (W59) as an example of self-centred attitude. In contrast, "*they were receptive to learning from students*" (B12) described student-centred attitude (*see also the previous sub-section for more examples*). However, the frequencies in which participants reported certain inferred attitudes differed between "best" and "worst" lecturers. For instance, over 80% of participants described their "best" lecturers with content I categorised in the themes 'Interested-enthusiastic' (88.7%; *see description above*) or 'Positive-relaxed-confident' (27.8%). The theme 'Positive-relaxed confident' attitudes captured content about "best" lecturers displaying a positive, happy, or relaxed approach. For example, one participant wrote, "*Also, they have a very relaxed approach to teaching which helps keep students calm*" (B53). Interestingly, many participants simply wrote the lecturer seemed positive or happy without any further elaboration on what they meant or how they reached this conclusion, e.g., "*The lecturer was positive towards students. Lecturer was also happy*." (B13). Overall, participants valued interested, enthusiastic, or positive lecturers. But the corresponding attitudes were

reported much less frequently in the "worst" lecturer descriptions, for example, content related to the themes 'Uninterested-unenthusiastic' (57.1%, *see analysis above*) and 'Negative-frustrated-insecure' attitudes (11.9%; *see Table 5*). The theme 'Negative-frustrated-insecure' included content about attitudes that implied some anxiety, frustration, negativity or insecurity of the lecturers described. For example, a participant commented, "*Lack of confidence in her teaching ability and subject knowledge*" (W7). Another participant mentioned, "*She might be fed up with teaching. Perhaps she thought being a lecturer would be different or perhaps she used to enjoy it but no longer does*", as an example of frustration, whereas "*had bad experiences with negative feedback so feels defensive*" (W5) hinted on a perceived insecurity of the lecturer.

The "worst" lecturers were more often described as lacking enthusiasm *about teaching* (mentioned by 40.5% of participants) rather than the lecturer being uncaring or uninterested in students (mentioned by 14.3%), whereas participants reflecting on their "best" lecturers mentioned both lecturers' enthusiasm about teaching (45.4%) and caring or interested attitude (49.5%) relatively often (*see Table 5*). This implies that students may value lecturers' caring or supportive attitude, but not especially mind the lack of this support. However, students may dislike it if lecturers come across as unenthusiastic about teaching. The difference between lecturers not caring about teaching in general versus not caring about students can be difficult to pinpoint. In the latter case, participants often discussed a lecturer's perceived lack of empathy, e.g., "…*also were uncaring of how the students may have felt.*" (W34) or "*did not care whether we were there or not*" (W10). Participants who commented on their "best" lecturers often referred specifically to lecturers caring about sharing knowledge or student learning, for example, "*Desire to make sure people did well in their subject*" (B8), "*Had an interest in making sure students understood the subject*." (B22) or "*Want to help students become their best*" (B1) or "*Care: they cared about making content accessible, making sure students were keeping up, helping students succeed*" (B96). Some comments were going beyond student learning and discussed caring on a more personal level, such as *"He was incredibly motivated to [research] and support students in this field He was empathetic and conscientious to student needs, having been in this situation" (B75), "Helping others and Feeling responsible for the well-being of their student" (B77).* These comments imply students want to feel as if their success matters to their lecturers and also feel understood.

Participants also reported their "best" lecturers were 'Engaging, funny and interactive' (16.5%) but interestingly no participant mentioned attitudes related to being

"boring" for their "worst" lecturers. Instead, "worst" lecturer comments frequently covered content related to a 'Self-centred' attitude (22.6%), such as being superior, arrogant, or close-minded, but comments about 'Student-centred' attitude appeared much less frequently (9.3%) for "best" lecturers.

My findings are in line with previous research that also found that students value if their lecturers are caring (Basow, 2000), enthusiastic or passionate (Basow et al., 2006; Tam et al., 2009), engaging (Reupert et al., 2009), or funny (Su & Wood, 2012) and approachable (Basow et al., 2006). Correspondingly, participants in my study reported all these attitudes (or lack thereof). In sum, students may especially appreciate if their lecturers seem passionate, enthusiastic, supportive, or positive, and dislike their lecturers' perceived lack of enthusiasm or a self-centred attitude.

5.3.2.3 Discrepancies between considered attitudes and behaviours

Reported attitude inferences mostly referred to the behaviours that my participants previously reported of their lecturers, although frequencies of attitude inferences differed from those of behaviours. For example, participants attributed to their lecturers certain attitudes that they mentioned much less frequently when describing their lecturers' teaching behaviours in the previous question. For example, only over 9% of participants reported a lack of passion when describing the *behaviours* of the lecturers, but the lecturer being 'Unenthusiastic in teaching' was the sub-theme with the highest frequency of comments in reported attitudes (related content mentioned by 40.5% of participants; *see Table 6*).

Similarly, participants appreciated if their "best" lecturers displayed attitudes related to being positive, happy or relaxed (26.8%) but did not comment on positivity at all when reporting their lecturers' *behaviours*. Furthermore, only 25.8% of participants (for "best" lecturers) and 7.8% (for "worst" lecturers) commented on their lecturers' supportive behaviours (or lack thereof). However, 49.5% of participants reflecting on their "best" lecturers and 14.3% of participants reflecting on their "worst" lecturers mentioned their lecturer's 'Caring' attitude or 'Interest in students' (or lack of, respectively). Participants perhaps inferred a lecturer's attitude as being 'Unenthusiastic', 'Positive-happy-relaxed' or 'Caring' from a range of different teaching behaviours. This suggests that students infer these attitudes (positive, supportive, unenthusiastic, caring) even from unrelated teaching behaviours, or potentially subtle behaviours that participants may have not reported, such as lecturers' specific tones of voice, gestures, or facial expressions

(e.g., lecturers smiling or frowning). However, observing one's behaviour may not correctly indicate one's attitude because numerous factors apart from attitudes can influence behaviours, and observed relationship between attitudes and behaviours is frequently weak (e.g., Schwarz, 2008). Students may also rely on their pre-existing conceptual schema rather than reporting what actually occurred when inferring their lecturers' attitudes (Shweder, 1977). For example, participants in different studies estimated correlations between items describing conceptually similar behaviours to be high, despite observing contrary events that did not show any correlations between actual observed behaviours (Shweder, 1975; Shweder, 1977). In sum, students may not be equipped to appropriately infer lecturers' attitudes from their teaching behaviours and correctly report them, which could explain these discrepancies.

5.3.2.4 Unexpected findings

There does not seem to be empirical research in which students described the general positive attitude of their lecturers as important, but it was frequently reported in my study. However, participants in a recent study remarked that female lecturers should be more "relaxed" or "smile more" (Adams et al., 2022), which could be seen as a part of a positive attitude. This suggests that students may value positivity in the behaviours and attitudes of their lecturers, which is consistent with my findings. A comment about female lecturers being required to smile more could be seen as sexist, because it implies that female lecturers must engage in additional emotional labour by regulating their own emotions (Adams et al., 2022). However, in my study, participants praised the positive attitude or happiness of both female and male lecturers.

Table 6

Comparison of Frequencies of Themes for Reported Teaching Behaviours and Attitudes

	"Best" lecturers	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**
% of participants mentioning content of this theme					
Behaviours			**Corresponding attitudes**		
Engagement	50.5	26.19	Engaging-funny-interactive	16.5	—
Passion/ enthusiasm	36	9.5	Enthusiasm about teaching (or lack of)	48.5	40.5
Communication (during lecture)	22	23.8	N/A	—	—
Organisation	10.3	14.3	Organised-structured (or unorganised)	11.3	8.3
Knowledge	9.3	9.5	N/A	—	—
N/A		—	Lazy	—	6
Support and care	25.8	7.1	Caring-interested in students (or uncaring)	49.5	14.3
Towards students (friendly-approachable or mean-patronising)	21.7	28.6	Student-centred (or self-centred attitude)	9.3	22.6
Answers to students' questions	4	14.3	N/A	—	—
Communication (with students)	3	4.8	N/A	—	—
N/A	—	—	Positive-relaxed confident (or negative-frustrated-insecure)	27.8	11.9
N/A	—	—	Potentially desirable (or undesirable)	5.2	11.9

Note. The table portrays how many participants (N = 181) mentioned teaching behaviours and attitudes related to the particular themes and sub-themes. Content related to teaching behaviours may be compared to the corresponding attitudes. For example, participants commented frequently on engaging behaviours, but much less on corresponding attitudes ('Engaging-funny-interactive').

5.3.2.4.1 Counter-intuitive attitudes

Some participants discussed attitudes that seemed counter-intuitive in regards to the "best/worst" categories. For example, although participants mostly described their "worst" lecturers' negative attitudes, some comments discussed attitudes that could be viewed as positive, such as 'Willing to help'. Similarly, participants reported negative attitudes of their "best" lecturers, such as being 'Frustrated-nervous-tired' (*see Table 5*). This could be explained in two ways. Participants may have misunderstood the question, even though the instructions were worded in the same way as those about their lecturers' behaviours, and they did not report any counter-intuitive behaviours (e.g., positive behaviours for their "worst" lecturers). Alternatively, participants perhaps did not view these attitudes in an entirely negative light. For example, lecturers being 'tired' could show dedication and 'frustrated' attitudes may imply that lecturers care about their students. Similarly, participants may have perceived their lecturers' 'Challenging' attitude negatively. Alternatively, participants might have acknowledged that even their "worst" lecturers may have some desirable (and similarly, the "best" lecturers undesirable) teaching attitudes.

5.3.3 RQ3: Which teaching behaviours and inferred teaching attitudes do students consider when rating their specific "worst" or "best" lecturers on SET items?

In a next step, I explored the reasons that participants provided for rating their actual lecturers on the seven SET items through open-ended answers about what concrete information participants considered (*see also Tables A1, A2 in Appendix for the overview of this information*). Many of these statements described their lecturers' teaching behaviours and inferred attitudes.

Table 7

Means and Standard Deviations for Participants' Ratings of their "Worst" and "Best" Lecturers

	"Worst" lecturers (N= 84)	**"Best" lecturers (N= 97)**
ITEM	**M (SD)**	**M (SD)**
Q1: Their module was well-organised.	2.69 (1.18)	4.44 (0.59)
Q2: The lecture has made the subject interesting.	2.38 (0.99)	4.41 (0.67)
Q3: The way their module was taught has helped me to learn.	2.64 (1.14)	4.28 (0.69)
Q4: I have received good support to manage my assessment workload.	2.45 (1.02)	3.85 (0.88)
Q5: Feedback has helped me develop and improve my performance.	2.73 (1.14)	4.02 (0.83)
Q6: This module has challenged me to do my best work.	2.67 (1.06)	3.96 (0.80)
Q7: Overall, I was satisfied with the teaching on this module.	2.39 (1.03)	4.36 (0.68)

Table 7 depicts all "worst" and "best" lecturers' overall scores on the seven SET items. I numerically recoded the verbal answer scale categories chosen by participants and calculated the means and standard deviations to comply with common standards in mainstream psychology. As expected, participants who rated their "worst" lecturer provided lower scores on all seven items compared to participants rating lecturers chosen as "best".

5.3.3.1 Considerable variations in lecturers' behaviours and attitudes that participants considered when rating their lecturers on the same item statement

Student participants frequently considered a wide range of behaviours and inferred attitudes for the same item. On average, participants of the sample considered per item between three to six themes (M = 4.93, SD = 1.07). Each theme contained between zero and five sub-themes (M = 1.90, SD = 1.44). However, each participant considered on average only between one to two pieces of evidence (M = 1.25, SD = 0.82). This suggests

that participants considered different information when rating their lecturers even on the same item (*see also Figure 7*). The themes slightly differed between items (*see Tables A3, A4 in Appendix*).

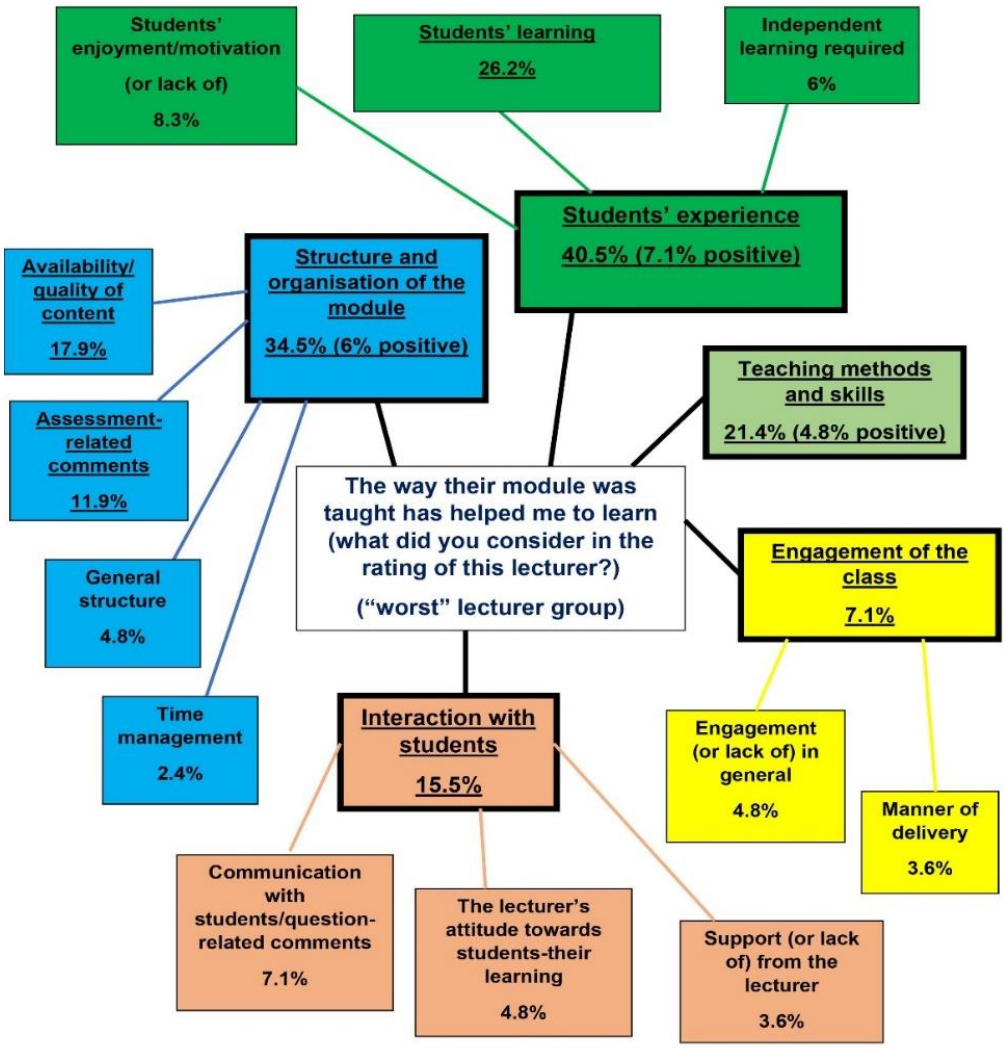

Figure 7

Different Information Considered by N = 84 Participants when Rating their "Worst" Lecturers on the "Effective Teaching" Item

Note. This figure portrays a diverse range of information that participants considered when rating their lecturers on the same item and the extent to which participants endorsed this information. The main themes are underlined and bold frames. They are connected to the related sub-themes. The percentages depict a number of participants who mentioned content related to the particular themes or sub-themes. The underlined sub-themes and themes indicate that content related to them was mentioned by over 10% of participants. Participants' comments could be categorised into several themes or sub-themes and percentages may, therefore, not add to 100%. Similarly, participants could have mentioned several sub-themes within the same theme, so percentages of the sub-themes may not always add to the total of the themes. Participants occasionally provided also positive comments and their frequencies are shown in brackets.

5.3.3.1.1 'Teaching skills and methods' as a reflection of teaching quality

SETs should enable students to evaluate *teaching quality*. Students should therefore mostly judge lecturers' teaching skills. In my study, participants considered 'teaching skills and methods' to some extent (for five SET items, between 13.4% and 41.2% of participants mentioned content categorised into this theme). Participants mentioned, "*The teacher was not willing to give good explanations*" (W62) and "*Lecturer talked us through problems thoroughly and showed us application of material to solving them. Explanation of background/ theory behind the practical techniques used. Clear and logical explanations*" (B10).

However, 'Teaching-related comments' was the most frequent theme only when participants rated the "overall satisfaction" item. Participants did not even mention teaching skills at all when evaluating their lecturers on the "support with assessments" and "feedback" items. Although the other identified themes (e.g., 'interaction with students', 'structure and organisation') may also contain ideas related to teaching quality, participants mentioned teaching skills relatively scarcely even when rating their lecturers on the items that appear to be closely related to teaching skills or methods, such as the "effective teaching" item or "level of challenge" item. This indicates that students may consider different referents (study phenomena; *see Chapter 3*) than intended by scale developers.

5.3.3.1.2 'Students' experience'

This theme included content related to students' reflections on their student experience. This could be seen in three different ways. For example, in the sub-theme 'Students' learning-performance-perceived usefulness of the module', participants mentioned their grades or learning process and whether they gained anything from the course. Another type of comments reflected students' motivation, enjoyment of the module, or feelings that I categorised into the sub-theme 'Students' interests-motivations-feelings'. Lastly, participants also commented on their ideas and experiences of 'Independent learning'. Participants generally considered their own experience when rating their lecturers on all but the "feedback" item. Some participants commented on their motivation, e.g., "*I didn't care enough*" (W77) but also on performance, e.g., "*I did well in the module*" (B23), "*My grades*" (W79).

Occasionally, participants even mentioned other students' behaviours or experiences as reasons for their ratings. For instance, on the "engagement" item, a participant provided, "*Opinions of other people on the course – attendance*" (B16) as a reason for their rating of their former lecturer. This provides evidence that some students may be influenced also by other students' opinions during their SET ratings. But this may result in biased ratings. For example, in another study, positive reputation (in terms of experience, qualifications, credentials, expertise, and overall reputation) of the evaluated lecturers artificially inflated their SET ratings (McNatt, 2022). Similarly, in another study, unfavourable information about the evaluated lecturer (in terms of their reputation, credentials and experience) negatively affected SET ratings (McNatt, 2010). This occurred even though participants attended the course for four months and, therefore, experienced teaching behaviours that were contrary to negative information they received about the given lecturer prior to evaluating them. This is problematic because it implies that student raters may consider opinions of other people more than teaching behaviours that have actually taken place. These opinions could also be influenced by stereotypical biases (*see also Introduction in this chapter*).

5.3.3.2 Differences in teaching behaviours and attitudes considered for "best" vs. "worst" lecturers

Student participants may prioritise different behaviours based on how they feel about their lecturers overall. For example, for their "worst" lecturers, participants

mentioned content related to the ‘Time and scheduling concerns’ on the “organisation” item more frequently compared to “best” lecturers (15.1% for “worst” versus 2.1% for “best”). This provides evidence that students may especially dislike if any problems arise in terms of time management, scheduling, or structure, but not particularly value if things run smoothly, similarly to assignment-related issues (*see above*).

Certain behaviours and attitudes were mentioned *only* for “best” but not “worst” lecturers. For instance, participants mentioned, content related to the ‘Passion-enthusiasm of the lecturer’ when rating the “organisation” and “effective teaching” items. When rating their lecturers on the “engagement” item and “overall satisfaction” item, participants mentioned content related to the ‘Use of humour’ and lecturers’ ‘Competence-confidence-expertise’ only when evaluating their “best” lecturers. Vice versa, participants mentioned ideas related to the theme ‘Lack of structure or clarity’ when rating “level of challenge” item only for their “worst” lecturers.

Occasionally, participants cited a certain teaching behaviour (e.g., ‘Engagement of the class’ and correspondingly, ‘Lack of engagement’), but the frequencies of the themes to which I categorised participants’ comments differed between “best” and “worst” lecturers. For example, when participants evaluated their lecturers on the “effective teaching” item, they mentioned teaching behaviours related to the ‘Engagement of the class’ more frequently for “best” (30.9%) as compared to “worst” (7.1%) lecturers.

In sum, participants appreciated if their lecturers exhibited passionate and enthusiastic behaviours or attitudes, used humour, or appeared competent, confident, or experienced. However, students rarely seemed to consciously penalise lecturers for a lack of these behaviours or attitudes. Similarly, many participants described engagement of the class as a factor that helped them to learn, but only a few considered a lack of engagement as hindering to their learning.

5.3.3.3 Unexpected findings

5.3.3.3.1 Behaviours mentioned may better fit the intended meanings of the other items

Student participants often mentioned behaviours that would be more suited to other SET items. For example, on the “organisation” item, participants frequently justified

their ratings with content related to the theme a lecturer's 'Interaction with students', which also included the sub-themes 'Support-guidance' and a 'Lecturer's attitude towards students and their learning'. This would seem to better fit the intended meanings of the other items, such as "support with assignments" item. Similarly, on the "effective teaching" item, participants explained their ratings more often with reasons related to the 'Structure and organisation' compared to 'Teaching methods and skills', mentioned only by 14.4% (for "best" lecturers) and 21.4% (for "worst" lecturers). But the wording of the item "the way their module was *taught*" implies that students should consider teaching (methods or skills) during rating. When participants rated lecturers on the "challenge" item, instinctively, it may be expected that students would consider content related to the 'Level of challenge', but it was one of the least frequent themes. Several participants also stated that item ("This module has challenged me to do my best work") did not apply to them because they strive to do their best regardless of lecturers, directly disagreeing with items that put them into passive roles. This is an important finding, because it emphasises the active role of students that students see for themselves but that these item statements diminish. The wording of this item shifts the responsibility for student learning to their lecturers. Specifically, this item is framed in the way that may suggest that lecturers should challenge students to do their best work without considering that students may perceive themselves as independent and active learners who want to achieve their best work. These examples provide evidence that meanings that students construct for SET items may therefore overlap between items and differ from the meanings that developers intended for these items.

5.3.3.3.2 Item statements not applicable to all students

Some item statements did not even seem to apply to all participants. For example, regarding the "support with assessments" item, some participants reported they did not need any support, e.g., "*I feel this is something you have to manage and organise yourself…"* (B68) or *"[Though support is available] managing assessment workload is mostly up to the student. I do what I can to manage the work*." (B72). However, students may think they must provide ratings, and this could then negatively affect SET ratings of their lecturers.

5.3.3.3.3 Counter-intuitive comments

Participants also frequently mentioned positive behaviours or attitudes for their "worst" lecturers or commented negatively on their "best" lecturers. This was especially obvious in comments about ratings of the "feedback" item. Participants commenting on their "best" lecturers mentioned content related to the theme 'problems with feedback' (unclear, vague, late; 10.3%), whereas those commenting on their "worst" lecturers reported information that I categorised into the theme 'Feedback was useful-interesting-organised' (16.7%), about a positive aspect of feedback, such as its constructiveness or a great amount of detail. For example, participants said, "*The feedback provided was useful, there were explanations on how to improve and a clear summary of expectations.*" (W52) or "*Feedback on my assignment was well structured and told me what I did well and what I could improve upon*." (W32), as examples of positive comments about feedback from their "worst" lecturers. In contrast, participants commenting on their "best" lecturers stated, "*I didn't receive much feedback for this assignment even though I received a low grade.*" (B50) and "*I have not received much detailed feedback*" (B9). However, feedback that lecturers could provide may also depend on how participants performed during their assignments. For example, if the essay was excellent, lecturers may struggle to deliver detailed feedback.

These contradictory teaching behaviours that participants mentioned may be explained in several ways. For example, participants could have appreciated some of their "worst" lecturers' teaching behaviours even if they were dissatisfied with the lecturer overall. Similarly, participants could have valued their "best" lecturers overall, but still dislike a specific teaching behaviour or other elements of the module (e.g., subject). Alternatively, participants may have misread the question, misunderstood the instructions or be distracted during completing SETs. This reflects potential problems that can also apply to real-life SETs, because students may misread the questions or los**e** their focus on completing SETs.

5.3.4 RQ4: In what ways do students interpret the item statements of SET scales in general, and what meanings do students generally construct for these statements?

I analysed participants' *general* interpretations of the seven SET items, specifically, their answers when asked to interpret the item content generally, thus not with regard to a specific lecturer. This enabled me to explore what meanings students construct for SET items.

5.3.4.1 Considerable variations in participants' general interpretations of the same item

Student participants (as a group) considered on average per item three to six themes (M = 4.79, SD = 0.80) and each theme with between zero and six sub-themes (M = 1.73, SD = 1.62). This suggests that the same SET item statement evoked a wide range of meanings (a field of meaning as described by the different themes) between participants, who did not seem to interpret SET items in standardised ways (*see Tables A3, A4 in Appendix*).

I identified several themes that participants considered frequently across several items. For example, participants mentioned content related to the theme 'Structure and organisation' (or its sub-themes such as 'Content-related comments') not only when interpreting the "organisation" item, but across all other SET items (apart from "the feedback" item). This theme included comments about content, general structure and organisation, assessments, and time management and scheduling. Similarly, participants almost always considered ideas related to the theme 'Students' experience', which included participants' experiences with learning, their feelings about the modules and opinions on independent learning. Participants frequently mentioned these ideas, especially when interpreting the "engagement", "effective teaching", "challenge" and "overall satisfaction" items. This further supports the idea that meanings which students construct for different SET items may often overlap between these items.

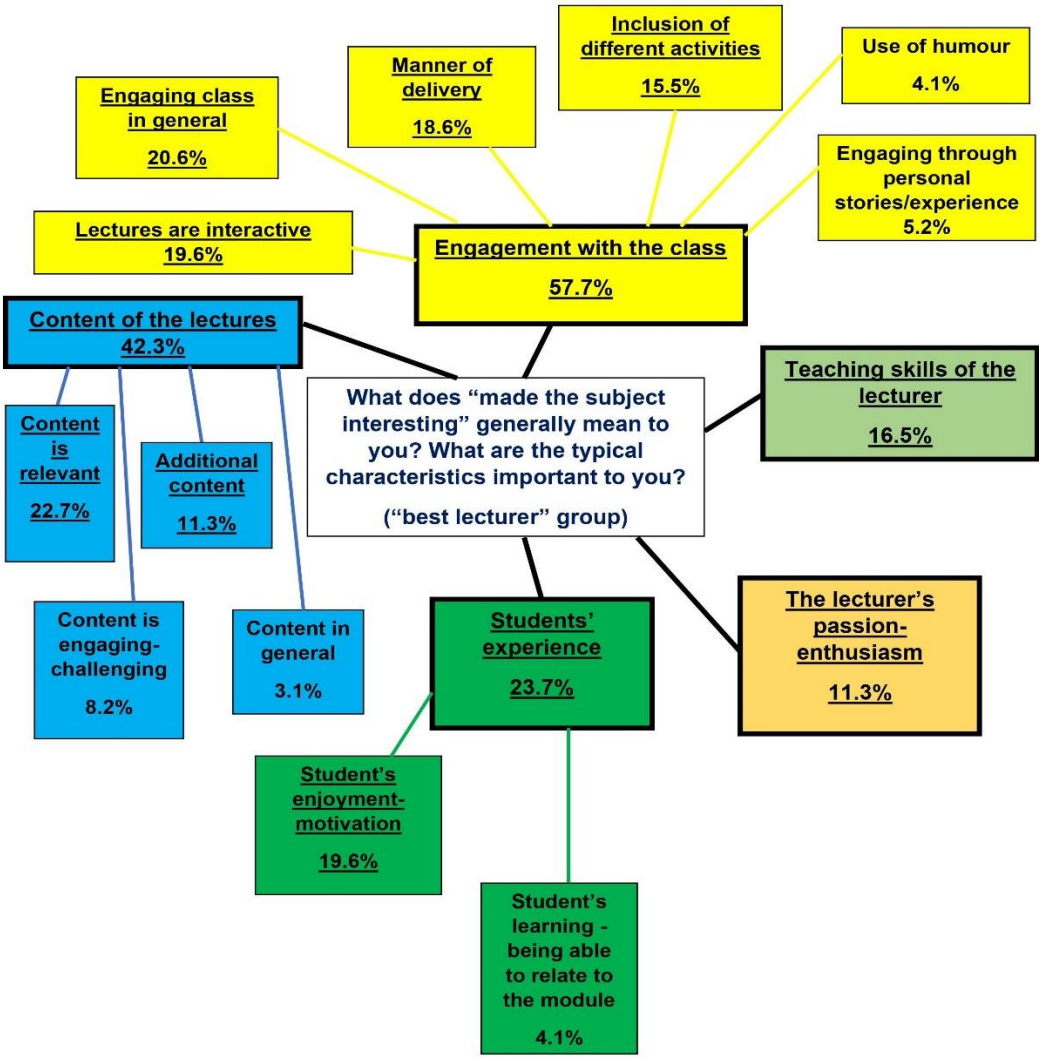

Figure 8

The Field of Meaning Constructed by N = 97 Participants for the "Engagement" Item

Note. This figure portrays a large field of diverse meaning that participants constructed for the same item. The main themes are in a bold frame. They are connected by black lines to the related sub-themes. The percentages depict a number of participants who mentioned content related to the particular themes or sub-themes. The underlined sub-themes and themes indicate that content related to them was mentioned by over 10% of participants.

Participants also sometimes provided comments that could be interpreted as very generic. For example, participants simply stated "*organisation*" (B2) as their interpretation of the "organisation" item, and "*notes on assignment*" (B44) and "*Challenging*" (W39) as their interpretations of the "feedback" and "level of challenge"

items without any further elaboration. This may imply that some students interpret item statements on a rather superficial level or were not motivated to write more.

Figure 8 illustrates how even the same SET item statement evoked a wide range of meanings between participants. In this example, participants considered different referents of the word ‘interesting’ (e.g., content, interaction, students’ own experience, manner of delivery, use of humour, teaching skills of their lecturers or lecturers’ enthusiasm and passion).

5.3.4.2 Unexpected findings

Interestingly, participants interpreted the “effective teaching” item less in the relation to the theme ‘Teaching skills and methods’ or even ‘Students’ learning’ and more often in terms of the ‘Structure and organisation’ or ‘Interaction with students’ themes (*see Figure 9*). This suggests that students may view structure, organisation, and successful interaction with students as a part of effective teaching. Furthermore, even when students evaluate their lecturers on items designed specifically to assess effective teaching (learning), students may prioritise other aspects of their experience.

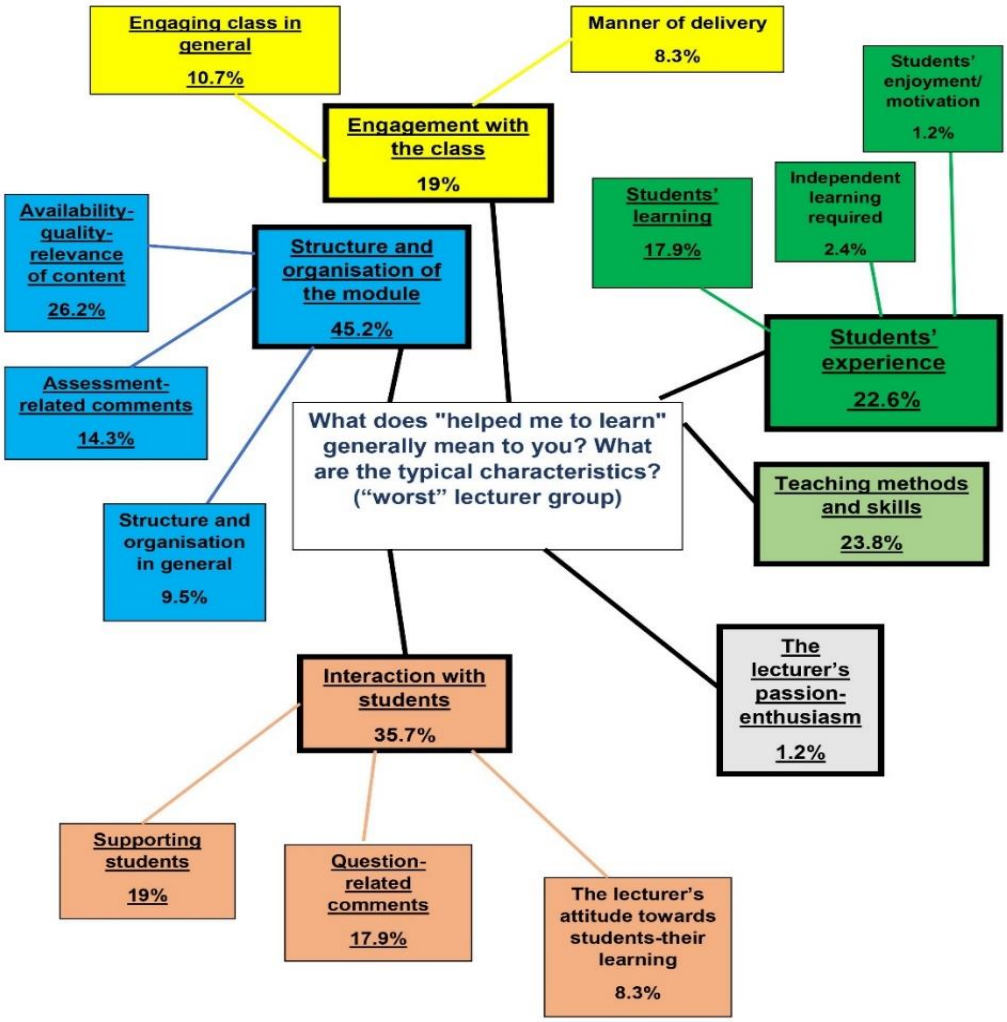

Figure 9

The Field of Meaning Constructed by N = 84 Participants for the "Effective Teaching" Item

5.3.5 RQ6: What are possible differences in the teaching behaviours and inferred teaching attitudes that students consider either generally or in SET ratings of their male versus female lecturers?

I analysed participants' ratings of their lecturers on the seven SET items for any significant differences between the scores of male and female lecturers with One-Way MANOVAs. Because missing values may negatively affect statistical power and lead to a

wrong interpretation of findings (e.g., Kang, 2013), I imputed missing values (4.71% of the overall dataset of participants evaluating their "best" lecturers and 3.1% evaluating their "worst") with EM (expectation-maximization) analysis. This is a method that enables addressing missing values in the dataset by imputing new values approximated through the maximum likelihood methods (Dempster et al., 1977; Kang, 2013).

5.3.5.1 Overview and interpretation of the findings

5.3.5.1.1 Differences in ratings based on a lecturer's gender

Overall, "worst" male and female lecturers did not receive significantly different ratings across all seven SET items, $F(7, 76) = 0.29$, $p = .957$, Wilk's $\Lambda = 0.974$, $\eta p2 = .026$. Therefore, to avoid examining data for statistically significant comparisons and finding a false positive effect, I did not consider univariate effects (Simmons et al., 2011). However, "best" male and female lecturers received significantly different ratings on the combined seven items, $F(7, 88) = 2.38$, $p = .028$, Pillai's trace = 0.159, with a large effect size, $\eta p^2 = .159$[24]. Exploring univariate effects revealed that participants rated their "best" male lecturers significantly higher ($M = 4.49$, $SD = 0.58$) than female lecturers ($M = 4.15$, $SD = 0.80$) on the "overall satisfaction" item, $F(1, 96) = 5.94$, $p = .017$, with a medium effect size, $\eta p^2 = .059$. In sum, participants did not rate their "worst" male and female lecturers differently. However, my findings show participants awarded to their "best" male lecturers overall significantly higher scores than to their "best" female lecturers, especially when participants rated their "overall satisfaction" with the module. However, this significant difference between ratings may have been influenced by only a minority of participants awarding significantly higher scores to their male than female lecturers, rather than most of the sample.

Similar patterns were observed in previous research. For example, participants chose lecturers of their own gender as "best" more frequently than expected (Basow, 2000; Sprague & Massoni, 2005), or only male participants did so (Basow et al., 2006), but these differences were not observed when participants chose their "worst" lecturers. Similarly, participants rated all lecturers high but provided more positive comments for

[24] I interpreted $\eta p^2 = .01$ as small, $\eta p^2 = .06$ as medium, and $\eta p^2 = .14$ as large *(Cohen, 1988; see Data analysis section*)

their male rather than female lecturers for the identical teaching performance (Khazan et al., 2019). Individuals implicitly associate brilliance with men more than with women (Storage et al., 2020), which could also influence SET scores. Consistent with the previous findings (Rivera & Tilcsik, 2019; Storage et al., 2016), perhaps participants in my sample judged female lecturers as very good overall, but still did not associate them with brilliance and therefore awarded them high but still lower scores than their male lecturers. These observed patterns (e.g., withholding of praise) imply that students may be influenced by subtle prejudice when evaluating their lecturers.

My findings could be also explained by the role congruity theory (Eagly & Karau, 2002). Specifically, the role of a lecturer could be associated with leadership and masculinity, which may contradict traditional expectations of a woman's gender role, such as empathy and warmth. Similarly, according to the status incongruity hypothesis (Rudman et al., 2012), when evaluating female but not male leaders, individuals may perceive an incongruence between ascribed status as a woman (which may be perceived as low) and achieved status of a position (seen as high). Because, according to these theories, leadership positions may be seen as less desirable for women, students may then perceive female lecturers less favourably than men and evaluate them slightly less positively. These findings are also consistent with a stereotype content model (Fiske et al., 2002), according to which men (but not women) are likely to be rewarded for displaying competence (*see Chapter 2*).

Participants awarding male lecturers higher scores than female lecturers on the "overall satisfaction" item is also consistent with the previous evidence (Basow, 1995; Boring, 2017). Importantly, this item is highly abstract and does not refer to any specific aspects that students should consider (unlike, e.g., "support with assessments"). This could lead to even broader fields of meaning that students constructed when they interpreted this item and what study phenomena they consider when rating their lecturers. Due to this abstract wording and lack of definition students must rely on their subjective interpretations that can be affected by different expectations and beliefs, including stereotypical beliefs. It is well-established in research that individuals interpret, consider and generate evidence in the ways consistent with their beliefs and expectations (Nickerson, 1998; Stanovich et al., 2013). This may include stereotypical biases and beliefs, as shown in empirical research (Dodson et al., 2008; Lenton et al., 2001; Sherman et al., 2003; Uher et al., 2013; Uher & Visalberghi, 2016). In fact, some authors (e.g.,

McCullough & Radson, 2011) specifically advise against including "overall" items in evaluations.

Furthermore, the broad field of meaning that students may construct for this item could make rating cognitively demanding for students. But individuals tend to rely more on their stereotypes when their cognitive resources are depleted (Curley et al., 2022; Petty & Cacioppo, 1986). In fact, evidence shows that being presented with ambiguous information may reinforce stereotypes (De la Fuente et al., 2003). According to the cognitive miser theory, to avoid this depletion of cognitive resources, individuals may apply mental shortcuts and rely on categorical and potentially stereotypical thinking (Fiske & Taylor, 1991; Macrae & Bodenhausen, 2001; Taylor, 1981). Instead of considering all available information about a target person, individuals may categorise this person according to stereotypical information about the member of their category (e.g., female lecturer; (Corcoran & Mussweiler, 2010; Macrae et al., 1994). Students in my sample may have, therefore, applied heuristics when judging their lecturers on this non-specific item and recalled stereotype consistent information more easily better than other information inconsistent with stereotypes. Other SET researchers also highlighted that this item is highly subjective, interpreted in different ways and should be avoided (McCullough & Radson, 2011). My findings support this view. However, they also show that pronounced variations in interpretations unfortunately relate to also other SET items (*see 5.3.4; RQ4*).

Table 8

Means and Standard Deviations for the Overall Participants' Ratings of their "Worst" and "Best" Female and Male Lecturers

	"Worst" lecturers					**"Best" lecturers**				
	Female		**Male**			**Female**		**Male**		
ITEM	***M* (SD)**	***N***	***M* (*SD*)**	***N***	ηp^2	***M* (*SD*)**	***N***	***M* (*SD*)**	***N***	ηp^2
Q1: Their module was well-organised.	2.63 (1.06)	40	2.75 (1.30)	44	.003	4.35 (0.72)	36	4.49 (0.50)	60	.014
Q2: The lecture has made the subject interesting.	2.40 (0.98)	40	2.36 (1.01)	44	.000	4.33 (0.72)	36	4.44 (0.65)	60	.006

Q3: The way their module was taught has helped me to learn.	2.60 (1.13)	40	2.69 (1.16)	44	.001	4.17 (0.68)	36	4.36 (0.69)	60	.018
Q4: I have received good support to manage my assessment workload.	2.55 (0.99)	40	2.36 (1.06)	44	.008	3.93 (0.85)	36	3.81 (0.91)	60	.005
Q5: Feedback has helped me develop and improve my performance.	2.73 (1.15)	40	2.74 (1.15)	44	.000	3.96 (0.80)	36	4.06 (0.86)	60	.003
Q6: This module has challenged me to do my best work.	2.60 (1.01)	40	2.74 (1.11)	44	.004	3.82 (0.84)	36	4.05 (0.77)	60	.019
Q7: Overall, I was satisfied with the teaching on this module.	2.34 (0.94)	40	2.43 (1.11)	44	.002	4.15 (0.80)	36	4.49 (0.58)	60	.059

Exploring Table 8 shows that male lecturers received higher scores than female lecturers on most items. But both "worst" and "best" male lecturers obtained lower scores than female lecturers on the "support with assessments" item, and "worst" male lecturers also on the "engagement" item, although these differences were not significant.

5.3.5.1.2 Differences in teaching behaviours or attitudes considered for male versus female lecturers based on the lecturer's gender

I used chi-square analyses to examine whether participants consider certain teaching behaviours or inferred attitudes more often for male or female lecturers when reflecting on two teaching behaviours or inferred attitudes of these lecturers. Specifically, I considered the frequencies by which participants mentioned content related to themes or sub-themes of these behaviours and attitudes. Participants provided this information after reflecting but *before rating* their "best" or "worst" lecturer on the scale. There were no differences for 97.7% of analysed teaching behaviours and attitudes[25]. However, over a third of participants considered 'engagement-related' behaviours significantly more frequently for their "best" male (36.08%) versus best female (13.4%) lecturers, $X^2(1)$ =

[25] Because my aim was to provide an overview of possible patterns for future exploration, I did not correct for multiple analyses in these frequency analyses.

4.444, p = .035. There were, however, no gender differences in behaviours or attitudes considered for participants' "worst" lecturers, although the 'assignment-related concerns' theme was mentioned more than twice as often for female (11.9%) rather than male (4.76%) lecturers, $X^2(1) = 3.818, p = .051$. This is consistent with findings by Sprague and Massoni (2005), in which students held their male (but not female) lecturers to an entertainer standard.

5.3.5.1.3 Differences in teaching behaviours or attitudes considered for male versus female lecturers based on the lecturer's gender during rating

I *used chi-square analyses* to explore whether participants considered information *across all items* similarly frequently for male and female lecturers to generate their ratings. Specifically, I studied whether participants mentioned certain behaviours and attitudes more frequently for male or female lecturers *during ratin*g. This analysis only examines the frequency of themes of teaching behaviours across all items, not how they relate to ratings (*see RQ7 for the relation to ratings*).

I examined teaching behaviours and attitudes related to the themes 'Structure and organisation', 'Teaching methods and skills', 'Interaction with students', 'Students' experience' and 'Engagement-related' comments. I chose these themes because content related to them appeared in several items and therefore seemed representative of the overall dataset. I also considered 'Quality of feedback' although participants only mentioned content related to this theme when evaluating their lecturers on the "feedback" item, to include a content representative of this item in my analysis. Overall, participants considered content related to the analysed themes similarly frequently for male and female lecturers (*see Table A5, Appendix*), with no differences in 91.7% of analyses[25]. The only significant association was between the gender of the "worst" lecturers and "poor quality of feedback", $X^2(1) = 5.624, p = .018$, with participants on average considering poor quality of feedback more often when rating their "worst" female (20.62%) rather than male lecturers (11.34%).

5.3.6 RQ7: How do students weight specific teaching behaviours and inferred attitudes when generating their ratings of the lecturers on seven SET items, and does this differ by the lecturers' gender?

In order to explore how much participants weighted the specific information (teaching behaviours, inferred attitudes) they considered when rating their *lecturers* and

whether it differed by their lecturers' gender, I conducted bivariate correlation analyses between the number of mentioned sub-themes and the ratings participants provided for the lecturers evaluated.

For example, participants could frequently mention positive teaching behaviours, but this may not relate to higher rating scores. In contrast, participants could consider negative behaviours sparsely, but still rate their lecturers lower. In terms of gender differences, participants might consider certain teaching behaviours (e.g., positive content about 'Teaching methods and skills') less frequently for lecturers of certain gender (e.g., female lecturers), but still award them higher ratings than the lecturers of the other gender. This would imply that how students weight teaching behaviours and attitudes in their ratings may differ by their lecturer's gender.

After applying the Benjamini-Hochberg procedure, the highest p values I deemed significant were $p = .014$ for ratings awarded to "best" lecturers and $p = .033$ for ratings of "worst" lecturers.

5.3.6.1 General overview

As expected, generally, the more positive content participants mentioned for their "best" lecturers, the higher they rated their lecturers. In contrast, the more frequently participants mentioned negative content, the lower they rated their "worst" lecturers.

Content related to the themes of certain teaching behaviours correlated more strongly with how participants rated their lecturers (compared to the themes of other teaching behaviours) on several items. For example, the more frequently participants considered 'Poor interaction with students', the lower participants rated their "worst" lecturers, both male ($M = 2.36$, $SD = 1.06$[26]), r (42) = -.58, $p < 0.001$ and female ($M = 2.55$, $SD = 0.99$), r (38) = -.59, $p < 0.001$, on the "support with assessments" item. Similarly, the more frequently participants considered 'Poor structure and organisation of the module', the lower participants rated their "worst" male ($M = 2.75$, $SD = 1.30$), r (42) = -.49, $p = 0.002$, and female lecturers ($M = 2.62$, $SD = 1.06$), r (38) = -.41, p =.008 on the "organisation" item.

Therefore, students may reflect on a wide range of negative teaching behaviours and attitudes but consider some of these teaching behaviours more strongly in their

[26] This shows the means and standard deviations of ratings that participants gave their lecturers.

ratings. In this study, this occurred if participants perceived lecturers interacted with them in an unsatisfactory manner or their module was poorly organised.

5.3.6.1.1 Differences in weighting of negative content for "worst" lecturers of a different gender

I explored the correlations between negative content that participants considered when they rated their lecturers and their ratings of these lecturers, and whether this differed for their male and female lecturers. This would suggest that participants may weight specific teaching behaviours differently in their ratings of lecturers of a certain gender.

Participants weighted negative content about the 'Lack of engagement with class' only in their ratings of their "worst" male lecturers (M = 2.36, SD = 1.01), r (42) = -0.34, p = .023 on the "engagement" item, but not female lecturers (M = 2.40, SD = 0.98), r (38) = 0.04, p = .788. Interestingly, participants considered more negative content for their female than male lecturers. However, participants weighted this negative content more strongly when they rated their male lecturers. Therefore, the more frequently participants mentioned lecturers engaged poorly with class, the lower they rated their "worst" male lecturers. Similarly, the more frequently participants mentioned 'Poor quality of feedback', the significantly lower they rated only their female (M = 2.73, SD = 1.15), r (38) = -0.34, p =.033, but not male lecturers, (M = 2.74, SD = 1.15), r (42) = -0.13, p = .386, on the "feedback" item (*see Figure* 10). These ratings are almost the same, however, participants weighted their considerations more strongly in their ratings of female but not male lecturers.

In sum, participants provided similar considerations and ratings when evaluating both male and female lecturers. However, participants weighted certain considerations differently depending on their lecturer's gender. Specifically, participants rated their male lecturers lower for lessons that they perceived to be boring, and their female lecturers lower if participants thought female lecturers provided feedback of a poor quality.

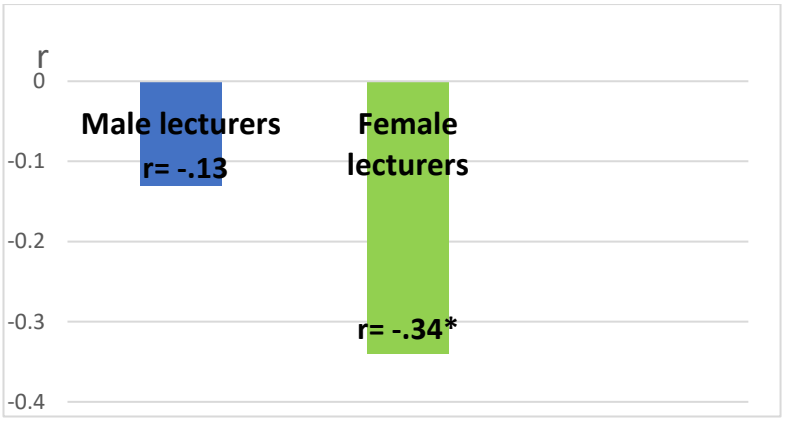

Figure 10

Correlation of Negative Content Related to Poor Quality of Feedback with the Ratings on the "Feedback" Item for the "Worst" Male and Female Lecturers

Note: * p <.05.

5.3.6.1.2 Differences in weighting of positive content for "best" lecturers of a different gender

Participants weighted positive content about the high quality of 'Structure and organisation', only in their ratings of "best" female lecturers (M = 4.35, SD = 0.72), r (34) = 0.47, p = .004 on the "organisation" item, but not male lecturers, (M = 4.49, SD = 0.83), r (58) = 0.20, p = .121.

Similarly, participants weighted positive 'Interaction with students' only in their ratings of female lecturers (M =3.93, SD = 0.85), r (34) = 0.41, p = .014, on the "support with assessments" item, but not male lecturers (M = 3.81, SD = 0.91), r (58) = 0.16, p = .218. Interestingly, participants considered positive 'Interaction with students' more frequently for their male than female lecturers, even though participants rated their female lecturers higher. Therefore, the more often participants mentioned positive content about how lecturers interacted with students or structured and organised the module, the higher participants rated but only their female lecturers.

The more frequently participants mentioned 'High quality of feedback', the higher they rated their "best" male lecturers (M = 4.06, SD = 0.86), r (58) = 0.33, p = .011, but not their female lecturers (M = 3.96, SD = 0.80), r (34) = 0.23, p = .172 on the "feedback" item. Participants mentioned high quality feedback less frequently for male than female

lecturers, even though participants rated male lecturers higher. Therefore, participants seemed to weight positive content related to feedback in ratings more strongly for their male lecturers.

In sum, participants may have weighted specific teaching behaviours and attitudes in their ratings differently, especially poor quality of structure of the module or poor interaction with students, that were related to lower ratings of all lecturers. But how participants weighted the considered information when rating their lecturers differed between male and female lecturers in some cases. For example, the more positive content participants considered about structure and organisation, as well as interaction with students, the higher they rated but only their female lecturers, whereas the more positive content about the high quality of feedback participants considered, the higher they rated only their male lecturers. In contrast, the more frequently participants mentioned poor feedback, the lower participants rated only their female lecturers, whereas if participants considered lack of engagement, they rated their male lecturers lower.

This is in line with previous findings that students tend to think of their "worst" male lecturers as "boring", and "best" female lecturers as "caring" (Sprague & Massoni, 2005). Participants penalised poor engagement with the class when they rated their "worst" male but not "worst" female lecturers. Similarly, participants considered positive interaction salient when rating only their "best" female lecturers. These findings are also consistent with the stereotype content model, which suggests that women may be praised for behaviours related to warmth (e.g., positive interaction with students). Previous evidence suggests that students may evaluate their female (but not male) lecturers on the favorability of feedback (Sinclair & Kunda, 2000), but participants in my research focused more on perceived quality of feedback.

5.3.7 RQ9: How do students regard SETs and approach their completion in general?

To answer this question, I asked participants to estimate time they think they spent on completing SETs and analysed these participants' statements about their opinions and approaches to SETs.

5.3.7.1 Overview and interpretation of the findings

A sub-sample of participants (N = 34) overall estimated spending approximately 11 minutes on student evaluations (M = 11.00, SD = 8.75). Most participants (70.6%) said

they complete evaluations spontaneously when asked, whereas a lower number of participants reported thinking about them in advance during the year (14.7%)[27]. However, past research has shown that spontaneous evaluations could be affected by many other factors, even unrelated to teaching quality, such as bad weather (Braga et al., 2014; *see Chapter 2 for a discussion of other factors*).

Most participants held positive opinions about SETs and believed that feedback from SETs can improve the lectures or that SETs are useful, important, and needed. Participants mostly expressed the belief that SETs can provide their lecturers with more information about students' perspectives, e.g., "*They can know what students are seeking from a lecturer.*" (B84), "*Very useful in order for lecturers to see and understand what makes a good teacher to help them do better in the future*" (B91) or "*It is very vital for every faculty/university to ensure student satisfaction. It would help improve or help the faculty in understanding where work needs to be put through as not everyone understands how opinions are generated. Feedback or evaluation of faculty is a nice technique to help improve the bond of the faculty and student, eventually increasing the value of the university.*" (W81). Participants also frequently emphasised that SETs should be taken more seriously, e.g., "*i think they should actually use them rather than just collecting them*" (W83) or "*They should be read out loud with the academic board and actions should be decided upon reading all feedback*" (B86), "*Whilst some of our course leaders have actively responded to a whole class regarding the feedback received, I believe few changes have been seen."* (B92). However, a considerable number of participants also emphasized negative elements of SETs, such as broad, unclear, and frustrating questions (11.76%). For example, participants commented, *"[they can be moderately helpful I suppose], but would be better if the questions were more clearly explained*" (B80) or *"[They should be ruthless and honest. A lot of teaching can be lackluster and] many evaluations ask too broad questions which are frustrating, unengaging and boring to answer."[sic]* (B82). Some participants also acknowledged a potentially problematic nature of using SETs as the only source of information, e.g., "*I understand why such evaluations are sort, but from my experience the response to evaluations is far too forceful rather than subtle and considered. Not everything a lot of students say will be correct. Not everything a lecturer believes will be true. Often, it is only these groups that*

[27] The remaining participants were not sure, provided general comments, or did not answer the question.

are heard" (W74) or "*They are necessary BUT students should not use them to needlessly criticise lecturers*" (W69). Participants also highlighted insufficient anonymity, e.g., "*I feel that they are often given in a way that you can not give your honest opinion as the lecturer is often in the room.*" (W70) or doubts of SET's effectivity, "*Nothing improves*" (W79).

Student participants provided some recommendations for improvement, such as SETs should be done "*after every lecture interactively*" (B81) or "*throughout the year as opposed to just at the end*" (B89; *see Figure 11*). However, these participants may be in minority, as universities frequently report a low number of respondents for SETs (*see Chapter 2*). Perhaps participants who chose to complete my study that involved a considerable amount of writing were more positively inclined towards SETs than other students and would like to complete evaluations frequently.

Participants generally reported putting an adequate effort into completing SETs and most stated they approach SETs spontaneously, which can be influenced by different factors (e.g., mood, weather, *see Chapter 2*). Overall, participants provided a mixture of perspectives on SETs. Participants overall perceived SETs positively, which is in line with previous evidence (Kite et al., 2015; Spencer & Schmelkin, 2002; Spooren & Christiaens, 2017). The majority also felt that SETs are important and should be used to improve lecturers. However, the others highlighted potential problems, such as a potential misuse of SETs by students, and insufficient anonymity. Other participants found SET questions to be too frustrating, vague, and unclear or felt like SETs do not lead to any impactful changes. Importantly, how participants perceive SETs, and their purpose may affect how much thought and effort they put into completing them or even their decisions to participate.

These findings suggest students differ regarding what they expect from SETs and how they perceive and understand SETs. For example, some participants conceptualised SETs as simply the source of information for the lecturer, or even a tool for strengthening the bond between lecturers and students. In contrast, some advocated SETs should be ruthless or read aloud at academic boards, with action taken based on student feedback. This implies that students likely vary in their understanding of how to use SETs, SETs purpose, as well as a general role of lecturers.

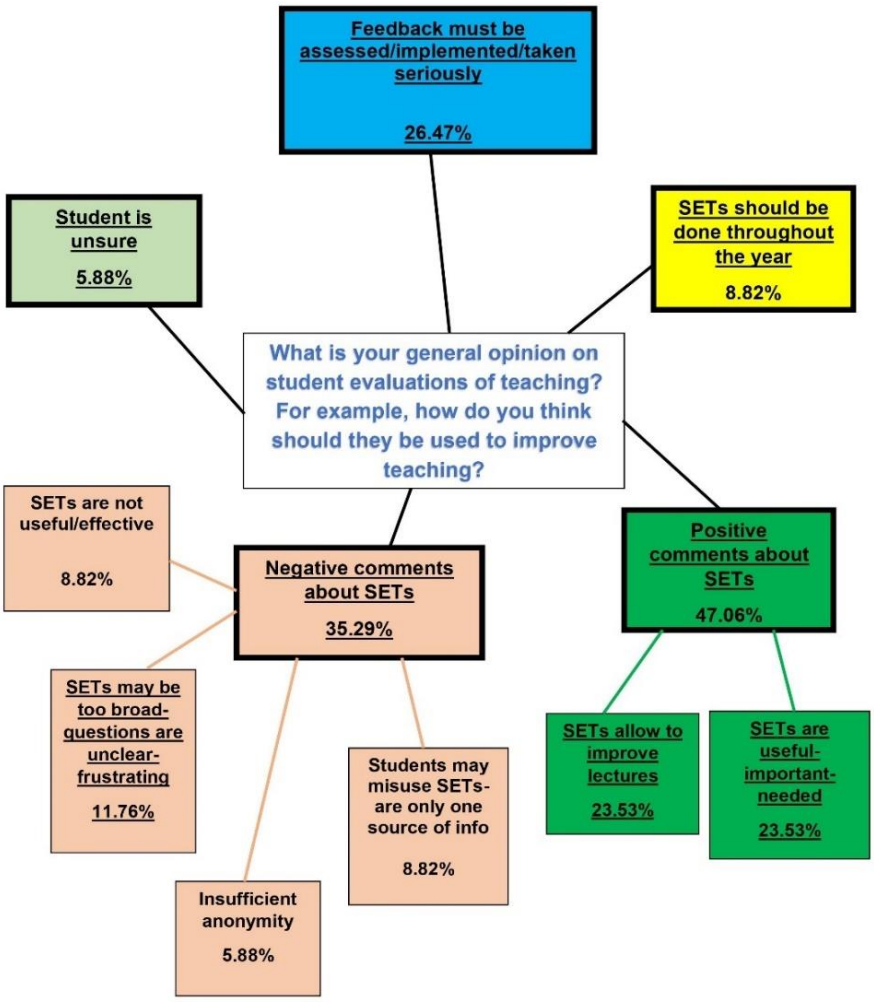

Figure 11

An Overview of Participants' General Opinions on SETs

Note. The themes are displayed in the bold frames.

5.4 General summary of this chapter

In sum, in Study 1 discussed in this chapter, I focused on students' interpretations of the item statements. Participants considered a wide range of teaching behaviours and attitudes when reflecting on or evaluating their lecturers. For each single item, participants as a sample constructed a large field of meaning. Importantly, participants interpreted

each standardised item in distinct ways. Participants also frequently constructed meanings that overlapped between items.

In Study 2, presented in the next chapter, I examine the answer scale categories and explore how students choose these categories. I also investigate the overall sample's field of meanings that students construct for answer categories and potential individual differences in how students interpret and use these categories.

Chapter 6: Study 2

In Study 1, I investigated how students interpret SET item statements; Study 2 examines how students interpret and use answer scale categories to rate lecturers.

The research aims of this study are: a) to investigate what specific information (teaching behaviours, inferred attitudes) students consider when rating lecturers described in scenarios by exploring types of reasons that participants provided for picking the answer categories (RQ3), b) to examine in what ways students interpret SET answer scale categories when completing SETs by studying general patterns but also potential individual differences in the participants' reasons for picking the answer categories (RQ5), c) to investigate whether students are influenced by gendered schemas and hold their lecturers to gender-related expectations when evaluating their lecturers (RQ6), d) to explore if, and if so, how students weight the reasons they considered for picking the answer category differently in their ratings of lecturers and whether this weighting differs by the lecturers' gender (RQ8). Additionally, I also aimed to explore whether students rate their female and male lecturers differently even for the same teaching behaviours and whether the direction of scale affects students' ratings.

6.1 Introduction

6.1.1 Interpretation and use of answer scale categories

Raters' answers may be affected by various response biases, such as extreme response style, random responding or socially desirable responding (e.g., Podsakoff et al., 2003; Wetzel et al., 2016). For example, extreme response style occurs when respondents prefer to choose extreme categories (the categories on the lowest or highest end of scale, such as "strongly disagree"; e.g., Hamilton, 1968; Wetzel et al., 2016). In contrast, some individuals report reluctance to choose the lowest or highest answer category(Block, 1998; Uher, 2017).

Random responding (also known as careless or non-contingent responding) happens when raters apply a seemingly random pattern, such as repeating the same response or response sequence over again without considering item content (Meade & Craig, 2012; Wetzel et al., 2016). Raters respond in a socially desirable manner when they reply in the ways that conform to socially acceptable values regardless of their true opinions (e.g., Podsakoff et al., 2003). This can heavily depend on culture, because cultural norms may differ between countries (Kemmelmeier, 2016).

When raters evaluate other individuals, raters may be influenced by further biases that are specific to assessment of others (Wetzel et al., 2016). These biases include the halo or horns effect (tendency to like or dislike all the features of the evaluated individual, including behaviours that raters did not observe; MacDougall et al., 2008; Murphy et al., 1993; Thorndike, 1920; Wetzel et al., 2016) and the leniency or severity effect (tendency to evaluate others too lightly or too harshly; Guilford, 1954; Podsakoff et al., 2003; Wetzel et al., 2016).

Importantly, people's judgements of others are always relative. Raters compare behaviours of the evaluated individuals to behaviours of others (called reference group), but the context of comparison is subjective and may differ between raters (Wood et al., 2012). For example, in academic context, students may perceive a lecturer who finishes classes on time in 90% of cases as managing their time effectively or poorly, depending on the behaviours of other lecturers. Social and cultural context can further influence students' judgements further. For example, individuals may differ in how they perceive time and punctuality across countries (Levine et al., 1980; van Eerde & Azar, 2020; White et al., 2011). Thus, because of different social and cultural norms, individuals from distinct backgrounds may use different contexts of comparisons.

Surprisingly, despite a frequent use of rating scales in psychology, only a few psychologists highlighted and examined potential variations in how raters interpret answer scale categories (e.g, Rosenbaum & Valsiner, 2011; Uher, 2018b, 2017). Their empirical findings showed that participants interpreted scale categories in distinct rather than standardised ways. Participants also frequently reported different reasons even for choosing the same answer category or provided identical reason for picking different answer categories.

In the context of SETs, research about how students choose, interpret and use their answer categories when rating their lecturers is rather limited (e.g., Benz & Blatt, 1996; Block, 1998; Robertson, 2004). However, consistently with the findings reported above, student participants in this research varied in how they interpreted and used answer scale categories (Block, 1998; Robertson, 2004). Some raters may even tick a scale mid-point (e.g., "neither agree nor disagree") instead of non-applicable. For instance, student participants in the U. K. provided scores for non-applicable item statements even if the "non-applicable" answer category was available (Ashby et al., 2011; Robertson, 2004, *see Chapter 3*).

Furthermore, little is known about whether the direction of rating scales (starting with "strongly agree" versus "strongly disagree") influences how raters interpret answer categories. Evidence shows that participants may prefer categories on the left side of the scale (Friedman & Amoo, 1999). However, there seems to be no research in the U. K. exploring this issue in SET context.

Students sometimes tend to rate female lecturers lower than male even if they demonstrate identical teaching behaviours (e.g., Arbuckle & Williams, 2003; MacNell et al., 2015) or use the same teaching materials (e.g., Mengel et al., 2019; Mitchell & Martin, 2018; Özgümüs et al., 2020, *see Chapter 2*). In Study 1, participants rated their specific lecturers who could have behaved differently. In this study, I explored how participants rate lecturers for *the same* teaching behaviours.

In sum, the main aim of this study was to examine how students interpret and use answer scale categories during completing SETs. Additionally, I aimed to explore whether participants would rate male and female lecturers differently if teaching behaviours described in the scenarios are the same. I also studied whether students' interpretations of answer categories varied depending on a different direction of standardised rating scales.

6.2 Method

6.2.1 Design

I used a multi-method 2x2 design (male versus female lecturers and two different directions of scale), in which qualitative and quantitative methods were both concurrently applied. The design of this study involved online surveys and four scenarios, two in each of the four conditions (male versus female lecturers and two different directions of scale) about fictitious lecturers. Student participants rated these lecturers on five pre-determined SET item statements using multistage answer scale categories, indicating their level of agreement with item statement.

In this between-participants design, participants read two different scenarios to achieve diversity of responses. This design had two independent variables with two levels, a) the scenario lecturer's gender (man or woman), b) the direction of the answer scale categories in the rating scales (starting with "strongly agree" or "strongly disagree"). Five dependent variables were the participants' ratings of lecturers for on each of the five chosen SET items.

Table 9 shows the overview of the design and each possible condition. I included different scenarios to elicit a wider range of responses to teaching behaviours described in

scenarios, but I did not analyse differences between types of scenarios that were only used to diversify teaching situations. All scenarios included a mixture of positive, negative and ambiguous teaching behaviours (rather than only positive or negative).

Unlike Study 1, in which participants rated specific lecturers they encountered in real life, participants in this study were (if in the same condition) all considering the same situation as described in the scenarios. Therefore, any potential gender biases or stereotypes may have been easier to identify.

I used the responses from the open-ended questions about participants' reasons for selecting specific categories to scrutinise how students use and interpret these scale categories and form their overall judgements.

Participants' answers to an open-ended question about whether they would judge male and female lecturers differently enabled me to analyse their comments for potential stereotypes regarding lecturers' gender. Participants consciously reflected on whether they may judge male and female lecturers differently. To avoid participants guessing that one of the purposes of the study was to explore gender-related beliefs, this was the last question presented only after they already provided their ratings.

Table 9

Overview of the Design and Conditions in Study 2

	Female lecturers		**Male lecturers**	
The direction of the scale				
Strongly disagree -> strongly agree	Scenario A&B	Scenarios A'&B' (Reversed scenarios A&B)	Scenario A&B	Scenarios A'&B' (Reversed scenarios A&B)
Strongly agree -> strongly disagree	Scenario A&B	Scenarios A'&B' (Reversed scenarios A&B)	Scenario A&B	Scenarios A'&B' (Reversed scenarios A&B)

Note. The table depicts possible conditions. Each participant was only exposed to one of the conditions (e.g., reading scenarios A&B, then rating female lecturers, with the scale direction starting from strongly disagree to strongly agree).

6.2.2 Participants

Out of 336 participants who took part, 226 identified as women, 102 as men, and eight participants identified differently. All participants were 18 years or above, with an age range between 18 and 63 years (M = 24.60, SD = 5.96). They were either enrolled as students at a U.K. university *or* had finished their studies within the last twelve months. Participants were either U.K. students (229), E.U students (59) or International students (48). There was a similar distribution of postgraduate students (166) and undergraduates (170). Participants came from different academic backgrounds, with Social Sciences, Business and Law and Psychology being the most common (*see Table 10*).

In March 2020, the COVID-19 global pandemic spread into the U.K., leading to a series of national lockdowns and a transition to online learning. My recruitment begun in April 2021, shortly after the last national lockdown and introducing easing of restrictions. I used opportunity sampling combined with snowball sampling (*see Chapter 4 for rationale*). I recruited participants through the SONA system at the University of Greenwich. However, to improve the representativeness of the sample, I also advertised this study on social media and websites used for this purpose (SurveyCircle, SurveySwap). Student participants were, therefore, recruited from various institutions across U.K. Because students from different parts of the U.K. may have different tuition arrangements (*see Chapter 2*), it cannot be determined whether all student participants paid tuition fees for their degrees. Furthermore, other differences between universities may influence student evaluations of teaching. For example, students attending institutions with the highest ranking might have higher expectations than students at lower-ranked universities. Similarly, universities in the cities could have a more diverse student population than rural universities. This could lead to a higher variation of what students expect of their lecturers or how they interpret teaching quality, for example, depending on their economic class or cultural background. Although these factors may have shaped participants' responses, I did not request specific information for ethical reasons, as these data may be seen as sensitive and were not crucial to my research.

Table 10

Percentage of Participants' Academic backgrounds (N = 336) Organised by Their Gender

Subject	Male students		Female students		Non-binary students	
	%	N	%	N	%	N
Art/Design	0.9	3	2.1	7	0.3	1
Business and Law	8.9	30	14.6	49	0	0
Computer Science	2.1	7	1.2	4	0	0
Engineering & technology	1.2	4	0.3	1	0	0
Humanities	0.9	3	1.2	4	0.3	1
Natural Sciences	3.3	11	3.3	11	0.6	2
Psychology	3.9	13	17.0	57	0	0
Social Sciences	5.4	18	21.1	71	0.6	2
Subjects Allied to Medicine	2.4	8	2.4	8	0.3	1
Other	1.5	5	4.2	14	0.3	1

Note. For an easier comparison between my two studies, I retained the original labels and the guidelines I used to create subject clusters in Study 1.

6.2.3 Materials

Participants in each condition read two different scenarios about fictitious lecturers' behaviours (either Scenarios A&B or Scenarios A'&B'). Scenarios A & B (or A' & B') presented to participants were identical in each condition and differed only by a lecturer's name, which reflected either male or female gender.

To capture a broad range of behaviours and beliefs and for realistic scenarios of lecturers showing a mixture of behaviours considered by student participants both positively and negatively, these scenarios described lecturers demonstrating particular constellations of both positive and negative teaching behaviours frequently reported by participants in Study 1. To make the situations more realistic in regards to participants' explanations, obtain more varied ratings and explore whether participants interpreted or used answer categories differently based on a lecturer's gender, I also included ambiguous

teaching behaviours. Descriptions of teaching behaviours in Scenarios A&B were changed to create "reversed" Scenarios A'&B' in which lecturers demonstrated corresponding teaching behaviours (e.g., 'Lack of visual elements' changed to 'Many visual elements' (*see Tables 11, 12*). My aim was to examine each of these behaviours in different (negative, positive, ambiguous) contexts.

I chose these behaviours based on two factors: teaching behaviours that: a) related to the specific items on which participants rated lecturers in each scenario, and b) participants frequently mentioned in Study 1. For example, for a scenario in which participants rated the "organisation" of the module, I included teaching behaviours related to 'Content', 'Structure', and 'Time management', which were frequently identified themes when participants rated the "organisation" in Study 1 (*see Tables 11, 12*).

To avoid using potentially unisex names for the lecturers, I selected names only if over 99.9% of people with this name were women (Emily, Sara) or men (Jacob, Nathan; (HowManyofMe, 2021). All four chosen names were popular between the 1970s and 2000s, reaching the top 100 of the most popular names in each of these decades (Social Security Administration, 2021). I chose the first names from the publicly available U.S. database. In terms of surnames, I selected "Taylor" (Dr Emily/Jacob Taylor) and "Jones" (Dr Sara/Nathan Jones, which are in the top three of the most popular surnames in the U.K.

I used the following five SET items derived from the SET questions used by the University of Greenwich: [28]

1.) The lecturer makes the subject interesting (the "engagement" item)

2.) The way their module is taught helps me to learn (the "effective teaching" item)

3.) This lecturer provides good support to manage my assessment workload (the "support with assessments" item)

4.) Their module is well-organised (the "organisation" item)

5.) Their feedback helps me develop and improve my performance (the "feedback"

[28] Similarly to Study 1 and although these items are frequently used in SETs, I included only a few selected items and not the full scale. The predominant aim is to scrutinise students' use and interpretations of answer scale categories with qualitative methods rather than focus on the differences between lecturers' ratings (see also Methodology chapter)

item)[29]

I asked participants to rate the lecturers from Scenario A (A') on item statements about "engagement", "effective teaching" and "support with assessments", and the lecturers from Scenario B (B') on the "organisation" and "feedback" items. All participants rated two lecturers from different two scenarios on all five item statements. The difference consisted in the types of scenarios (A&B vs. A'&B').

Table 11

Teaching Behaviours Depicted in Scenarios A&A'

Teaching behaviours	**Scenario A**	**Scenario A'**
Engagement (delivery, humour, slides)	Steadily reading slides	Lively voice, humorous remarks
	Lack of visual elements, long text	Many visual elements
	A few video clips played	No video clips played
	Lack of elaboration	Sufficient elaboration
Supporting students (help with exam preparation)	Questions to test knowledge & exam preparation	No questions to test knowledge & no exam preparation
Teaching methods & skills (explanation, clarity of the lecture)	Provides clear explanation	The explanation seems quite unclear
Supporting students (answers to students' questions, email communication)	Directs student to the learning platform for further info	Clarifies answers about assignments in more detail

[29] This item was originally worded as "Feedback helps me develop and improve my performance." However, student participants seemed to comment on helpfulness of feedback in general. Therefore, I reworded the item statement (by adding "their": I chose this word because SET items usually use gender-neutral language) to clarify that this is about feedback from the lecturer in the scenario.

Table 12

Teaching Behaviours Depicted in Scenarios B&B'

Teaching behaviours	Scenario B	Scenario B'
Organisation (structure, content, time management)	Structured slides	Unclear and slightly vague slides
	Detailed and relevant but slightly dry content	Interesting but not fully applicable content
	Two reference links are missing	All reference links are provided
	Finishes five minutes late	Finishes on time
	Online resources easy to find, with links to the library and information for exam prep	Online resources difficult to find, with no links to the library nor information for exam prep
	Difficulty retrieving articles outside the reference list	Articles from the reference list retrieved easily
Feedback (detail, consistency, suggestions for improvement)	Short, vague feedback, only partial explanation for mark	Detailed, specific feedback, fully explains the mark
	Some of the mistakes contradict the guidelines	Consistent with the guidelines
	Three very useful suggestions provided	One useful suggestion provided

6.2.4 Procedure

All participants read two scenarios, both featuring a different lecturer of the same gender (either two men or two women, which was indicated by name and the pronouns used to describe the lecturer) to avoid inferences on the gender focus in this study.

Participants read either Scenarios A & B (N = 167) or Scenarios A' & B' (N = 169). All participants rated the two lecturers on five SET items overall, choosing one of five answer scale categories. The direction of the scale started either from "strongly disagree" or "strongly agree" and was randomised between participants but stayed consistent for each participant throughout the whole procedure.

After each rating, participants answered open-ended questions to justify their choice of a specific answer scale category instead of another (e.g., selecting “agree” instead of “neither agree nor disagree” or “strongly agree”)

Afterwards, I asked them whether they would generally judge male and female lecturers differently and to explain the reason for their answer (“Do you think that you would judge male and female lecturers differently? Why or why not? Please describe briefly.”)

Participants provided their demographic information: their own gender, study subject, undergraduate or postgraduate study. They also indicated whether they classified as U.K., E.U. or International students. All participants could provide feedback about the study.

6.2.5 Data analysis

To explore reasons that student participants constructed for picking answer categories when rating lecturers on five items, I analysed all participants’ justifications for choosing specific answer categories with latent, predominantly deductive thematic analysis (TA). I also followed most principles outlined by Robinson (2022) that focus specifically on analysing brief texts usually provided by larger samples.

When I analysed participants’ justifications for picking the answer categories, I coded to answer my specific research questions instead of letting specific research questions evolve through my coding process. This type of deductive approach is more suitable for focusing on a particular aspect of data (Braun & Clarke, 2006). My coding was therefore guided by the themes detected in Study 1, but I also identified a few original themes when I noticed a pattern in data that captured important information.

However, when I coded participants’ answers to an open-ended question about their judgements of lecturers who were either men or women, I used an inductive (data-driven) approach. I chose this approach because I aimed for broader data analysis and tried to minimise reflecting my personal beliefs about the subject on data for this one question (Braun & Clarke, 2006; Kiger & Varpio, 2020).

Because I was searching for the underlying meanings behind the participants’ words, I used a latent approach (Braun & Clarke, 2006). Participants frequently wrote partially vague or unclear comments or provided limited context. I had to look beyond the meanings on the surface and aim to capture ideas underlying data. Interpretation, therefore, formed a crucial part of the coding process. I analysed participants’

justifications for the chosen particular answer categories for each item separately because each item related to a different context. The meanings of the identified themes or categorisation of the themes and sub-themes could subtly differ between items.

Additionally, to analyse whether the lecturers' gender and the direction of the rating scales affected SET scores on five item statements, I conducted two One-Way MANOVAs.

6.2.6 Special ethical considerations

Participants could feel stressed after reading about the negative behaviours of lecturers. However, scenarios only described everyday teaching situations unlikely to cause stress, and participants were informed that they could withdraw at any time[30].

After participants provided ratings and their reasons for choosing answer categories, I asked them whether they would judge their male and female lecturers differently, which could potentially upset some participants. Participants who were not comfortable answering this question could proceed without providing any answer.

One participant named a specific module at their university. This information was censored and deleted from data to protect anonymity of lecturers.

6.3 Findings and discussion

6.3.1 RQ3: Which teaching behaviours and inferred teaching attitudes do students consider in their SET ratings?

I analysed the justifications that participants provided for choosing answer categories across all items and scenarios, and conditions. This enabled me to explore types of reasons for picking the answer categories that student participants generally constructed across all answer categories. Exploring these types of reasons enabled me to identify teaching behaviours and attitudes that participants considered when rating target lecturers described in scenarios.

6.3.1.1 General overview and patterns

Participants mostly justified their choices by commenting on teaching behaviours depicted in the scenario or teaching attitudes they inferred from these behaviours. I categorised justifications (teaching behaviours, inferred attitudes or other types of reasons) that participants as a sample provided into seven to fourteen types of reasons per

[30] I outline the general ethical considerations in Chapter 4.

rated item (*M* = 10.80, *SD* = 3.27). Participants constructed a wide range of reasons for picking the same answer category. Each participant, however, mentioned on average just between one and two (*M* = 1.40, *SD* = 0.92) reasons for choosing the category per item. This indicates individual differences in how participants interpreted and used answer categories and teaching behaviours they considered (*see also 6.3.2., RQ5*).

The most frequent types of reasons outlining teaching behaviours or attitudes concerned 'Lecturer did/did not explain the topic well-used clear teaching style' (11.5% of content mentioned by participants was related to this type of reasons), 'Lecturer did/did not elaborate in detail' (11.0%), and 'Lecturer did/did not support or show interest in students' (9.0%). All three types of reasons were mentioned in both positive and negative context. Table 13 depicts other frequent types of reasons with content related to teaching behaviours and attitudes that participants considered in their SET ratings.

Participants generally focused more on negative rather than positive aspects (*see Table 13*). The three types of reasons with the highest percentage of negative context concerned 'Lecturer read the slides-long text', 'Lecturer did/did not prepare students for exam- tested-interacted with students' and 'Lecturer did/did not elaborate in detail'.

In contrast, most participants mentioned behaviours that are related to 'Lecturer encouraged independent learning', 'Lecturer provided a variety of methods' and 'Lecturer provided useful content-resources' in positive context. Across all items and scenarios, 53.8% of participants provided negative and 45.4% positive comments[31].

Interestingly, participants frequently mentioned lecturers' attitudes or intentions that they inferred from the behaviours described, especially unsupportive behaviours. For instance, "*[the lecturer] pointed to the online learning platform to get rid of you*" (MOSD18[32]).

Participants also commonly assumed from the lecturer's behaviour that the lecturer wants students to work independently. Specifically, participants believed lecturer acted in certain ways to encourage students to take responsibility for their learning without over

[31] A small part of provided comments was ambiguous. When possible, I categorised them into positive or negative type of reasons, but for 0.8% comments (all in mid-scale category), the context could not be determined and I, therefore, categorised them as neutral.

[32] Indicates label for participant. First letter indicates female (F) or male (M) lecturer, second letter original (O= Scenario A & B) or reversed mirror (M= Scenario A' & B') condition, and the rest the direction of the scale, starting with strongly disagree (SD) or strongly agree (SA).

relying on the guidance of the lecturer, e.g., "*He wanted his student to read the material that he gave hence the attitude*" (MOSD9).

Some participants also commented on their personal rating tendencies or feelings or discussed their personal experiences of teaching. Occasionally, participants provided comments that were quite difficult to interpret, or suggested participants may have not entirely understood the instructions, e.g., "*I agree and disagree but mostly disagree*" (MMSD27) or "*Some tutors have tips to make the lecture more interesting*" (FOSA16).

Table 13

The Overview of the Most Frequent Types of Reasons for Choosing Answer Categories that Described Teaching Behaviours and Inferred Attitudes

THEMES	**Total**		**Positive**		**Negative**		**Examples (Participant quotes)**	
Theme content	**%**	**N**	**%**	**N**	**%**	**N**	**Positive**	**Negative**
Lecturer did/did not:								
Explain the topic well or used clear teaching style	11.5	260	33.8	88	66.2	172	"*because he broke the information down in order for me to understand it fully*" (MOSA19)	"*Confusion over explanations"* (FMSA35)
Elaborate in detail	11.0	250	24.4	61	**75.6**	**189**	"*Because once asked to elaborate she clarified things in more detail*" (FMSD36)	"*She doesn't give you enough information*" (FMSD17)
Support-care about students	9.0	204	28.4	58	71.6	146	"*Gave help on assignments*" (FMSD13)	"*he doesn't help enough*" (MOSA22)
Provide useful-consistent feedback or marking	6.4	146	60.3	88	39.7	58	"*Contradictory marking with the criteria*" (MOSD29)	"*she provides useful and valid feedback on how I could have improved my grade*" (FMSD37)

Prepare students for exam-interacted-asked test questions	6.0	136	14.7	20	**85.3**	**116**	"*gives information about the assessment*" (MOSA11)	"*She did not test our knowledge of the topic and did not give much information on how to prepare for the exam*" (FMSD33)
Provide useful content-resources	5.6	127	**65.4**	**83**	34.6	44	"*provide good information and useful resources*" (MOSD31)	*"[The inconsistencies are annoying] and she should be on top of the materials she is producing*." (FOSD9)
Answer students' questions	**5.5**	125	38.4	48	61.6	77	"*Dr Taylor answered the student [and provided support and guidance]…"* (MOSD13)	"*you should be able to receive comprehensive answers to questions*" (FMSA28)
Provide accurate-clear guidelines	5.1	116	45.7	53	54.3	63	"*the guidelines/information are structured nicely*" (FOSD33)	"*If feedback contradicts the guidelines then the guidelines were not written with much attention to detail*" (MOSD36)
Provide variety of methods	4.1	94	**74.5**	**70**	25.5	24	**"***A variety of different techniques used, e.g. clips, text, visual elements*" (FOSD16)	***"****I perceive using video recordings as a lazy method of teaching in lectures."* (MOSA33)
Ensure easy navigation-access	3.6	81	24.7	20	75.3	61	"*Dr. Sara Jones seems to had put a lot of work in organizing her teaching material…*" (FMSD8)	"*files are difficult to find*" (MMSD27)

Engage students	3.0	68	63.2	43	36.8	25	"*Seems like he taught in an engaging way*" (MOSA20)	*"[The extended explanation later will help with learning] but to just reading of a ppt is not very interesting*" [sic]" (FOSA27)
Lecturer did/did not relate content to real life-applicability	3.0	67	37.3	25	62.7	42	"*Not fully applicable to practice*" (MMSD41)	"*The lecture possibly provided a up to date lecture on current practices that may of benefited students in the future*" [sic] (MMSD42)
Lecturer read the slides-provided long text	2.8	63	4.80	3	**95.2**	**60**	"[*Dr Taylor has provided both visual stimuli and text], even though the text passages are long, the topic is complex so may need to be*." (MOSD18)	"*Reading me the slide doesnt help me teach me the subject*" [sic] (MOSD2)
Lecturer did/did not provide accurate references	2.6	60	28.3	17	71.7	43	"*If the reference list is well managed, It makes the lesson organised*" (MMSD26)	*"[yes but] missing references are annoying; i would email her to get the refernces*" [sic] (FOSD25)
Lecturer encouraged independent learning	2.2	50	**84**	**42**	16.0	8	**"***It encourages independent learning and promotes a level of self confidence***"** (MOSA25)	**"***Because I have to study more by myself at library*" (FMSD42)

Note. The total column indicates the overall frequency of types of reasons that participants provided for choosing answer categories and the overall percentage. The percentages do not add to 100 because I only present the overview of the 15 most frequent types of reasons. The 'negative' and 'positive' indicate percentages of negative or positive

comments for each given type of reasons and will therefore add to 100. In bold are the frequencies of the types of reasons with the highest percentages (three most frequent types of reasons for positive and three for negative content). The most frequent types of reasons overall are described in more detail below.

6.3.1.2 The most frequent types of reasons for choosing particular answer categories

6.3.1.2.1 'Lecturer did (did not) explain the topic well-use a clear teaching style'

This important most frequent type of reasons (11.5% overall) involved all comments about teaching manner, style, and presentation or explanation and clarity of teaching. Out of all comments mentioning this theme, 66.2% were negative and discussed an unclear, insufficient or not provided explanation or simply stated that it could be improved, for instance, "*She doesn't explain her slides [and has to be prompted to give answers]. If a student is left with questions, then the professor hasn't explain it entirely or given them the opportunity to understand*" [sic][33] (FOSA37), "*he is not clear in how he describes things, making it difficult*" (MMSD10).

In contrast, 33.8% of participants appreciated explanation was provided or useful, e.g., "*He gives useful explanation when being asked about assignment*" (MMSD20) and "*She clarifies the topic further, showing that she is useful*" (FMSA6).

6.3.1.2.2 'Lecturer did (did not) elaborate in detail'

Overall, 11% of participants considered ideas related to 'Elaboration in detail' (or lack thereof). This included elaborating or adding detail in terms of content, instructions, responses to students' questions or provided feedback.

Out of participants who mentioned content related to elaboration, 75.6% of participants primarily focused on insufficient elaboration, e.g., "*I disagree because she did not elaborate while explaining the lecture but only did when I asked her to*" (FOSA42), "*…but I think its their responsibility to elaborate further if somebody doesn't understand something*" [sic] (MOSA27).

[33] I kept all participants' quotes in their original form (without correcting spelling or grammar errors) to maintain authenticity.

In contrast, 24.4% participants mentioned 'elaboration' in a positive context, e.g., "*Because once asked to elaborate she clarified things in more detail" (FMSD36) and "he gave extra detail when the student was still unsure*" (MMSD27).

6.3.1.2.3 'Lecturer did (did not) support or show interest in students'

This type of reasons included all behaviours related to supporting students, but also a lecturer's perceived caring or interest in students or their learning. It was the third most frequent type of reasons (9% overall).

Out of participants who mentioned this type of reasons, 71.6% commented on lack of support. Participants especially disliked being directed to the online platform by their lecturers. For instance, "*telling students to go look at the online learning platform is not helpful as the likelihood is they have already looked there and either need additional help or some guidance to where the content is on the page*" (FOSA13), "*The lecturer could help put by guiding instead of just stating module material is available on platform*" [sic] (MOSA40) or "*I do not think a student is supported with their work if just told to read the online learning platform*" (FOSA5).

Participants also reported feeling like they had to put a lot of effort into getting their lecturer's support, e.g., "*Would have to chase her for help, which takes up time and energy and would be stressful for me if i had other assignments to work on*" [sic] (FMSD39) or "*he needs to be chased up multiple time to obtain support*" (MMSD11), "*a lot of effort on the behalf of the student is required to receive sufficiently detailed support*" (FMSD18).

This implies that most participants disliked behaviours that they felt indicated the lecturer's lack of interest or lack of effort in supporting students.

In sum, participants mostly focused on clarity of explanation and sufficient detail. They disapproved of lecturers' behaviours that could be seen as dismissive or unsupportive, such as being directed to the online platform. Participants also complained about having to "chase" lecturers for support.

6.3.1.3 General patterns of teaching behaviours and students' expectations observed in types of reasons that participants provided

This section involves also types of reasons with lower frequencies. I discuss patterns identified in teaching behaviours and inferred attitudes that participants considered when completing SETs as well as participants' expectations of their lecturers.

6.3.1.3.1 A lecturer as an entertainer vs educator

Although only 1.9% of content overall related to the type of reason 'Lecturer used humour', it was together with 'Lecturer provided a variety of methods' the most mentioned behaviour (12.7%) when participants read the scenario depicting the lecturer making humorous remarks and rated the "engagement" item. This theme involved any content mentioning the use of humour or describing lecture as fun.

Humour was always mentioned in a positive context. For example, participants commented, "*humorous remarks can have a significant impact on effectiveness*" (MMSD5), "*the funny comments may help the students stay engaged during the lesson*" (FMSA31), and "*includes witty humour*" (MMSD21). This may imply that many students expect their lecturers to fulfil the role of an entertainer (Wong & Chiu, 2019).

In contrast, 2.8% of the whole sample provided content related to 'Lecturer read the slides-long text'. Interestingly, this was the most frequently mentioned behaviour (18.6%) when participants read about the lecturer steadily reading from the slides and rated lecturers on the "engagement" item. Participants overwhelmingly (95.2%) disliked if the lecturer provided slides with predominantly long text or simply read the slides, e.g., "*Lecture slides with long text passages are boring*" (FOSD19), "*I personally don't find just reading from slides very engaging*" (MOSD6), "*Only reading from the slide and having loads of text on a slide is quite lazy and does not make it very interesting*" (FOSA10).

This suggests students may feel that lecturers are not providing any additional value by simply reading the content that students could read on their own. Students may also infer lack of caring or effort related to teaching from the lecturer from this behaviour.

6.3.1.3.2 Independence of university students

Out of the whole sample, 2.2% of participants also commented that 'Lecturer encouraged independent work or learning'. Independence was seen as students not over relying on their lecturers' support but taking their own responsibility for their learning. Interestingly, opportunity for independent work was often framed as positive, and many participants demonstrated an active approach to learning. For instance, "*Asking extra questions forces you to learn more, and it is not the responsibility of the lecturer to inform you of everything regarding a topic. It is for them to lay the foundations so you can go away and learn more*" (FMSA39), "*Once I have a small understanding I tend to seek my*

own information. Usually, library services or journal databases. At University I don't expect to be given everything, I expect to find a lot of the information for myself" (MMSA19).

Participants frequently mentioned independent work when rating "support with assessments." They acknowledged, "*Assignments are mostly the student's duty*" (FMSA16), "*the encouragement for independent learning helps*" (MMSA22), *"[She's not offering much help towards the assignment] however as a university student I should be expected to interpret information myself*" (FOSD9).

Occasionally, participants defended the lecturers, e.g., "*Uni isn't about being spoon fed*" (FMSA41), *"[he could elaborate on the question] but he is right that he says you should read the material*" (MOSA41), "*She can not spoon feed the students all the answers, student need to be able to discover and interpret the information provided by themselves*" [sic] (FOSA27). These observed patterns are encouraging because they imply that some participants may have reflected on their responsibilities as university students.

Only a minority (16%) of participants framed a need for independent learning in negative context, for example, "*I prefer to have all the information given to me, rather than having to find it myself*" (MOSA4) and [student chose disagree] "*Because I have to study more by myself at library*" (FMSD42).

Other comments were ambiguous, where participants acknowledged that independent learning is an essential part of university study but finished their comments by listing teaching behaviours they disliked. For instance, "*Higher education is more independent than other learning [but students still need a good amount of guidance as the content is new to them]"* (FOSA13).

This shows an interesting contrast in participants' answers. On the one hand, they seemed to dislike spending additional effort to obtain further support or answers to their questions. On the other hand, other participants positioned themselves as active learners and seemed to prefer working independently. This may indicate individual differences in students' expectations in terms of university work and from their lecturers.

Another frequent type of reasons associated with independence concerned 'Lecturer answering students' questions' (5.5% overall), which included any responses to students' questions but also email communication with lecturers. This also relates to independence because although students may need to clarify certain aspects with lecturers, the general expectations are that they should take initiative of looking for answers first (e.g., by checking online platform).

When participants mentioned ‘Lecturer answering students’ questions’. 61.6% of comments were negative. Participants complained about having to “chase” the lecturer for answers, unhelpful or untimely answers, e.g., “*He has told the student to go online and look at the answers instead of giving the student the answer himself*” (MOSA1), “*Would be helpful if information was clear initially without having to ask twice*” (FMSA36), “*She isn’t helpful in answering any questions therefore I would feel ill prepared*” (FMSD2).

Positive comments (38.4%) included lecturers providing the answers to students, such as “*answered the question about exam requirements the way he should have - in a useful way without giving the straight answer*” (MOSD30), “*She gave all the answers to my questions*” (FOSD18), “*Emily has responded to email questions which is helpful*” (FOSA40).

This further supports the idea that many students expect their lecturers to provide help directly rather than looking for answers independently. It also implies individual differences in what students expect from their lecturers (e.g., in terms of support).

6.3.1.3.3 The importance of assignments

Participants also frequently mentioned content related to ‘Lecturer providing accurate-clear guidelines’ (5.1% overall). This included different aspects of guidelines such as clarity and consistency or how useful participants perceived them to be for their assignments. Many participants (45.7%) appreciated being provided with assessment guidelines and praised them for being easy to understand or helpful**,** e.g., “*helpful instructions to help aid us with the exam*” (MOSA26), “*The assessment guidelines were helpful and also updated with improvements*” (FMSA40).

Other participants (54.3%) disliked contradictory, unclear, inaccurate or confusing guidelines. For instance, “*his instructions went against the marking guidelines*” (MOSD29), “*It is unfair to be penalised if the brief was incorrect*” (FOSD21).

Interestingly, most participants commented on guidelines positively when rating target lecturers on the “organisation” but in a negative way when rating on the “feedback” item. This may suggest that students’ expectations for particular teaching behaviours may change across different items. For example, when rating “organisation”, students may appreciate guidelines are simply provided. However, when rating “feedback”, students may focus more on perceived problems with the guidelines, such as lack of consistency.

Another type of reasons related to assignments concerned ‘Lecturer providing improvement suggestions’ (1.9% overall) that lecturers provided in their feedback.

Interestingly, 94.6% of participants perceived lecturers' offered suggestions for improvement in a positive context, usually stating these suggestions were provided or useful, "*He does provide useful suggestions for improvement*" (MOSD4) or *"[While the feedback may be short and vague], it seems like the recommendations would make up for it*" (FOSA17). The rest of the participants reported that suggestions for improvement were insufficient, e.g., "*only one answer was provided in improvement*" (FMSD15). This suggests that students generally may be open to improvement suggestions and value them.

Participants also emphasised the importance of feedback in general. Out of the participants who commented on feedback in general (2.1% overall), majority (97.9%) provided positive comments, stating that feedback is important and useful. Participants appeared to discuss general feedback rather than specific feedback provided by the lecturers in the scenarios, e.g., "*I personally agree, but it do not seem to be happened in the scenario 2*" [sic] (MOSD15). This suggests that the item was worded confusingly, making participants think they should comment on the general helpfulness of feedback rather than specific feedback from the lecturer (*see also Method section).*

Interestingly, participants who commented on content related to 'Lecturer providing useful-consistent feedback-marking', predominantly focused on quality (e.g., clarity, consistency) of the feedback and marking rather than received grades.

In sum, participants highlighted teaching behaviours related to assignments, such as lecturers providing accurate and consistent guidelines, as important. Participants also valued provided feedback and, especially, any improvement suggestions.

6.3.1.3.4 The same teaching behaviours (scenario) interpreted in different ways

Participants as a sample (N =336) interpreted teaching behaviours described in scenarios in distinct ways, the reported teaching behaviours or inferred attitudes could be categorised into seven to fourteen types of reasons per item (M = 10.80, SD =3.27). However, participants considered on average only between one and two different reasons (M = 1.40, SD = 0.92) per item and rater. This suggests individual differences in how participants interpreted teaching behaviours depicted in scenarios.

Surprisingly, participants frequently interpreted even the same teaching behaviours and attitudes considering different contexts. For instance, participants who read Scenario A commented on behaviours categorised into the theme 'Lecturer did/did not explain topic well-used clear teaching style' in 102 instances. Even though participants considered the same situation, "explanation-teaching style" was interpreted in 53.9% of cases as

"unclear, insufficient, badly timed", and 46.1% cases as "good, clear, engaging, provided or useful" (*see Table 14*). Participants said, "*helps students easily understand the more complex areas*" (FOSD13), "*He explains clearly*" (MOSD35) but also "…*she does not offer to explain the coursework properly*…" (FOSD29) and "*if he is not able to explain the lecture, then the student's learning is challenged*" (MOSD15) about the same scenario.

This implies that participants perceived the same described teaching behaviour in various ways and interpreted these behaviours in a distinct context after reading the same scenario. The scenario only provided rather short descriptions of two particular teaching situations. However, in real life, students evaluate lecturers over a long period after observing them in various situations and contexts. Students may also be affected by many other factors, such as age, appearance or ethnicity of the lecturer. Interpretations may, therefore, vary even more in real life. Student participants were also reminded of the scenario each time they rated the lecturer. In real life, students would have to rely on their memories of the situations and may be influenced by ideas they develop about their lecturers over time.

Table 14

A Range of Teaching Behaviours and Attitudes Mentioned in Types of Reasons that Participants Provided for Choosing Answer Categories (N = 167) when Reading the Same Scenario and Rating the "Engagement" Item

Theme content	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Lecturer read the slides-provided long text (42)	Did (3), neg. context	Did (28), neg. context	Did (7), neg. context	Did (4), 1 in pos. context	
Lecturer did/did not provide variety of methods (34)	Did not (1)	Did (7), 2 in neg. context Did not (2)	Did (4), 1 ambiguous	Did (19)	Did (1)

Lecturer did/did not explain the topic well-use clear teaching style (31)	Did not (5)	Did (1) Did not (8)	Did (2) Did not (5)	Did (9)	Did (1)
Lecturer did/did not support students (23)	Did not (4)	Did (1) Did not (3)	Did (3) Did not (3)	Did (2) Did not (7)	
Lecturer did/did not elaborate in detail (21)		Did (2) Did not (15)	Did not (3)	Did not (1)	
Lecturer did/did not engage students (18)	Did not (2)	Did not (4)	Did (4) Did not (3)	Did (3) Did not (1)	Did (1)
Lecturer did/did not prepare students for exam-interact-ask test questions (16)		Did (1) Did not (13)		Did not (1)	Did (1)
Students described their own feelings-preferences-perceptions (11)		(5)	(6), neutral context		
Lecturer did/did not provide professional examples (10)		Did not (3)	Did (3)	Did (3)	Did (1)
Lecturer did/did not show caring-effort-enthusiasm (9)	Did not (3)	Did (1) Did not (5)			

Lecturer provided average-adequate lesson-standard (8)	(1)	(5), neutral context	(2)

Note. In bold (theme content) are the summary abbreviations of teaching behaviours and attitudes that participants mentioned when they provided reasons for picking answer categories. The numbers indicate their frequency. Negative context is indicated with blue colour and positive context with red colours. Participants frequently disagreed on context for the same behaviour, therefore, some types of reasons indicate how many participants perceived the lecturer demonstrating the behaviour and how many perceived the lack of it (did/did not). Occasionally, the predominantly positive type of reasons involved also a few negative comments (or vice-versa). For example, the 'Variety of methods' was provided by a lecturer, but some participants saw it in a negative context, or participant did not mind the 'Lecturer reading slides'. This is indicated in additional brackets with 'neg.' or 'pos.' context. Comments with no colour are neutral, such as describing lecture standard as 'average' or participants stating their preferences or feelings (e.g., stating more information is needed to judge).

6.3.1.4 Brief discussion

In sum, the main teaching behaviours and attitudes that participants considered were linked to how clearly lecturers explained the topic, whether they elaborated in detail and supported students, as well as if they provided a high quality of feedback and marking. Participants considered how well lecturers prepared them for assignments, answered their questions and the quality of guidelines and content or resources that lecturers provided.

Participants frequently inferred their lecturers' attitudes from supportive or engaging teaching behaviours described in the scenarios or any perceived opportunities for independent work (or lack thereof) often by considering different contexts.

The COVID-19 pandemic occurred predominantly between Study 1 and Study 2, with its last stages during data collection for Study 2 that I started in April 2021. Therefore, all student participants experienced a shift to online learning due to the national lockdowns before participating in this study and some participants may have still not

transitioned back into in-person teaching. This could have potentially shaped participants' responses and expectations they had of their lecturers or their understanding of teaching quality. Interestingly, no participant reflected on the lockdowns or the COVID-19 pandemic, although a few participants described their (positive or negative) stances on online learning. Because the scenarios involved fictitious lecturers, student participants perhaps did not reflect on real-life contextual factors, such as transitions to online learning. Alternatively, participants may have experienced no deterioration in teaching quality during the COVID-19 pandemic. For example, 34 interviews with tutors and educational leaders or mentors in England revealed they responded creatively to challenges during the pandemic and used them to develop resilience as well as enhance and diversify their teaching practices (Towers et al., 2023).

The sample as a whole reported a broad range of behaviours to justify their choices of answer categories. However, each participant considered on average only between one and two reasons, indicating pronounced individual differences in the ways in which students interpret and use the answer scale categories. Similarly, these findings indicate individual differences in participants' expectations (e.g., some praised the opportunity for independent work, whereas others disliked it).

This suggests that participants perceived even the same situation in subjective ways and considered very different behaviours and attitudes and saw them in varying context. Specifically, participants constructed very different meanings for the same situation, describing the same study phenomena. In real-life SETs, this situation may be even more pronounced because students need to rely on their memories and remember lecturer's behaviours across different times and contexts. These subjective perceptions can lead to the potential influence of gender biases and stereotypes on students' ratings.

6.3.2 *RQ5: In what ways do students interpret and use the answer categories of SET scales when completing teaching evaluations?*

In the previous section, I analysed participants' justifications for choosing answer categories to explore what teaching behaviours and attitudes they considered salient. In this section, I analyse general patterns in how participants interpret and use these answer scale categories. I also explore the types of reasons that participants constructed for each *specific answer scale category* in the sample.

6.3.2.1 General overview

Deductive thematic analysis of key reasons that 336 participants provided for choosing answer categories yielded five key themes describing types of reasons participants provided for choosing answer categories. Participants provided together 2352 reasons (excluding miscellaneous reasons not categorised into types of reasons) that I categorised into 28 themes across all five items (*see Table 15*).

Overall, justifications for male and female lecturers were of similar frequency. Participants provided slightly more negative justifications for female (52%) compared to male (48%) lecturers, and more positive justifications for male (52.7%) than female lecturers (47.3%).

The frequency of reasons was also similar for the different directions of the scales, for both negative (48.6% for starting with "strongly disagree", and 51.3% for "strongly agree"), and positive (49.1% for starting with "strongly disagree", and 50.9% for "strongly agree) justifications. This suggests that participants were likely not influenced by priming or anchoring effect, therefore, apart from investigating whether a direction of scale affected students' ratings, I did not conduct further analyses. Overall, participants did not rate lecturers differently across all five items when using the scales with different directions, F (5, 327) = 1.80, p =.112, Pillais' Trace = .027, with effect size ηp2 =.027 (*see Table B1, Appendix for descriptive statistics*).

Table 15

The Overview of the Ten Most Frequent Types of Reasons Identified Across All Scenarios and All Items for Each Answer Category

THEMES: Q1-Q5	**Strongly disagree**		**Disagree**		**Neither agree nor disagree**		**Agree**		**Strongly agree**	
Theme content	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
Lecturer did/did not:										
Explain the topic well or used clear teaching style (260)	**12.6**	19	**13.3**	86 (9+)	**14.9**	65 (20+)	**9.7**	78 (47+)	5.2	12+

Elaborate in detail (250)	**11.3**	17	**14.2**	92 (2+)	**10.1**	44	**8.5**	68 (30+)	**12.5**	29+
Support-show interest in-care about students (204)	**21.2**	32	**12.8**	83 (7+)	**7.4**	32 (14+)	5.6	45 (25+)	5.2	12+
Provide useful-consistent feedback or marking (146)	5.3	8 (1+)	4.3	28 (4+)	5.5	24 (10+)	**7.7**	62 (50+)	**10.3**	24 (23+)
Prepare students for exam-interacted -asked test questions (136)	9.3	14	10.2	66 (1+)	4.6	20 (5+)	4.2	34 (12+)	0.9	2+
Provide useful content-resources (127)	4.0	6	3.9	25 (4+)	6.4	28(16+)	6.2	50 (45+)	**7.8**	18+
Answer students' questions (125)	8.6	13	6.0	39 (3+)	6.7	29 (15+)	4.9	39 (26+)	2.2	5 (4+)
Provide accurate-clear guidelines (116)	6.0	9	3.5	23 (2+)	6.7	29 (12+)	5.6	45 (29+)	4.3	10
Provide variety of methods (94)	2.6	4	2.6	17 (9+)	2.3	10 (6+)	6.9	55 (47+)	3.4	8+
Ensure easy navigation -access (81)	2.6	4	5.6	36 (2+)	3.0	13 (5+)	2.7	22 (8+)	2.6	6 (5+)

Note. In the bold are the ten most frequent types of reasons for choosing answer categories describing teaching behaviours and attitudes. A sign + indicates a number of positive comments. Percentages depict the frequency of types of reasons within each answer

category. Because I only included ten most frequent types of reasons, percentages do not always add to 100%.

6.3.2.1.1 Individual differences between participants when providing reasons for choosing answer categories

At first, I explored how many different types of reasons participants provided and analysed general patterns in how participants chose answer categories. Each justification that I had categorised into a theme (*see 6.3.1, RQ3*) was considered to be a theme of reasons (e.g., if a participant mentioned behaviours I had categorised into the 'Explanation-teaching style' and 'Elaboration-detail' as a justification for selecting an answer category that would count as two themes of reasons).These themes of reasons could be specific teaching behaviours, inferred attitudes, participants' feelings, or any other reasons. Miscellaneous justifications not categorised into any types of reasons were not included in this analysis.

Participants as a whole sample (N =336) provided from seven to fourteen different types of reasons per item statement. Each individual participant provided between zero to seven types of reasons, with each participant considering on average between one and two different reasons (M = 1.40, SD = 0.92) per item and rater. This suggests that most raters only focused on one or two key reasons rather than considering or averaging all pieces of evidence instead of considering the whole field of types of reasons. This is consistent with the previous findings (Uher, 2018a).

Exploring general patterns in ratings revealed that when participants rated their lecturers on the "engagement" item, a majority (40%) considered only one main theme reason (M = 1.58, SD = 0.93), whereas approximately a third of participants (29.5%) weighted or averaged several pieces of evidence. Even fewer participants (23.8%) considered multiple (two or more) reasons (*see Table 16*). The rest of participants either did not provide any justifications or their justifications were too infrequent to be categorised into any themes.

I assumed that participants weighted or averaged the evidence if I could clearly infer this from their justification. Weighting involved cases when participants specifically mentioned why they chose a specific category instead of another one (e.g., 'I would choose strongly agree if [specific elements were better]'). Averaging involved cases when participants considered several pieces of evidence, positive and negative and chose the mid-category, e.g., "*Although she is enthusiastic about the topic, not much information is*

provided and explanations seem unclear" (FMSD11). Participants who considered two or reasons but simply listed them without providing the evidence of weighting are categorised in 'Considered several reasons'. However, some participants may have applied this reasoning mentally without writing it down. This could occur if participants perceived completing this study as cognitively taxing (Dunegan & Hrivnak, 2003) or not exerted all their effort into completing it (Uijtdehaage & O'Neal, 2015). To avoid cognitive overload, participants may have also relied on heuristics (mental shortcuts) to make a quick automatic judgement (Kahneman, 2003; Merritt, 2008).

Table 16

General Patterns in Participants' Ratings on the "Engagement" Item (N =336)

	Strongly disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
	(n= 19)	**(n=105)**	**(n=58)**	**(n=140)**	**(n=14)**
Participants					
Considered one reason	(12)	(46)	(17)	(46)	(6)
Considered several reasons	(6)	(31)	(2)	(34)	(7)
Weighted or averaged evidence		**(21)**	**(31)**	**(46)**	**(1)**
Uncategorised	(1)	(5)	(7)	(8)	

I then looked beyond general patterns and considered specific meanings that participants constructed for answer scale category (across whole dataset). Participants' reasons for choosing the same category frequently overlapped (e.g., participants provided different reasons to justify choosing "disagree"). This implies that the answer categories have no disjunctive meanings as would be required for converting them into numbers.

In 96% of cases, participants (as a group) who read the same scenario, but with a different gender of the lecturer protagonist and different direction of the scale provided at least two different reasons for using the same answer scale category. The remaining 4% of

cases were non-applicable[34] Participants who rated the lecturer of the *same* gender in the *same* scenario with the *same* scale direction almost always provided at least two or more different justifications (as a group). In the whole dataset (*N* = 336), the vast majority of participants reading *the same* scenario provided different justifications for selecting the same answer category if rating the lecturer in the same condition (gender, scenario, scale direction). In 89% of cases, participants as a group interpreted answer categories differently. The remaining 9.5% involved non-applicable cases.

These findings therefore suggest that students' interpretations and use of answer categories may become broader in more distinct contexts (e.g., a different gender of a lecturer; *see also Table 15 for the field of reasons for all items, Table 17 for the different reasons for picking the same answer categories for the "engagement" item*).

Table 17

Individual Differences in Participants' (N = 336) Interpretations and Use of Answer Categories when Rating Engagement

	Strongly disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
Lecturer did/did not:					
Provide variety of methods (71)	Did not (2)	Did (9), 2 in neg. context Did not (5)	Did (5), 1 ambiguous	Did (37) Did not (10), 1 in pos. context	Did (3)
Explain the topic well-use clear teaching style (64)	Did not (7)	Did (1) Did not (19)	Did (3) Did not (11)	Did (12) Did not (9)	Did (2)

[34] For example, if only one participant within the condition chose the particular category or only one participant provided comments categorised into these types of reasons.

Support-care about students (45)	Did not (5)	Did (1) Did not (7)	Did (4) Did not (4)	Did (6) Did not (16)	Did (2)
Elaborate in detail (43)		Did not (1)	Did (2) Did not (21)	Did not (6)	Did (4) Did not (9)
Read the slides-provided long text (42)	Did (3)	Did (28)	Did (7)	Did (4), 1 in pos. context	
Prepare students for exam-interact-ask test questions (38)	Did not (2)	Did (1) Did not (20)	Did not (3)	Did (2) Did not (9)	Did (1)
Show caring-effort-enthusiasm in general (38)	Did not (3)	Did (2) Did not (6)	Did (4) Did not (1)	Did (17)	Did (5)
Engage students (37)	Did not (3)	Did-could (1) Did not (6)	Did-could (7) Did not (3)	Did-could (15) Did not (1)	Did-could (1)
Use humour (35)		Did (1)	Did (3)	Did (24)	Did (7)
Provide professional examples (32)		Did (1) Did not (3)	Did (3)	Did (21)	Did (4)
Use lively voice-manner (21)		Did (1)	Did (1)	Did (15)	Did (4)
Students described their own feelings-preferences-perceptions (17)		(6)	(10), neutral context	(1)	
Answer students' questions (11)	Did not (1)			Did (3) Did not (6)	Did not (1)
Provide average-adequate		Did (1)	Did (5), neutral context	Did (2)	

lesson-standard (8)

Note. In the bold are the types of reasons student participants provided as justification for ticking answer boxes. The numbers indicate the frequency of these reasons. Negative context is indicated with blue colour and positive context with red colours. Comments with no colour are ambiguous or neutral.

Table 17 shows that participants reported a whole range of different types of reasons for choosing the same answer categories. For example, participants who ticked a middle answer category (here 'Neither agree nor disagree') provided 13 distinct types of reasons for choosing this category when rating the "engagement" item. Furthermore, participants mentioned these types of reasons in positive as well as negative and ambiguous context or used mid-category as non-applicable option.

Vice versa, participants frequently used the same reasons for picking distinct answer scale categories. For instance, participants justified choosing four distinct answer categories (ranging from 'disagree' to 'strongly agree') with the 'lecturer using humour'.

This suggests individual differences in how participants encode, interpret, and use answer scale categories. Participants also constructed meanings that were frequently qualitative rather than quantitative and overlapped between different answer categories.

However, psychologists commonly assume raters interpret and use answer scale categories in standardised ways. Psychologists attempt to create quantitative meaning by recoding answer categories into numerals. On this five-point scale, the mid-scale category would often be converted to a numeral '3'. But this mistakenly creates an assumption that this numerical value may reflect homogenous meaning (Uher, 2022a), even though participants constructed a wide range of type of reasons for choosing this category, and similarly, used the same types of reasons for different categories. This indicates overlap between category meanings and indicates it is erroneous to assign quantitative meanings to numerical scores of recoded answer categories.

Furthermore, numerals are frequently interpreted as numbers. But the answer categories recoded into numerals do not have the same quantitative properties as numbers (*see also Chapter 3*). For example, some participants who chose 'disagree' considered several types of evidence, e.g., "*she is quite vague in her explanations and does not provide enough exam/assignment support*" (FMSD37). In contrast, participants who chose

mid-category stated lack of evidence, e.g., "*There is not much information about support for workload in the above description, so it is difficult to make a judgement*." (FMSD20). Psychologists would convert these categories into numerals '2' (disagree) and '3' (neither agree nor disagree). If researchers interpret these numerals as numbers, '3' would indicate a higher quantity than '2'. However, as seen in the example above, these quantitative relations do not apply to relations between verbal categories recoded into these numerals. In fact, a participant who chose "disagree" considered more pieces of evidence than participant who chose "neither agree nor disagree". Not only these participants did not consider different quantities of the same quality, but one also actually considered a different quality.

This suggests that relations between the recoded answer categories do not adequately represent relations between the quantities of the encoded empirical relational system (Uher, 2018b). For example, participants encoded different empirical information (e.g., 'unclear explanation', 'lack of support') into the same verbal answer scale categories (e.g., 'disagree'), recoded by researchers into one numerical value (e.g., "2"). However, the same numerical symbol must always represent the same quantity of the same study phenomena to qualify as measurement. In this example, participants did not even consider the same phenomena, let alone the same quantity.

6.3.2.1.2 Using the same reasons for selecting different answer categories

Participants frequently provided the same reasons even if they selected different answer categories. Nearly all identified themes feature as a reason for choosing all answer categories (*see Table 20*). The only difference is the context. As expected, when participants rated target lecturers on a scale starting with "strongly disagree", the justifications for choosing the higher ends of the scale are more likely to be positive, whereas the opposite applies to the lower ends of the scale (and the contrary if participants rate on a scale starting with "strongly agree"). This especially applies to the "strongly disagree" and "strongly agree" categories. Only $N = 1$ (0.7%) reason for choosing "strongly disagree" was positive and $N = 5$ (2.2%) reasons for selecting "strongly agree" negative. The remaining three answer scale categories are more balanced, with participants providing a mixture of positive and negative comments (*see Table 20*). This results in one–to–many assignment relations with answer scale categories (Uher, 2023). But this means that created values (e.g., here numerals '2' and '4' corresponding to answer categories 'disagree' and 'agree') may represent the same information. For

example, participants provided almost identical reasons, "*The explanation could have been clearer in the first place*" (FMSD38), "*His explanations are unclear and vague*" (MMSD20), and "*unclear explanations*" (FMSD21), to explain their choices of three answer scale categories, ranging from 'disagree' to 'agree'). However, 2 and 4 recoded as numerals and interpreted as numbers indicate different quantitative meanings. This, however, contradicts the interrelations between raters' use of the answer scale categories.

Therefore, participants encoded the same empirical information into different verbal answer scale categories that researchers then recoded into numerical values, or, vice versa, encoded different empirical information to the same verbal answer categories. This makes it impossible to trace information back to the study phenomena and the information that participants wanted to convey.

Importantly, to qualify as measurement, the same quantities of the same study phenomena must always be encoded with the same numerical symbols. But in the example discussed above, numerals '2', '3' and '4' do not even represent different qualitative information. Specifically, participants provided the same type of reasons (unclear explanation) for picking three different categories. Exploring their answers show no evidence that participants considered different quantities of the same quality. Again, interrelations between how participants used these answer categories do not represent interrelations between "2", "3" and "4". It would therefore be erroneous to ascribe them quantitative meanings.

6.3.2.2 Further patterns observed in participants' use of answer categories

6.3.2.2.1 Participants providing ratings and then contradicting themselves

Participants occasionally provided a rating and then contradicted their own rating in their justification of this choice. For instance, some participants only mentioned negative aspects but still provided a relatively "high" rating, such as "agree". Similarly, participants sometimes praised the lecturer but rated them on the given item with "disagree". One participant commented "*seems like a very well organized professor, wish all of mine had been like that*" (FOSA39), but chose "strongly disagree" on the item "the lecturer made the subject interesting". Another participant wrote, "*The long text passages would discourage my interest and he does not elaborate each video clip shortly after they have played. I feel like attention usually drops towards the end of the lecture and choosing to elaborate on the slides at the end would be a dread*" (MOSA26), which seems like a negative comment, but chose "agree" on the statement that the lecturer made subject

interesting. This could be because participants showed different rating tendencies (e.g., to choose higher ends of the scale) or simply made mistakes (*see below*).

Occasionally, participants clarified their choice (e.g., I would have chosen "strongly agree" if [specific aspect] was better). For instance, "*I chose agree because she provides insights, humorous remarks, etc., which makes the topic engaging. However, I would not choose strongly agree because there is a limit to how interesting a topic can be if you are not able to fully comprehend it by having your doubts solved, and she seems unable to do that*." (FMSA20). Some participants may have applied the same reasoning in their mind but not written it down.

Participants may have also changed their minds when asked to think about their choice. For example, one student stated "*I actually changed my answer to I disagree- because her lecture structure needs improvement, she clearly is unaware of that. Also, the fact her marking is confusing and answers are vague suggests she isn't a star teacher*" (FMSA26). Alternatively, participants may have made a mistake and chosen the wrong answer category without realising it (*see below*).

6.3.2.2.2 Obvious mistakes in choosing or justifying choice of answer scale categories

Out of all participants, only 1.79% explicitly stated that they had accidentally selected the wrong answer category. For example, "*I was mean to press strongly disagree. I expect all learning resources to work as intended and the learning to be engaging*" [sic] (MOSD12) or "*it is NOT well designed, but i cant go back to the last question*" [sic] (MOSD37).

Some participants may have also made this mistake without mentioning it, which could explain why some justifications contradicted participants' ratings. I identified 2.38% of these cases. For example, the participant stated: "*He provides detail and relevant materials*" (MOSD41) but ticked "disagree" (that the module is well-organised).

Furthermore, 2.1% of participants explained selecting the answer category by either repeating category, e.g., "agree with it" they chose or simply stating another answer category, e.g., typing "agree" even though they ticked "neither agree nor disagree". In addition, 1.4% of participants provided answers that suggested they did not fully understand the question, e.g., replying "yes" or "no" or repeating a part of the statement "help to learn".

This suggests that some students may misunderstand the instructions, be distracted or rush to complete the study. Importantly, these cases would go unnoticed in formal

SETs, as students may not realise they made a mistake. These mistakes may, however, negatively affect lecturers' ratings.

6.3.2.2.3 Insufficient information

Occasionally, participants also mentioned they did not have sufficient information to judge. Out of all participants, 2.68% mentioned this reason as a justification for choosing at least one answer category. Most of these participants provided this reason for choosing the "neither agree nor disagree" category (72.7%), but also for "disagree" (18.2%) and "agree" (9.1%) categories. For example, participants wrote, "*I felt that there was not enough information provided about her delivery of the subject to know whether or not she made the subject interesting*" (FMSD20), or "*I cannot infer whether his module is structured*" (MMSD32). Most participants who mentioned insufficient information chose mid-scale category. This implies that some students may use mid-category as interchangeable with a non-applicable category.

6.3.2.2.4 Middle category as non-applicable and ambiguous item statements

Similarly, 4.76% of participants commented at least once that they were not sure or did not know why they chose a particular category or stated that the situation did not apply to them, or they did not fully understand the question. For example, "*I don't really know what is meant by the question*" (FMSD11), or "*Does not apply to me*" (FMSA24). Out of these 4.76% of participants, most again provided this reason for choosing "neither agree nor disagree" category (87.5%).

This may imply that students did not know why they made this choice but also that students may have used the middle category as a substitution for "non-applicable" (*see also* Ashby et al., 2011). Researchers may recode this category to indicate an average score (here recoded into a numeral '3'). However, some participants clearly chose this category to indicate 'non-applicable' rather than 'more' agreement than participants who chose category 'disagree' (here recoded into a numeral '2'). Therefore, perceived quantitative relations between the numerical scores of recoded answer categories do not match semantic meanings of these answer categories (Uher, 2022a, 2022b, 2023).

6.3.2.3 Brief discussion

In sum, the entire participant sample constructed a large variety of heterogeneous reasons for picking the answer categories. However, individual student participants

considered on average only between one to two theme reasons from that field of reasons when reporting why they chose specific answer categories in their justifications. Therefore, different participants ticked the same category for distinct reasons, related to different observations and judgements. This precludes establishing an unbroken chain of connections between study phenomena and results, precluding data generation traceability, thus making it impossible to trace back the data to the information to which they refer.

Participants frequently used the same justification even for selecting distinct answer categories. This shows that the meanings of answer categories overlap, and the categories do not have disjunctive meanings.

Fields of meanings are context sensitive. i.e., participants considered different context even for the same scenario. Similarly, there were individual differences in how participants interpreted even the same answer categories. Participants' interpretations became broader in more distinct contexts (e.g., a different lecturer's gender).

Some of the answers and patterns suggested that participants selected the wrong answer category or changed their mind after reflecting further. Other participants mentioned insufficient information to judge, most of them chose mid-category. In fact, mid-category seemed to be used by some participants as "non-applicable" option. This shows some potentially problematic patterns (students making mistakes, students using mid-category as non-applicable) that could also occur in SETs in real life.

This suggests that some reasons for ticking the answer categories do not represent information related to item content. Similarly, results depended on the subjective and flexible ways in which participants used and interpreted answer categories, including any mistakes that participants made (e.g., choosing a wrong answer category). The results, therefore, did not represent the same quantitative meaning across different contexts. The generated values were not connected to any reference quantity as required for public interpretability of quantitative meaning. This precludes numerical traceability and thus establishment of measurement (Uher, 2020, 2022a, 2023).

6.3.3 RQ6: What are possible differences in the teaching behaviours and inferred teaching attitudes that students consider in SET ratings when judging male versus female lecturers?

I explored whether the frequencies of types of reasons for choosing answer categories that participants considered across all items differed for male versus female

lecturers. This enabled me to investigate any potential differences in the teaching behaviours and attitudes that participants considered when judging their male and female lecturers.

6.3.3.1 Overview and interpretation of findings

I examined whether participants commented on particular teaching *behaviours* or attitudes more frequently than expected for male versus female lecturers (*see Table 18*). Specifically, I explored the four most frequent types of reasons regarding whether lecturer did (or did not) demonstrate particular teaching behaviour or inferred attitude: 'Explained the topic well-used clear teaching style", 'Elaborated in detail', 'Supported or cared about students', 'Provided useful-consistent feedback or marking'. I also added two further frequent types of reasons: 'Provided accurate-clear guidelines' and 'Answered students' questions'. These types of reasons were only the seventh and eight most frequent overall, however, they were the next most frequent (after four main ones) when considering only the first three items ('Answers to students' questions') or last two items ('Guidelines). I, therefore, added them to provide a broader general overview across items.

Chi-square analyses revealed that participants commented on good 'Explanation-teaching style' significantly more frequently than expected for their male (14.3%) than for female (8.0%) lecturers, $X^2(1) = 7.895$, $p = .005$ (*see Figure 12*). This suggests that students may appreciate high-quality explanation more from their male versus female lecturers.

There were no gender-related differences for the remaining (91.7%) analysed types of reasons. In sum, participants considered teaching behaviours and attitudes similarly frequently for male and female lecturers, apart from providing high-quality explanation, which they considered more salient for male lecturers. This provides support for stereotype content model, because the good explanation may be associated with competence (frequently perceived as masculine characteristics). Figure 12 depicts how many times student participants praised explanation or teaching when the lecturer was a man versus a woman.

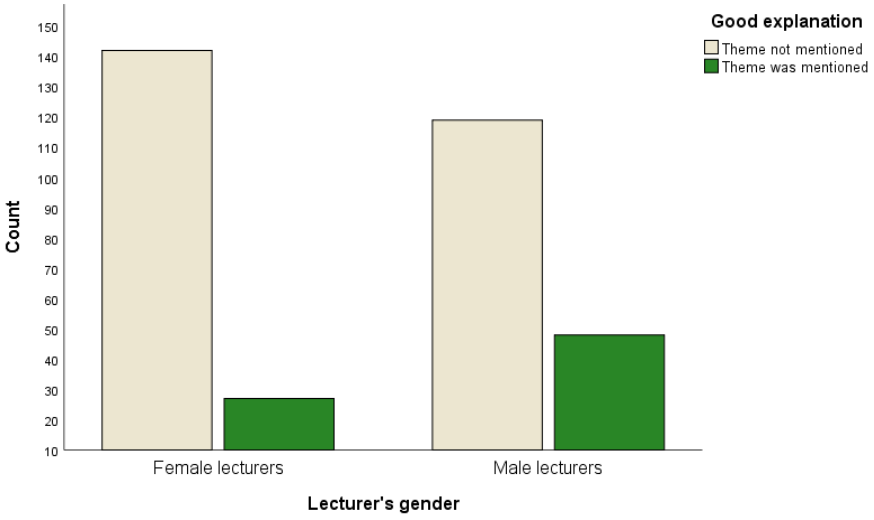

Figure 12

The Frequency with which Participants Mentioned High Quality 'Explanation-teaching style' for Lecturers of Different Gende

Table 18

The Overview of the Frequency of Types of Reasons for Choosing Answer Categories Identified Across All Items Categorised by a Lecturer's Gender

THEMES: Q1-Q5	**Strongly disagree**				**Disagree**				**Neither agree nor disagree**				**Agree**				**Strongly disagree**			
Theme content	**Men**		**Women**		**Men**		**Women**		**Men**		**Women**		**Men**		**Women**		**Men**		**Women**	
Lecturer did/did not:	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
Provide clear explanation-teaching style-lesson	7.3	11	5.3	8	6.8	44 (5+)	6.5	42 (4+)	7.8	34 (14+)	7.1	31 (6+)	5.6	45 (30+)	4.1	33 (18+)	3.0	7+	2.2	5+
Elaborate in detail	5.3	8	6.0	9	6.6	43	7.6	49 (2+)	5.3	23	4.8	21	4.5	36 (16+)	4.0	32 (14+)	6.9	16+	5.6	13+
Support-care about students	11.3	17	9.9	15	5.4	35 (1+)	7.4	48 (6+)	4.1	18 (7+)	3.2	14 (7+)	3.7	30 (15+)	1.9	15 (10+)	2.2	5+	3.0	7+
Provide useful-consistent feedback or marking	2.6	4 (1+)	2.6	4	2.0	13 (2+)	2.3	15 (2+)	3.2	14 (7+)	2.3	10 (3+)	3.4	27 (23+)	4.4	35 (27+)	6.9	16+	3.4	8 (7+)
Prepare students for exam-interacted-asked test questions	4.0	6	5.3	8	3.7	23 (1+)	6.5	42	2.3	10 (3+)	2.3	10 (2+)	2.0	16 (5+)	2.2	18 (7+)	0.9	2+	0	0
Provide useful content-resources	2.0	3	2.0	3	1.9	12 (3+)	2.0	13 (1+)	3.9	17 (10+)	2.5	11 (6+)	4.0	32 (27+)	2.2	18+	5.2	12+	2.6	6+

Answer students' questions	4.6	7	4.0	6	3.1	20	2.9	19 (3+)	2.3	10 (5+)	4.4	19 (10+)	3.1	25 (17+)	1.7	14 (9+)	2.2	5 (4+)	0	0
Provide accurate-clear guidelines	2.6	4	3.3	5	1.4	9	2.2	14 (2+)	1.8	8 (5+)	4.8	21 (7+)	4.1	33+ (21+)	1.5	12 (8+)	2.2	5+	2.2	5+
Provide variety of methods	1.3	2	1.3	2	1.7	11 (5+)	0.9	6 (3+)	0.9	4 (2+)	1.4	6 (4+)	2.9	23 (21+)	4.0	33 (25+)	1.3	3+	2.2	5+
Ensure easy navigation-access	1.3	2	1.3	2	17 (1+)	2.6	2.9	19 (1+)	1.6	7 (2+)	1.4	6 (3+)	1.6	13 (5+)	1.1	9 (3+)	1.7	4+	0.9	2 (1+)

Note. This table shows participants' answers when asked to justify their choice of answer category. It depicts how frequently participants mentioned content categorised (on the left, in bold) for each answer category, for each lecturer's gender. The percentages indicate how many participants mentioned specific content in the answer category. The sign + depicts justifications I interpreted as positive, for example, 4 (1+), indicates four justifications of which, one was positive

6.3.3.1.1 Differences in participants' explicit reports about their judgements of male versus female lecturers

At the end of the survey, I explicitly asked student participants whether they would judge their male and female lecturers differently and to explain why, or why not. Inductive thematic analysis revealed that most participants reported they would not judge their lecturers differently (84.7%), but some would (10.7%), and others were not sure (4.6%).

However, 3.3% of participants who said "yes" or "maybe" mentioned that this would be subconscious or unintentional. For example, "*I'm sure I would since I would certainly be biased according to stereotypes. That doesn't mean I would want to judge them differently, though*" (MMSA6). This shows that some participants in my sample had some awareness of issues related to biases and stereotypes.

In contrast, 7.4% of participants provided reasons that suggested sexism or bias against female or male lecturers. For instance, "*Male teachers [are] usually more simple and helpful*" [sic] (MOSD2), "*Yes, I may find male lecturers more intimidating but female lecturers more annoying [but this doesn't always apply. It depends more on the person]*" (MMSA8). "*I find female lecturers tend to explain better and often provide suggestions or examples*" (FMSA10) or "*Male lecturers [are] more rationale*" [sic] (FMSA1).

Participants often emphasised their personal experience, e.g., "*yes I probably do. Like all women, I have had more experiences of men underestimating my intelligence and patronising me, so I tend to respond differently to male teachers, a bit defensively or like I have something to prove. I feel more guilty when I miss lectures or assignments set/run by female lecturers, for some reason I imagine the are more likely to remember and hold it against me*" [sic] (MMSD35).

The key themes I identified were 'No difference between genders/teaching skills' (20.2%), 'Expectations from the lecturers (same/different)' (12.5%), 'Gender does not matter-I would consider other things' (53%), 'I would/would not due to my own gender' (2.1%), 'Awareness of sexism, bias, discrimination' (6.8%), 'My past experiences with lecturers' (8%).

6.3.4 RQ8: What are the potential differences in the ways in which students use and interpret typical answer categories on SET Likert scales when judging male versus female lecturers?

I explored whether types of positive reasons that participants mentioned resulted in higher ratings and types of negative reasons in lower ratings of the evaluated lecturers. I also investigated whether this differed for female versus male target lecturers.

I conducted bivariate correlation analyses between the number of positive versus negative justifications (themes) and the rating scores participants provided for the lecturers evaluated. I analysed the number of justifications only of the most frequent type of reasons per each item. However, for the "engagement" item, I chose later in the coding process to combine the themes (Lecturer provided/did not provide 'Different methods', 'Visual elements' and 'Videos') into one theme, 'Variety of methods'. This theme of type of reasons became the most frequent for this item. I, therefore, considered the most two frequent types of reasons for this one item. I conducted Benjamini-Hochberg method to correct for multiple analyses[35] with 5% False Discovery Rate (Benjamini & Hochberg, 1995). After applying the Benjamini-Hochberg procedure, the highest p value I deemed significant was $p = .009$.

I also analysed participants' ratings of their lecturers on the five SET items for any significant differences between the scores of male and female lecturers with One-Way MANOVA.

6.3.4.1 General overview and interpretation of findings

As to be expected, the more negative reasons (e.g., lack of 'Support' provided by lecturers, poor quality of 'Explanation-teaching style') participants mentioned, the lower they rated the target lecturers. In contrast, the more positive reasons (e.g., sufficient 'Support' provided, high quality of 'Explanation-teaching style') for choosing categories participants mentioned, the higher they rated the lecturers.

Participants considered some reasons more strongly in their ratings of particular items. The more frequently participants provided affirmative comments about 'Lecturer explaining topic well- using clear teaching style' or 'Supporting or showing interest in students, the higher they rated all lecturers (specifically on the "effective teaching" and "support with assessment" items). In contrast, the more frequently participants mentioned

[35] P-values that would previously demonstrate significant differences, $p = .013$ and $p = .045$ were rejected after conducting this method.

poor quality of provided ‘Explanation-teaching style’, or that ‘Lecturer did not support or show interest in students’, the lower they rated all lecturers on the same items.

The more frequently participants considered ‘Lecturer provided poor overall feedback-marking’, on the “feedback” item the lower they rated both male and female lecturers. Similarly, the more participants mentioned that ‘Lecturer provided variety of methods’ or ‘Lecturer provided useful content-resources’, the higher they rated the target lecturers on the “engagement” and “organisation” item.

When rating the target lecturers, participants, therefore, appreciated high quality of explanation, teaching style, provided support, and content, but disliked lack of support, unclear explanation or poor quality of feedback. This suggests that although participants may think about different types of reasons when evaluating their lecturers, participants may consider only some of these reasons more strongly in their ratings.

6.3.4.2 No effect of lecturer’s gender on overall ratings

Overall, participants did not rate male and female lecturers differently across all five SET items, F (5, 327) = 0.41, p =.843, Pillais’ Trace = .006, with effect size ηp^2 =.006.[36] Therefore, univariate effects of gender on rating scores were not considered (*see Table B2, Appendix for descriptive statistics*).

These findings contradict some of the previous research which showed bias against female lecturers who received lower ratings than male lecturers even for the identical teaching behaviours (e.g., MacNell et al., 2015; Mengel et al., 2019). Contradictory findings could be explained by different context (e.g., country, discipline). These contextual aspects could interact with gender but not show in rating scores (Bachen et al., 1999; Gelber et al., 2022; Laube et al., 2007; Renström et al., 2021).

6.3.4.3 Differences in weighting of the types of reasons for lecturers of a different gender

After applying corrections for multiple analyses, participants weighted 75% of the analysed types of reasons similarly for both studied genders. However, participants weighted

[36] I analysed the scores to comply with mainstream psychology standards. Importantly, the analysed rating scores represent answer categories recoded into numerals that can represent a wide range of meanings (as demonstrated above). These numerals do not have mathematical properties that researchers commonly ascribe to them (Uher, 2022a)

25% of the teaching behaviours (described in the analysed types of reasons for choosing categories) differently for lecturers of different genders.

The more frequently participants mentioned that 'Lecturer provided poor explanation-used unclear teaching style', the lower they rated only male lecturers (M = 3.13, SD = 1.05), on the "engagement" item, r (165) = -.28, p < .001, but not female lecturers (M = 3.04; SD = 1.06), r (167) = -.16, p < .045[37], on the same item.

Similarly, the more frequently participants mentioned that 'Lecturer provided useful overall feedback-marking', the higher rating scores they awarded to only male lecturers (M = 3.71; SD = 1.13) on the "feedback" item, r (165) = .21, p = .007, but not female lecturers (M = 3.69; SD = 1.17), r (167) = .10, p = .204.

The chi-square analyses revealed that participants considered these types of reasons ('provided explanation' and 'feedback-marking') similarly frequently for male and female lecturers when rating them on these specific items[38]. This is a positive finding because it means that, on average, student participants in this sample made similar considerations about teaching behaviours of their male and female lecturers.

6.3.4.4 Brief discussion

In sum, participants mostly used answer categories similarly when judging male and female lecturers, however, there were differences only in a small subset of the analysed types of reasons.

Specifically, participants weighted some considerations more strongly in their ratings for lecturers of certain gender. The more frequently participants mentioned lecturer provided poor 'Explanation-teaching style' on the "engagement" item, the lower participants rated only male lecturers. Participants also weighted a high quality of 'Overall feedback-marking' on the "feedback" item only when they rated male lecturers. When rating "engagement", participants appeared to penalise male lecturers more for poor explanation than women.

Interestingly, when rating "effective teaching", participants weighted both negative and positive comments about 'Explanation-teaching style' strongly in ratings of all lecturers. This provides some support for stereotype content model, as explanation skills could be

[37] The p value was no longer considered significant after applying Benjamin-Hochberg procedure to correct for multiple analyses.

[38] Although participants mentioned good quality of lecturer's explanations significantly more frequently for male than female lecturers (see RQ8), this was in the whole dataset overall and not for specific items explored here.

associated with competence (often perceived as masculine characteristic). Perhaps participants in this sample expected their male lecturers to explain content to a high standard and considered any deviation from the expected performance salient, even when rating male lecturers on items not directly related to "effective teaching".

Furthermore, participants weighted more consideration of high quality of feedback and marking only in their ratings of male lecturers, specifically, the more frequently participants mentioned affirmative comments about feedback and marking, the higher they rated but only their male lecturers. This implies that participants may have appreciated high quality of feedback more from their male as opposed to female lecturers. This finding is quite novel and consistent with findings from Study 1.

In sum, although participants considered some teaching behaviours similarly frequently for both male and female lecturers, participants weighted some considerations more strongly in their ratings for lecturers of certain gender. This implies that participants in this study applied in some fewer cases gendered schemas when evaluating their lecturers.

My findings are consistent with recent findings by Gelber et al. (2022) and Renström et al. (2021). Their research did not show any significant differences in ratings of male and female lecturers, but detected that participants were subtly applying gendered schemas, although participants in my study only did so in a minority of cases.

Despite scenarios describing the identical situations, participants weighted some fewer considerations in their ratings differently for male and female lecturers. However, in real life, students may be affected by various other factors (e.g., academic subject, lecturer's voice, ethnicity, age) and forced to also rely on their memories as they may need to recall teaching behaviours over time. This demonstrates subjectivity in students' choices, which may lead to the potential influence of gender biases or stereotypes.

6.4 General summary of this chapter

In Study 2, I investigated how students interpret and use the answer scale categories. I also examined any potential gender differences in how students rate lecturers of a different gender for the identical teaching behaviours and how students weight these behaviours in their ratings. I found extensive variations in participants' interpretations and use of answer scale categories. The sample provided the same reasons for choosing different answer categories, and, vice versa, different reasons for choosing the same categories. Some participants contradicted themselves, made obvious mistakes in their ratings or used a mid-category instead of a non-applicable.

Overall, students did not rate male and female lecturers significantly differently across five SET items. Although students weighted 25% of the analysed considerations differently in their ratings, this applied to their ratings of both male and female lecturers. There were no gender differences in how frequently students considered 91.7% of the analysed types of reasons for male or female lecturers. This is a promising finding that suggests there was no evidence of pervasive gender bias in this sample. Perhaps these participants were more aware of the potential influence of sexism or gender stereotypes. When answering the open-ended question about judging their lecturers, over 6% of participants provided reasons that suggested their awareness of sexism, gender biases or stereotypes.

In the next chapter, I discuss general key patterns in my findings from Study 1 and Study 2, relate them to the previous literature, and highlight relevant contribution of my research to the existing literature.

Chapter 7: Overall analyses and discussion

In this chapter, I discuss my key research findings. I present each research question separately. At the end of each section reviewing research questions that addressed a specific topic, I discuss the original contribution of my findings to the existing research. In the second part of this chapter, I outline the practical and theoretical implications of my research and recommendations for practice. I summarise the overall contribution of this PhD research to the field and briefly discuss its limitations. Afterwards, I provide the directions for future research and present an overarching conclusion of my PhD research.

7.1 Students' views on lecturers' teaching behaviours and inferred attitudes

In research questions through which I explored this key topic, I aimed to examine which teaching behaviours and inferred attitudes of their lecturers do students generally consider salient. It is critical to differentiate between teaching behaviours and attitudes, because behaviours can be observed, but attitudes only inferred (i.e., from these behaviours, *see Chapter 3*).

I aimed to answer this question in two ways. First, I explored teaching behaviours and attitudes that students considered when reflecting on their specific lecturers in general (before providing any rating*; see Chapter 5*). I instructed participants to reflect on teaching behaviours and underlying attitudes of their lecturers separately, which was rarely done in the previous research. Second, I examined what teaching behaviours and inferred attitudes students considered *when rating* lecturers (both specific and described in scenarios) on SET items (*see Chapters 5, 6)*.

7.1.1 Key findings and relation to the previous literature

7.1.1.1 RQ1: Which of their lecturers' teaching behaviours do students generally consider important?

The whole sample in Study 1 reported a wide range of teaching behaviours they considered important (*see Chapter 5*).

Consistently with the previous findings, my participants also focused predominantly on teaching-related behaviours during the lectures (Emmelman & DeCesare, 2007). Participants in my study valued if lecturers engaged students (Chang-

Kredl & Colannino, 2017; Reupert et al., 2009; Slate et al., 2009; Tam et al., 2009), demonstrated passion and enthusiasm (Anghelcev et al., 2023; Basow et al., 2006; Tam et al., 2009), had good communication skills (Slate et al., 2009), supported and cared about students (Basow, 2000; Chang-Kredl & Colannino, 2017; Slate et al., 2009), and used humour (Anghelcev et al., 2023; Slate et al., 2009; Su & Wood, 2012).

Interestingly, student participants considered several teaching behaviours assignments salient only when they encountered some problems in this area. Similarly, participants reported good engagement much more frequently than lack of engagement. Students may therefore emphasise certain teaching behaviours and attitudes differently depending on the context in which students perceive them. Perhaps students have a certain threshold of what they expect of their lecturers and may notice any teaching behaviours that deviate from these expectations (*see Chapter 5*).

Participants in my sample frequently described their "best" lecturers as active, upbeat, using their body language, walking around, or acting energetically. This is an interesting finding that was also reported in Benz and Blatt (1996), but did not appear frequently in other research, although participants in Anghelcev et al. (2023) reported energetic behaviours and also framed it as a sign of a lecturer's passion, the same as participants in my sample. Therefore, energetic behaviour may be seen as a demonstration of 'passion or enthusiasm', which is commonly mentioned in other studies. Perhaps students generally infer an 'enthusiastic' attitude from the observable energetic behaviours of lecturers, but distinguishing between behaviours and attitudes in my study led them to report 'energetic' behaviours separately. This could disadvantage lecturers who cannot present energetically (e.g., because of disability), hindering their ability to demonstrate passion or enthusiasm to students.

7.1.1.2 RQ2: Which attitudes that students ascribe to their lecturers do students generally consider important?

In line with the previous findings, my participants in Study 1 (*Chapter 5*) appreciated if lecturers seemed friendly, approachable, enthusiastic and passionate, or student-centred (Anghelcev et al., 2023; Basow et al., 2006; Chang-Kredl & Colannino, 2017; Emmelman & DeCesare, 2007; Mogan & Knox, 1987; Slate et al., 2009; Sprague & Massoni, 2005; Tam et al., 2009).

In contrast, my participants disliked mean, uncaring, and unsupportive lecturers (Basow et al., 2006; Chang-Kredl & Colannino, 2017; Emmelman & DeCesare, 2007;

Mogan & Knox, 1987; Slate et al., 2009; Sprague & Massoni, 2005). A rather unanticipated finding in my study was how often student participants praised their lecturers for positive attitude (e.g., being generally positive, relaxed, confident, or happy). This attitude differed from enthusiastic attitudes that participants also frequently mentioned that I categorised into a different theme. Although some research reports students valued relaxed atmosphere or accommodating lecturers (Chang-Kredl & Colannino, 2017; Emmelman & DeCesare, 2007), my participants commented specifically on general positivity and happiness of their lecturers. The most similar finding was reported in a recent study in which participants mentioned that lecturers should have been more relaxed, less negative or smile more (Adams et al., 2022). In contrast, instead of solely disapproving of lecturers' negative attitudes, participants in my sample more frequently highlighted positive attitudes of their lecturers. Perhaps students became more aware of the importance of well-being in recent years and considered positive attitudes to be more salient compared to the previous generations of students. This finding, however, implies that lecturers may need to engage in additional emotional labour to appear happy to students.

Surprisingly, some participants also mentioned counter-intuitive attitudes (e.g., potentially undesirable attitudes for their "best" lecturers). Similarly, some participants contradicted themselves in their answers. Participants also sometimes listed attitudes, although I instructed them to report behaviours (and vice-versa).

Similarly to prior research (Benz & Blatt, 1996), some of my participants considered study phenomena that were absent. There was a discrepancy between several teaching behaviours and inferred attitudes that participants reported. That suggests participants may have inferred lecturers' attitudes even from unrelated teaching behaviours or subtle behaviours (smiling, tone of voice) that participants may have noticed but not reported or even from discussions with other students.

7.1.1.3 RQ3: Which teaching behaviours and inferred teaching attitudes do students consider in their SET ratings?

In both studies, participants as the whole sample considered a wide range of teaching behaviours and attitudes, even for the same item statement (*Chapter 5*) or an answer category (*Chapter 6*). However, each participant considered on average only between one to two teaching behaviours or attitudes, or reasons for picking the answer category. This shows students may have in mind various referents (e.g., teaching

behaviours) for the same item statement. From the measurement point of view, it remains unclear which specific referents (study phenomena) students considered even for the same item thus precluding data generation traceability (Uher, 2023). For example, researchers could assume that the item evaluates specific teaching behaviours, even though participants considered different referents. This would remain unclear if participants simply ticked a box without explaining what they considered while rating their lecturers. This finding has serious implications. If students have a different understanding of the item, their ratings cannot be compared with each other. The results (symbolic system) cannot be traced back to study phenomena (empirical system) that students considered when rating their lecturers.

Student participants differed in the ways in which they provided answers. Some participants provided long and descriptive responses, some simply repeated what items stated without providing any context, and others referred to their previous answers even when rating their lecturers on different items. This implies that some participants might have been influenced by the halo or horn effect (tendency to like or dislike all the features of the evaluated individual; MacDougall et al., 2008; Murphy et al., 1993; Thorndike, 1920; Wetzel et al., 2016). Furthermore, it suggests that participants may have interpreted meanings of different item statements as interchangeable.

In fact, the meanings that participants constructed for item statements frequently overlapped between items. Occasionally, participants interpreted items with the meanings that seemed better suited to other SET items. Therefore, when students rate their lecturers on certain items, students may have in mind teaching behaviours that relate more to other items and differ from the meanings that developers intended for these items. Several participants directly disagreed with items that put them into passive roles (*see Chapter 5*).

An interesting frequent theme involved students' ideas about 'independent learning'. Some participants in my sample highlighted the active role that students should play in their learning and did not rely exclusively on support from lecturers. This contradicts older research in which participants attributed responsibility for their learning to lecturers (Benz & Blatt, 1996). However, views that the participants in my PhD research expressed remained mixed. Some participants acknowledged that the university should emphasise independence whereas others reported reluctance to seek information by themselves. This shows individual differences in students' expectations of their lecturers.

Overall, there was a pattern that suggested students may expect from their lecturers to engage or entertain them. Similarly, participants strongly disliked a traditional teaching method of reading from slides. Perhaps participants perceived lecturers reading off the slides as lacking knowledge. Alternatively, participants may have expected more dynamic approach from lecturers, and found it dull to passively listen to the read content.

Participants sometimes interpreted even the same standardised scenario (describing teaching situations, e.g., lecturers presenting the topic and then answering students' questions) in very distinct ways. The scenarios were ambiguous and contained both positive and negative behaviours. Therefore, the differences in how participants interpreted the scenarios may be expected to certain extent. But some participants framed even the same behaviours in a negative whereas other participants in a positive context. This was a common pattern in my data that again highlights an active role of raters and has serious implications. Participants judged even the same teaching situation differently, which indicates subjectivity in their judgements. In this study, I minimised information that was available to participants, but findings still suggest many differences in participants' perspectives. In real life, students may also consider a lecturer's age, ethnicity, nationality, type of subject, many other factors unrelated to teaching quality in subjective ways. Students consider and interpret a variety of contexts that may invoke even more diverse perspectives. The scenarios were rather short, whereas in real life students usually complete SETs at the end of the term and must think about teaching situations that occurred throughout the year. Furthermore, in a real-life context, students must recall this information from their memories. This memory recall is prone to errors and can be influenced by many biases (Dodson et al., 2008; Lenton et al., 2001; Sherman et al., 2003). Furthermore, this task may be even more complex if students need to consider teaching behaviours of several lecturers (e.g., in real-life contexts, multiple staff may teach the same module).

Interestingly, participants sometimes framed the same teaching behaviour in a positive context when rating one item, but a negative context for another item. This suggests students may have higher expectations of their lecturers when rating them on certain items.

Student participants, therefore, considered different study phenomena (e.g., teaching behaviours – empirical relational system) when interpreting the item statements (the same symbolic relational system). Similarly, participants also interpreted, perceived

and framed standardised scenario (description of the same teaching situation) in different ways. This shows the important role that raters have in generating data with rating scales, specifically, their subjective interpretations of items which may let them consider different study phenomena than intended. This task would be even more complex in real life in which students experience a variety of teaching situations over a long period, some of which may be ambiguous.

7.1.2 Original contribution to the existing research.

This PhD research provides an overview of teaching behaviours and attitudes that students generally considered salient when reflecting on their lecturers and in teaching evaluations. It also offers important insights into what specific information students considered when rating their lecturers on different SET items, which was rarely examined before, especially in the context of universities in the U. K.

7.2 Students' interpretations and uses of item statements and answer scale categories

The following research questions aimed to examine the ways in which students interpret and use item statements and answer scale categories *in general*. This analysis focuses less on the meaning of the themes of teaching behaviours and attitudes examined in the previous section and instead discusses general patterns in the ways in which participants interpreted these key elements of rating scales (*see Chapters 5, 6*).

7.2.1. Key findings and relation to the previous literature

7.2.1.1 RQ4: In what ways do students interpret the item statements of SET scales in general, and what meanings do students generally construct for these statements?

Consistently with the previous empirical findings (Arro, 2013; Lundmann & Villadsen, 2016; Rosenbaum & Valsiner, 2011; Uher, 2018; Uher & Visalberghi, 2016), participants in Study 1 developed a wide range of meaning for the same standardised item statement.

Similarly, my findings are in line with previous research in SETs, which also showed that student participants interpret SET items broadly (R. Bennett & Kane, 2014; Benz & Blatt, 1996; Block, 1998; Robertson, 2004). Importantly, the meanings that participants constructed for different SET item statements frequently overlapped between items.

Interestingly, participants frequently mentioned ideas related to 'students' experience'. This, again, emphasises the active role of raters. Participants frequently reported that they interpreted meanings of item statements in relation to how participants themselves felt about the module, or their learning. This provides further evidence that participants may interpret SET items in subjective ways.

In line with the previous evidence (Block, 1998; Uher, 2018), even though my participants as a sample constructed a large field of meanings for the same item statements, each participant, on average, considered only one to two meanings per item. This shows substantial individual differences in how students interpret SET items. Specifically, different participants considered very different referents for the same item (e.g., teaching behaviours and attitudes). This finding can be analysed from the viewpoint of two measurement principles, data generation traceability (a requirement that an unbroken, traceable chain of connections that establishes proportional relations between the study phenomena and the results is implemented) and numerical traceability (a requirement that assigned numerical values must be connected to a known quantity standard in transparent and defined ways; *see also Chapter 3*). Applying these measurement principles, this means that raters are encoding information about different elements of the empirical system into the same elements of the symbolic study system (an item statement), which precludes comparing ratings generated even for the same item (Uher, 2023). These variations in how student participants interpreted SET items mean that unbroken traceable connections chains that establish quantitative relations between the results (SET scores) and the measurands (different teaching behaviours) could not be established. The crucial principle of measurement, data generation traceability, was, therefore, not fulfilled and it could not be ascertained whether results showed valid information about study phenomena. Results, therefore, could not be justifiably attributable to measurands, failing the first crucial criterion of measurement (*see Chapter 3*).

Sometimes participants simply repeated a part of the item statement without clarifying what it actually meant to them. This implies that some students may interpret item statements without carefully considering their meaning and could result in students using mental shortcuts to make quick but superficial decisions. Alternatively, participants may have rushed to complete the study without considering these questions, or even felt that the items were self-explanatory. Many participants' quotes contained grammar or

spelling mistakes, which indicates that participants may have been in a hurry or completed the studies in a sloppy manner. All of these explanations have problematic implications, because if student participants exhibited these behaviours in my study, some students may also do so when completing SETs in real life at universities.

7.2.1.2 RQ5: In what ways do students interpret and use the answer categories of SET scales when completing teaching evaluations?

Participants in Study 2 overall provided types of reasons for choosing categories with similar frequencies for both male (52.7% positive, 48% negative) and female lecturers (47.3% positive, 52% negative). Similarly, participants provided both positive and negative types of reasons for picking answer categories when rating lecturers. Participants also did so with similar frequencies for both scales starting with "strongly disagree" (49.1% positive, 48.6% negative) and scales starting "strongly agree" (50.9% positive, 51.3% negative; *see Chapter 6*).

In line with the previous findings (Block, 1998; Uher, 2018), participants as a sample also provided a wide range of reasons for picking the answer categories. But similarly to raters' interpretation of item statements, each participant on average considered only between one to two types of reasons for choosing an answer category per item statement This implies that student participants usually focused on only one or two key pieces of evidence to make their overall judgement

To explore this idea further, I examined general patterns in how participants rated lecturers using one item as an example. The majority of participants provided only one reason, although almost a third of participants weighted and averaged several pieces of evidence and over 20% of participants considered several reasons. This is partially consistent with findings by Uher (2018b) who found that most participants did not weight or average all evidence available, although a higher number of participants did so in my research. I used scenarios instead of videos. Therefore, participants in my sample had to consider fewer pieces of evidence and may have found it easier to weight it. Similarly, in my study, more participants based their decision on only one reason. This further supports the idea that raters consider only a part of the field of meaning. Interestingly, only over 2% of participants in my study (when rating target lecturers on this item) mentioned not having sufficient evidence to make a judgment. This is quite surprising, considering that participants had to make their judgements based on information provided in a short scenario. Considering that, compared to real life, participants only judged brief

information described in scenarios, it would be expected that more participants report they did not have sufficient information to judge.

The variations between how participants interpreted answer scale categories were slightly broader when participants rated the same lecturer on the same scenario but with a different direction of a rating scale, or different lecturer's gender (compared to rating lecturers of the same gender with the same direction of a scale). Therefore, making context even slightly different broadened variations of participants' interpretations.

Interestingly, some participants rated the fictitious target lecturers and then contradicted themselves when providing a reason for picking an answer category (for example, by rating a lecturer with "strongly disagree", but then providing a very positive comment). This supports the patterns observed in earlier SET studies (Benz & Blatt, 1996; Robertson, 2004). Perhaps participants felt ambiguous about some of the teaching behaviours described or started questioning themselves once they were asked to justify their choice. Benz and Blatt (1996) theorised that students may initially make a quick judgement (e.g., with the use of heuristics), but upon being asked for justification, they start thinking more analytically about their choice and change their opinion. The evidence from my study shows that at least some participants may have done so because a few participants explicitly stated they changed their mind when asked to explain their choice of answer category.

Participants also made some obvious mistakes in choosing answer scale categories and some explicitly stated they made a mistake. Some participants mentioned not having sufficient information to make a choice, usually as a reason for picking a mid-category ("neither agree nor disagree"). This may imply that participants may see this category as interchangeable with 'non-applicable'. This idea is further supported by another observed pattern in data. Several participants reported the situation did not apply to them, they did not know what was meant by the question or felt unsure about why they picked this category. Most participants who provided these reasons also picked a mid-category. This shows two important points. First, some students may struggle with understanding instructions or questions when completing SETs. Second, some students may mistakenly use 'mid-category' instead of 'non-applicable'. Participants may have been distracted, fatigued, or perhaps simply rushed to complete the study. Importantly, many of these mistakes likely occur when students complete SETs in real life, but this may not be directly apparent from obtained data.

My findings showed that participants as a sample frequently provided different types of reasons for choosing the same answer category. This was a common pattern that was repeated throughout the whole dataset. Participants also sometimes provided more types of reasons for choosing an answer category on a lower end of the scale than a category on a higher end. But relations between these numbers do not represent relations between how participants used and interpreted these answer categories (Uher, 2022a, 2023) with regard to the teaching behaviours they had in mind.

Similarly, participants provided the same types of reasons for picking different answer categories. This implies that answer categories have overlapping rather than mutually exclusive meanings, which contradicts the quantitative meanings of the numerical values into which researchers recode these verbal categories, and thus the idea of measurement (Uher, 2022b, 2023).

Considering the two basic measurement principles, these variations in how participants interpreted and used answer scale categories entail that the results (numerical values) did not always represent the same information about study phenomena. Results were not connected to any known reference quantities and lacked transparent quantitative meanings. Specifically, the generated numerical values represented different information that depended on subjective perspectives of participants. All of this precludes numerical traceability. Consequently, the second crucial criterion of measurement, the public interpretability of the results' quantitative meaning, could not be established (*see Chapter 3*).

7.2.2 Original contribution to the existing research

In sum, I conducted one of the first extensive examinations into how students in the U. K. interpret and use key elements of rating scales (item statements and answer scale categories). I studied the meanings that participants constructed for specific items and answer scale categories as well as general patterns in the ways in which participants used rating scales and analysed them through the view of the two basic measurement principles applicable across sciences. Furthermore, I examined whether participants used answer scale categories differently based on the gender of a target lecturer and the direction of the scale used (*see also 7.3*)

7.3 Gender differences between lecturers

7.3.1. Key findings and relation to the previous literature

7.3.1.1 RQ6: What are possible differences in the teaching behaviours and inferred teaching attitudes that students consider in SET ratings when judging male versus female lecturers?

Contrary to the previous research (MacNell et al., 2015; Mengel et al., 2019; Özgümüs et al., 2020), participants in Study 2 did not rate female and male lecturers differently for the same teaching behaviours (described in scenarios; *see Chapters 2, 6*).

Overall, there were also no significant differences in how participants in Study 1 rated their "worst" specific male and female lecturers. However, participants rated their "best" male lecturers significantly higher than "best" female lecturers across all seven SET items, and on the "overall satisfaction" item (*see Chapter 5*).

When reflecting on teaching behaviours *before providing any rating*, participants generally considered specific behaviours similarly frequently for their "best" male and "best" female lecturers. Participants only commented significantly more frequently than expected on 'engagement-related behaviours' of their "best" male but not "best" female lecturers (*see also RQ7*). For "worst" lecturers, there were no significant gender differences in how frequently participants considered content related to teaching behaviours and attitudes of lecturers of a different gender (*see Chapter 5*).

A higher number of participants commented on good 'explanation-teaching style' significantly more frequently than expected for their male (14.3% of the sample) than for female (8% of the sample) target lecturers *(see Chapter 6*). These patterns imply participants commented more frequently than expected on the positive behaviours of the male but not female lecturers. However, in both studies, overall, the samples considered teaching behaviours or attitudes similarly frequently for both male and female lecturers.

When explicitly asked whether participants would judge male and female lecturers differently, over 80% of participants reported they would not, over 10% of participants would, and the rest were unsure. Interestingly, some participants mentioned subconscious or unintentional stereotypes that could affect them or otherwise indicated their awareness of sexism and gender issues.

7.3.1.2 RQ7: How do students weight specific teaching behaviours and inferred attitudes when generating their ratings of the lecturers on SET items, and does this differ by the lecturers' gender?

I explored how participants in Study 1 weighted specific content (teaching behaviours, attitudes) in their ratings. Specifically, I explored if the more frequently participants mentioned certain teaching behaviour, the higher or lower they rated their lecturers (depending on whether behaviours were positive or negative) and whether this differed depending on the gender of a lecturer (*see Chapter 5*). Importantly, although I found some gender-related differences in the ways in which participants considered teaching-related behaviours for lecturers of a different gender and weighted these considerations in their ratings, these differences formed only a very small proportion of the analysed cases. Therefore, in most cases, participants in this sample did not seem to be influenced by gender schemas when judging their lecturers, which is an encouraging finding. However, a few gender differences emerged for the following teaching behaviours.

The more frequently participants mentioned lecturers engaged poorly with class, the lower they rated their "worst" male lecturers. This supports evidence that students may expect their lecturers to entertain them (Anghelcev et al., 2023; T. Bennett, 2021; Strage, 2008). But consistent with the previous findings (Sprague & Massoni, 2005), student participants in my sample held only their male but not female lecturers to a standard of an entertainer (*see also RQ6*). Evidence suggests that individuals may hold a stereotype of men being funnier than women (Hooper et al., 2016; Mickes et al., 2012). This corresponds to participants in my sample potentially being affected by this stereotype and considering engagement-related behaviours salient only for male lecturers.

The more frequently participants mentioned high 'quality of feedback', the higher they rated their "best" male lecturers. In contrast, the more frequently participants considered poor 'quality of feedback', the lower they rated but only their "worst" female lecturers (*see also RQ8*).

Participants in my sample weighted positive 'interaction with students' significantly more strongly in their ratings of their female than male lecturers, suggesting they valued warmth and interpersonal characteristics of their female more than male lecturers, consistent with previous findings (S. K. Bennett, 1982; Kierstead et al., 1988; Mitchell & Martin, 2018; A. H. Wu, 2017). This provides some support for the stereotype

content model and benevolent sexism (*see Chapter 2*), although it must be highlighted that I found no other gender differences related to participants emphasising warmth or interpersonal characteristics of their female lecturers in the rest of the analyses. In fact, before providing ratings (in Study 1), my participants also mentioned 'interesting-caring' attitude three times more for their male than female lecturers. Although this difference was not significant, it shows an interesting pattern, as participants frequently praised also interpersonal characteristics of their male lecturers.

A slightly surprising finding is that the more often participants mentioned content related to good 'organisation', the higher they rated female lecturers. This contradicts previous findings in which students considered organisation skills salient for male but not female lecturers (Basow, 2000; Basow et al., 2006). In general, 'organisation skills' could include preparing content of high quality, which could be associated with competence. For example, students rated male lecturers higher than female for the identical teaching materials (Mengel et al., 2019; Özgümüs et al., 2020). Perhaps student participants in my sample interpreted and framed this behaviour in different ways compared to participants in the earlier studies.

7.3.1.3 RQ8: What are the potential differences in the ways in which students use and interpret typical answer categories on SET Likert scales when judging male versus female lecturers?

Overall, participants in Study 2 did not rate male and female lecturers significantly differently across all five SET items when reading the scenarios about the same teaching situations. After reading the scenarios, participants weighted 75% [39] of the analysed themes of reasons for picking particular answer categories similarly in the ratings of male and female lecturers (*see Chapter 6*).

When providing reasons for picking answer categories, my participants praised their male but not female lecturers more frequently than expected for providing a great explanation. However, participants weighted 'poor explanation' more strongly in their ratings of male lecturers. This shows a partial support for the stereotype content model because good explanation skills could be associated with competence, which in this model is traditionally perceived as masculine characteristic.

[39] Specifically, out of 12 analysed types of reasons, there were differences between male and female lecturers for three of them.

In contrast, the more often participants mentioned high quality of feedback when providing reason for choosing the answer scale category, the higher they rated, but only their male lecturers. This is the only gender-related pattern replicated in both studies. Previous research showed that students may consider favourability of their feedback more salient for their female but not male lecturers (Sinclair & Kunda, 2000). In contrast, participants in my study focused on the perceived quality of feedback. Quality or usefulness of feedback could be seen as a teaching behaviour associated with traditional female stereotypes, such as being nurturing (Boring, 2017), for example, by providing detailed useful suggestions for students. Although participants in Study 1 penalised female lecturers for poor feedback, participants in both studies also appreciated male lecturers providing 'high quality of feedback'. Boring (2017) observed that participants rated lecturers of their own gender higher on the 'usefulness of feedback'. However, although over 70% of participants in my research were women, participants overall still seemed to appreciate high-quality feedback more from their male lecturers. Perhaps participants automatically expected their female lecturers to provide useful feedback and penalised them for deviating from these expectations. In contrast, participants may have not expected detailed feedback from their male lecturers and appreciated receiving it.

7.3.2 Original contribution to existing research.

In addition to examining potential differences in the ways in which participants rated their male and female lecturers, I also studied whether participants considered certain information more frequently than expected for lecturers of either gender, both, when reflecting on teaching behaviours and inferred attitudes of their lecturers in general, and in the context of the specific SET items in particular.

Importantly, I also explored how specific information that participants considered related to their ratings of lecturers, and whether this differed by their lecturer's gender. Although SET research frequently examines potential gender biases in rating scores, what specific information students actually consider and how is this related to their ratings of lecturers has hardly been explored. The nature of my research in this aspect is predominantly exploratory and aims to provide some foundation for future research.

7.4 Students' approach and views on SETs

7.4.1 RQ9: How do students regard SETs and approach their completion in general?

7.4.1.1 Key findings and relation to the previous literature

Generally, participants in Study 1 reported spending approximately ten minutes on completing teaching evaluations, and 70.6% participants stated they completed them spontaneously, instead of thinking about them throughout the year (*see Chapter 5*).

In previous research, over half of participants reported spending sufficient time on SETs (Marlin, 1987). Ten minutes may seem like sufficient time; however, students frequently need to rate their lecturers on over ten item statements. For example, the National Student Survey 2023 contains 27 core questions, and the guidelines on the website estimate students should complete this section in approximately ten minutes. Students following these official instructions, would, therefore, read and interpret item statements and answer scale categories, recall teaching behaviours from their memories, and rate their lecturers on almost three item statements per minute. This may result in students rushing with their answers and relying on mental shortcuts.

Consistently with the previous research (Kite et al., 2015; Spencer & Schmelkin, 2002; Spooren & Christiaens, 2017), my findings showed that over 47% of participants found SETs useful, important and believed SETs can improve lectures. However, participants who completed these studies (including my research, which involved answering open-ended questions), may not be representative of the general student population. Specifically, these participants may have liked completing surveys and perhaps viewed SETs more positively compared to other students.

Over a third of my participants perceived SETs as ineffective. Others emphasised that feedback should be taken seriously, potentially implying that is it not, which is in line with the previous findings (Spencer & Schmelkin, 2002; Surratt & Desselle, 2007). But students reported their main motivation in completing SETs is a belief that lecturers will consider their feedback or an improvement in teaching (Y. Chen & Hoshower, 2003; Iqbal et al., 2016). However, if students believe their SETs will not have an impact, they may exert less effort into completing them (Dunegan & Hrivnak, 2003; Heilman, 2012; Uijtdehaage & O'Neal, 2015).

Almost a third of participants in my study mentioned that feedback must be taken or implemented seriously (26.5%), whereas others reported that SETs should be done

more regularly (8.8%) or that they were unsure about their opinions on SETs (5.88%). Consistently with recent research in Norway (Borch et al., 2020), some participants in my study (11.8%) mentioned that SET questions are too broad, frustrating, or unclear. This is an important finding that could be related to low participation rates in SETs. As discussed earlier, researchers frequently word item statements in abstract ways to apply them to a wide range of phenomena. But several participants mentioned items should be more specific or better explained. This suggests that some participants may struggle to determine what study phenomena (e.g., teaching behaviour) they should consider when interpreting some SET items.

7.4.1.2 Original contribution to the existing research

In addition to examining how students perceive SETs in general, I also explored their approach towards SETs, such as in terms of the estimated effort they put into completing SETs. My findings showed that although most participants in my sample perceived SETs positively, they had different perspectives and expectations of SETs, indicating a lack of consensus in terms of the overall purpose of SETs. Most participants approached completing SETs spontaneously, instead of thinking about them in advance. These findings also summarised potential problems that students highlighted when thinking about SETs, such as lack of specificity in some SET items.

7.5 Overarching research question

In this section, I summarise key findings and answer the overarching question of this PhD research. I also outline the overall contribution of my PhD research to the literature.

7.5.1 RQ10: Can students' underlying potential stereotypical gender beliefs and expectations of their lecturers influence how students evaluate their lecturers with rating scales?

7.5.1.1 Key findings

In sum, my findings showed extensive inter- and intra-individual variations in the ways in which student participants interpreted and used item statements and answer scale categories. Participants generally constructed large fields of meanings for each standardised item or reasons for picking the answer category.

My findings revealed significant gender differences in how student participants rated their "best" lecturers across seven SET items. However, this significant difference may have been driven by a minority of a sample awarding higher ratings to male than female lecturers. Furthermore, participants weighted content related to teaching behaviours and attitudes similarly in their ratings of male and female lecturers in most cases. Therefore, overall, this sample did not display pervasive stereotypical gender biases. Perhaps asking participants to reflect on why they chose a specific answer category activated more analytical thinking process. This may have led to participants not relying on heuristics to make their judgements. These findings cannot be generalised but are nevertheless encouraging, because they indicate that sexism does not apply in all contexts.

I did find a few gender differences, but only in a few cases. I identified key patterns in how participants weighted teaching behaviours related to 'engagement' and 'quality of feedback' in their ratings of lecturers of a different gender. Participants penalised only male lecturers for not engaging students but seemed to value high quality of feedback only from their male lecturers. Similarly, participants penalised only female lecturers in their ratings for providing poor quality of feedback.

This suggests that, in some fewer cases, participants may be influenced by gendered schemas when rating their male and female lecturers on certain SET items.

7.5.1.2 Overall contribution of my PhD research to the existing literature.

I investigated whether the methodological and methodical limitations of rating scales may enable rather than impede the manifestation of gender biases. This was explored empirically in the context of SETs at universities in the U. K., making this contribution novel.

This research also provided further insight into the ways in which raters, specifically students, interpret and use key elements of rating scales and generate data when evaluating their lecturers. I applied several principles from the TPS-Paradigm to scrutinise the use of SETs from the methodological point of view and examined whether SETs fulfil principles under which data generation could constitute measurement.

I explored connections between information that students considered when evaluating target lecturers and their ratings of these lecturers, which was rarely done in previous research. When examining these connections, I also considered an additional context of whether students' judgements might have been influenced by gender

stereotypes. These findings cannot be generalised, but provide a useful scope for the aspects that may warrant deeper exploration.

In sum, I conducted an extensive empirical investigation that generated new knowledge about how students evaluate their lecturers, interpret, and use rating scales for SETs, and whether students may be influenced by gender stereotypes during this evaluation.

7.6 Practical implications

Exploring SETs through the view of measurement-theoretical principles revealed that even if SETs meet psychometric criteria, data generation through SETs does not constitute measurement. Therefore, numerical values cannot be justifiably attributed to evaluated lecturers nor publicly interpretable due to lacking transparent quantitative meanings.

To highlight a few specific key problems identified in my PhD research, participants, on average, considered only one or two key pieces of evidence. Participants also frequently constructed meanings that overlapped between item statements, or meanings that may have differed from those intended by scale developers. Similarly, participants provided the same justifications for choosing distinct answer scale categories, indicating an overlap in meanings between these answer categories. Some participants reported items were not applicable to their situation, or were too broad, vague and frustrating. Others misunderstood the instructions or made an obvious mistake in picking an answer category. All these problems may apply to students using SETs in real-life contexts but go entirely unnoticed, because the only outcome on rating scales is usually a 'tick' on a box.

These answer categories chosen by students are usually recoded into numerals. Researchers frequently interpret them as numbers, which imposes fundamental logical errors, because the relations between these numbers do not represent the relations between how raters used these answer categories (Uher, 2022a, 2023), as also shown in my research. In an attempt to create quantitative meanings, different cases are compared (e.g., modules in two different years). But the results do not depend on any known reference quantity, only a specific sample (*see Chapter 3*). These results, therefore, do not represent the same quantity across different contexts and should not be compared, as is frequently done. Furthermore, universities frequently calculate averages on ordinal data, although

this procedure is considered inappropriate and misleading (Hornstein, 2017; Jamieson, 2004; Stark & Freishtat, 2014). The means of SET scores may be reported to faculty and SET administrators and even used to compare lecturers' performance (Hornstein, 2017). This may involve an administrator declaring that a lecturer with a mean score of 3.5 is better than a lecturer with a mean score of 3.4 (Franklin, 2001). This approach ignores several methodological and methodical severe problems related to data generation with standardised rating scales discussed in this thesis. These findings have broad practical implications for the educational community, suggesting the necessity to develop a more suitable framework and tools for assessing teaching quality.

Lecturers may be strongly impacted by receiving evaluations that are largely subjective, influenced by the pronounced individual differences between students and their interpretations and use of rating scales, as well as possible mistakes that students may make during rating (e.g., misunderstanding instructions). This can negatively affect lecturers' careers, as well as their well-being, especially if institutions treat SET results as information about lecturers' teaching performance. Institutions should be aware that the use of SETs, especially for summative purposes, may be problematic and, sometimes, discriminatory.

Students may also be affected by the use of SETs. Students' opinions collected through the NSS are considered in the TEF (Teaching Excellence Framework). This may affect the university's placement in the league tables but also tuition fees that universities can charge per year and student (*see also Chapter 1*). Students may, therefore, over-rely on SET results, and this may affect their choices when applying to universities. Similarly, this could negatively impact student experience if lecturers feel compelled to alter their teaching behaviours or course content in order to obtain satisfactory SET scores.

In the next section, I provide several recommendations that could partially improve the situation. But, importantly, much larger structural reforms would be needed to fully remedy these problems. SETs, one of the most frequently used tools for assessing teaching quality, are built on foundations that are methodologically and methodically flawed. Universities should attempt to move towards other methods that enable more holistic approaches to assessments of teaching quality (*see also 7.7*).

7.7 Recommendations for practice

This research revealed several problems with the use of SETs, including pronounced variations in students' interpretations and use of elements of ratings scales. These findings also showed that students may, sometimes, be influenced by gendered stereotypes when rating their lecturers. Participants framed even short identical teaching situations in different contexts, which further highlights the subjectivity in students' evaluations of lecturers. These problems do not concern only SETs but rating scales in research in general. Rating scales are one of the most standard methods used in psychology (Baumeister et al., 2007). Psychologists frequently assume rating scales enable measurement despite serious methodological flaws in their foundations (Uher, 2018b, 2023). This shows the need for serious reforms in psychology as a science (Uher, 2022a, 2023). Similarly, standardised rating scales are predominant in student evaluations of teaching. Institutions generally attach major importance to obtained scores despite numerous researchers emphasising the problematic nature of SETs. Similarly, my research revealed further problems related to SETs (*see 7.6*).

I provide several recommendations that might partially improve the situation. However, the situation requires more definite structural changes at the macro level of educational development. For instance, the participation National Student Survey that also involves conceptual and methodological problems discussed here (e.g., variations in students' interpretations of rating scales, reliance on students' memories, differential analyses) is currently mandatory for universities in the U.K. (*see also Chapter 1*). Therefore, my broad recommendation to institutions is to engage in critical debates about the application of these scales as a tool for evaluating teaching quality. Furthermore, at least at an institutional level, universities should consider applying alternative, more holistic methods that do not have the same conceptual limitations as SETs and may be more informative than SET scores, to obtain students' views about their modules.

I grouped these recommendations into three different themes: a) modifications in a broader educational context, b) exploring alternative approaches to SETs, c) amendments of existing approaches (standardised rating scales), d) students' approach to SETs.

Modifications in a broad educational context involve changes that need to be implemented by policymakers in institutions and researchers. This theme concerns larger structural changes that should be implemented further than just at an institutional level and includes the following recommendations:

1. Terminology surrounding SETs should be carefully reviewed.

The term "student evaluations of teaching quality" is misleading and should not be used. It conceals the subjectivity of the SET process and disregards that students also consider various ideas not related to teaching quality. Using terms such as "student perspectives/opinions/views/ideas of/about this module", would highlight that these are merely student reflections of their lessons. Similarly, a term "measurement" should not be used in regards to results obtained with SETs. As discussed here, SETs failed to meet the necessary principles of measurement and therefore do not allow for justified attribution of results to target lecturers and publicly interpretable quantitative meanings of numerical results.

Terms 'teaching quality' and 'teaching effectiveness' should not be conflated and must be clearly defined. Because my research focused on evaluations of 'teaching quality' specifically, I provide a more elaborated conceptualisation of this term developed from my findings: Teaching quality can be defined against three areas: a) strong teaching methods and skills through which lecturers *explain* material to students, engage them during lectures and facilitate their learning, b) the quality of interaction with students through which lecturers may demonstrate their support for students but also inspire or motivate them, c) the high value and consistency of information that lecturers provide through resources but also feedback. In sum, teaching quality can be defined as: clear teaching methods through which lecturers explain material and provide an opportunity for students to engage; supportive and respectful interaction with students; and high-quality feedback and resources. Therefore, the definition broadens from teaching behaviours occurring solely during class to encompass further behaviours such as feedback that lecturers provide or supportive communication. An important point is that students are not perceived as passive 'receivers' of education; lecturers should provide a supportive environment, high-value teaching, and opportunities to learn, but students need to play an active role in their education. Institutions may conceptualise teaching quality differently, but theoretical definitions should always be provided. 'Teacher quality' usually applies to pre-university contexts (e.g., secondary schools) rather than SETs focused on university lecturers and I will, therefore, reflect on this term only briefly. As discussed earlier, this term is seen as broader compared to 'teaching quality' (Snoek, 2021; Towers et al., 2023)

and encompasses interpersonal characteristics of lecturers. My research focused on student evaluations of 'teaching quality', but participants frequently mentioned inferred teaching attitudes that reflected lecturers' personal or professional characteristics (e.g., passionate attitude) and could be seen as aspects of 'teacher quality'. Participants, therefore, may have considered teaching characteristics that were broader than those contained in the definition of 'teaching quality'. This supports the point raised by Towers et al. (2023) that narrow formal definitions and policies may omit components that are in fact crucial to students or teaching. It also demonstrates the need for a clear review of the terminology used in SETs.

2. SETs should not be used for summative purposes, i.e., in any personnel decisions regarding promotion or salary.

SETs might be used informally to notify lecturers of students' opinions about how to enhance lectures. However, because of their many methodological and methodical flaws, and the potentials for manifestation of bias, institutions should avoid consulting SET results when making decisions that may impact lecturers' careers.

3. Universities should develop a comprehensive framework for the use of SETs.

This might include specifying what construct SETs are intended to assess (e.g., determining whether the focus is on teaching quality, student learning or student satisfaction) and clearly define it (*see also recommendation 1*). It should be outlined what academic values underlie specific SET items and justifications for including these specific items. Faculty should be consulted during this process, in an effort to align the expectations of students, lecturers, and institutions.

4. Feedback should be conducted in the way that enables lecturers to implement changes.

To ensure that lecturers can remedy any potential issues, students could provide lecturers with informal feedback during mid-term. Students could outline aspects of the module they enjoy and aspects that they think could be improved. Lecturers may then consider these suggestions in the overall context of their

teaching approach and decide whether to implement changes or address any misunderstandings of context by managing students' expectations (e.g., by explaining their rationale for certain teaching decisions).

5. **Students should be provided with a brief explaining the purpose of SETs before completing them.**

Students seem to have rather different expectations and understandings of SETs. A brief session could be conducted to explain their purpose and answer any students' questions.

6. **Universities should familiarise themselves with problems surrounding SETs, carefully monitor relevant research and engage in discussions about SETs.**

Despite the prominence of SETs (e.g., the importance placed on the results from the National Student Survey in the U. K.), universities should critically reflect on whether the use of SETs at an institutional level constitutes the best practice. Numerous researchers highlighted problems with SETs, but only a few institutions implemented changes. At a minimum, universities should acknowledge these problems (e.g., subjectivity, biases, methodological and methodical limitations), and provide a statement outlining their position, in order to demonstrate their commitment to upholding research values. Ideally, to enable more prominent structural changes, universities should also initiate these discussions at a macro (e.g., national) level.

Next, I outline recommendations related to alternative approaches to SETs (conducted without the use of standardised rating scales). These approaches are devoid of many methodological and methodical flaws associated with standardised rating scales discussed here and could provide useful information about student feedback. Universities should explore these methods as potential replacements to traditional SETs.

7. SETs should not be used as the only source of information about teaching.

Institutions should not over-rely on SETs, but also assess teaching through other sources, such as peer observations, a comprehensive review of lecture materials, a number of teaching awards, or lecturers' self-reports.

8. Open-ended answers could be analysed with NLP algorithms.

In contrast to SET scores, open-ended answers may provide lecturers with richer and more nuanced feedback. Nowadays, natural language processing algorithms may be used to summarise important topics from large datasets (Arnulf et al., 2021). For example, topic modelling may be used to extract important points from information encoded by students. These key points could afterwards be reviewed by lecturers to put them in the context of their modules.

Furthermore, NLP could be applied to analyse data to detect potential biases. For example, a sentiment model was applied to investigate potential gender bias in evaluations of male and female medical trainees (Andrews et al., 2021). Potentially, these two approaches could be combined. Although these procedures are novel, their use (e.g., for a formative purpose) should be investigated in the context of SETs.

9. Universities should explore the use of conversational agents (AI bots).

Others recommended substituting traditional SETs with dialogue-based methods such as interviews or focus groups (Borch et al., 2020; Darwin, 2012; Steyn et al., 2019). A frequent concern is that procedures such as conducting interviews with students in order to obtain their views could be time-consuming (Constantinou & Wijnen-Meijer, 2022). To address this concern, universities could explore using conversational agents to obtain student feedback, as done in recent research (Pérez et al., 2020; Wambsganss, Winkler, Schmid, et al., 2020; Wambsganss, Winkler, Söllner, et al., 2020). However, instead of using conversational agents to obtain rating scores as done in the above-mentioned research, institutions could aim to obtain richer and actionable feedback from students. For example, students could provide open-ended answers to questions (e.g., about how to improve the module) and be prompted by CA to reflect or

elaborate on their answers. Student answers could afterwards be analysed by the researcher, or potentially with NLP algorithms discussed above (e.g., for larger size courses). This form of obtaining feedback could also boost student engagement in terms of higher participant rates. This is an innovative approach, and universities would need to exercise caution before fully implementing it. However, with a rapid development of technology, it may provide a potential future alternative to traditional SETs and should be considered.

Next, I describe amendments to current standard approaches to SETs. However, I advise caution, as this PhD research highlighted several methodological and methodical flaws related to the use of standardised rating scales that cannot be remedied by implementing these suggestions. I include the following recommendations solely for contexts that compel the use of rating scales.

10. Students could be instructed to reflect on their choices and justify them.

To potentially reduce students' reliance of heuristics, students could be instructed to think carefully about their choices even in traditional SETs. Students could also be prompted to justify their choices. For example, students could be asked to outline what specifically they considered (e.g., what teaching behaviours) when rating their lecturer or to justify why they picked a specific answer category.

11. It could be verified whether students intended to choose a mid-category.

If students choose a mid-scale category, an additional question could verify whether students intended to pick this category, rather than using it instead of non-applicable option.

12. A non-applicable category should be included in SETs.

Similarly to a previous point, a non-applicable category should be added and clearly marked. This could prevent students from feeling obligation to tick a box,

and potentially reduce their use of mid-scale category for non-applicable situations.

13. Non-specific items, such as items evaluating 'overall satisfaction', should be avoided.

Such ambiguous and non-specific items may be even more prone to contributing to the manifestation of biases than other items. Therefore, I would recommend excluding them from SETs.

14. Students' understanding of SET instructions could be verified.

An additional question could verify whether students understood instructions provided in SETs. This would provide students with an opportunity to highlight any problems with their understanding of instructions or items.

7.8 Limitations and future directions for research

Over 70% of my participants were women. This made a gender ratio of participants disproportionate and prevented studying a potential influence of a student's gender on how students judge their lecturers. Male and female students may differ in what they expect from their lecturers and evaluate them in different ways. However, women are, in general, more likely to respond to surveys than men (Becker, 2022; M. J. Wu et al., 2022), and this also applies to participation in SETs (Porter & Umbach, 2006; Porter & Whitcomb, 2005; Sax et al., 2008). Therefore, the disproportionate gender ratio observed in my sample may actually be representative of trends observed in SETs that students complete in real life.

Furthermore, it is unclear whether all student participants paid tuition fees for their university studies and whether participants attended teaching-intensive or research-intensive universities (*see Chapter 5*). Students with different fee-paying arrangements or attending specific types of universities may differ in what they expect from their lecturers and how they perceive teaching quality. This knowledge would have, therefore, provided additional context to my studies. Changes related to Brexit mean that tuition arrangements changed for E.U. students starting in autumn 2021, as they were no longer eligible for home-fee status. This likely did not affect participants in my research as data collection

for Study 2 finished prior to this change, but future studies could explore any potential consequences. My research may have also been influenced by unexpected contextual factors. For instance, due to the industrial action occurring during 2018-2021, some student participants may have experienced strikes during their university studies. Similarly, all participants in Study 2 likely underwent changes associated with their university studies because of the impact of the COVID-19 pandemic.

To study potential gender biases of target lecturers, I applied a simplified binary approach in my research, considering solely male or female lecturers (*see also Chapter 2 for a discussion of terminology regarding gender and sex)*. However, students could also be influenced by stereotypes when evaluating their non-binary lecturers. I fully acknowledge that gender is complex and forms a spectrum, comprising also individuals with non-binary gender identities. Examining gender stereotypes against lecturers with non-binary gender identities was beyond the scope of this research, but should be investigated in future work. Similarly, future research could investigate whether students hold stereotypes against lecturers on the basis of other characteristics (e.g., lecturers' sexual orientation, ethnicity, nationality, physical characteristics or lecturers with non-binary identities). It could be examined how numerous student characteristics (e.g., type of subject, age, ethnicity, country of origin, a social class) relate to the potential influence of gender stereotypes in their evaluations. These are important issues that were studied elsewhere. I focused on a lecturer's gender, but future research could incorporate these factors as well.

It could be explored how students interpret different SET item statements (other than those included in my studies), and whether potential stereotypes may influence how students rate lecturers on these different items. Future studies could also examine whether revised instructions for students (e.g., clearly defining item statements, outlining study phenomena that students should consider, instructions to contemplate questions carefully) may to some extent mitigate potential biases.

I explored how students interpreted and used Likert scales, but future research could also investigate how students interpret answer categories in different rating scales. For example, the use of answer categories that students pick to estimate the occurrence of certain teaching behaviours (e.g., rating an item statement, 'How often do lecturers make subject engaging?', with categories ranging from 'rarely' to 'very often') could be examined.

Future studies should continue to explore whether the use of rating scales may promote rather than reduce potential gender biases. The focus could be on potential gender differences between how students evaluate their "best" lecturers, non-specific items evaluating overall experience, and teaching behaviours related to engagement and feedback. Furthermore, written scenarios may have not activated potential gendered beliefs. Future research could explore the use of videos instead of scenarios. Furthermore, future research should further investigate the potential of using innovative methods such as NLP algorithms or Conversational Agents to obtain student feedback.

7.9 Conclusion

My research expanded knowledge on how students make their rating decisions, as well as interpret and use item statements and answer scale categories. I examined what teaching behaviours and inferred attitudes students considered and whether they applied gender schemata when evaluating university lecturers through rating scales. No previous research explored these issues applying the methodological concepts from the TPS-paradigm to scrutinise whether data generation under SETs meets the principles under which it would constitute measurement.

My findings revealed several problems with the application of SETs. These include extensive variations in students' interpretations and use of item statements and answer scale categories as well as large fields of meanings that participants constructed for each standardised item statement or types of reasons for picking the answer scale category, despite each participant, on average, considering only between one and two pieces of evidence. Participants also applied gendered schemas in some fewer cases.

Therefore, despite the prominence of SETs, data generation with SETs does not constitute measurement and SET results cannot be justifiably attributed to teaching quality of the evaluated lecturers and lack transparent and publicly interpretable quantitative meanings. Institutions, in general, attach disproportionate importance to data generated with SETs, in lieu of considering alternative, more holistic and informative methods of teaching evaluation or collecting students' views on teaching.

References

Abran, A., Desharnais, J. M., & Cuadrado-Gallego, J. J. (2012). Measurement and quantification are not the same: ISO 15939 and ISO 9126. *Journal of Software: Evolution and Process*, *24*(5), 585–601. https://doi.org/10.1002/smr.496

Adams, S., Bekker, S., Fan, Y., Gordon, T., Shepherd, L. J., Slavich, E., & Waters, D. (2022). Gender Bias in Student Evaluations of Teaching: 'Punish[ing] Those Who Fail To Do Their Gender Right.' *Higher Education*, *83*(4), 787–807. https://doi.org/10.1007/s10734-021-00704-9

Adisa, T. A., Harrison, M., Sani, K. F., Mingazova, D., & Kypuram, J. (2023). The National Student Survey and the 'customerization' of university students: a qualitative study of UK higher education. *Higher Education*, *86*(2), 449–466. https://doi.org/10.1007/s10734-022-00943-4

Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*, *40*(8), 1291–1315. https://doi.org/10.1002/smj.3015

Al-Maamari, F. (2015). Response Rate and Teaching Effectiveness in Institutional Student Evaluation of Teaching: A Multiple Linear Regression Study. *Higher Education Studies*, *5*(6), 9–20. https://doi.org/10.5539/hes.v5n6p9

American Psychological Association. (n.d.). *Sexism*. Retrieved November 16, 2020, from https://dictionary.apa.org/sexism

Andersen, K., & Miller, E. D. (1997). Gender and Student Evaluations of Teaching. *PS: Political Science & Politics*, *30*(2), 216–219. https://doi.org/10.2307/420499

Anderson, K. E. (1948). Measurement in Science. In Edward N. Zalta (Ed.), *School Science and Mathematics* (Vol. 48, Issue 6, pp. 429–432). The Metaphysics Research Lab Philosophy Department Stanford University Stanford, CA 94305-4115. https://doi.org/10.1111/j.1949-8594.1948.tb06596.x

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, *27*(2), 184–201. https://doi.org/10.1177/0739986304273707

APA. (2021). *The APA Dictionary of Psychology - Consumer Psychology*. American Psychological Association.

Arbuckle, J., & Williams, B. D. (2003a). Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations. *Sex Roles*, *49*(9–10), 507–516. https://doi.org/10.1023/A:1025832707002

Arbuckle, J., & Williams, B. D. (2003b). Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations. *Sex Roles*, *49*(9–10), 507–516. https://doi.org/10.1023/A:1025832707002

Arro, G. (2013). Peeking into Personality Test Answers: Inter- and Intraindividual Variety in Item Interpretations. *Integrative Psychological and Behavioral Science*, *47*(1), 56–76. https://doi.org/10.1007/s12124-012-9216-9

Arthur, L. (2020). Evaluating student satisfaction - restricting lecturer professionalism: outcomes of using the UK national student survey questionnaire for internal student evaluation of teaching. *Assessment and Evaluation in Higher Education*, *45*(3), 331–344. https://doi.org/10.1080/02602938.2019.1640863

Ashby, A., Richardson, J. T. E., & Woodley, A. (2011). National student feedback surveys in distance education: An investigation at the UK Open University. *Open Learning*, *26*(1), 5–25. https://doi.org/10.1080/02680513.2011.538560

Ashford, R. D., Brown, A. M., & Curtis, B. (2018). Substance use, recovery, and linguistics: The impact of word choice on explicit and implicit bias. *Drug and Alcohol Dependence*, *189*, 131–138. https://doi.org/10.1016/j.drugalcdep.2018.05.005

Baber, K. M., & Tucker, C. J. (2006). The social roles questionnaire: A new approach to measuring attitudes toward gender. *Sex Roles*, *54*(7–8), 459–467. https://doi.org/10.1007/s11199-006-9018-y

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, *48*(3), 193–210. https://doi.org/10.1080/03634529909379169

Bagby, R. M., Goldbloom, D. S., & Schulte, F. S. M. (2005). The Use of Standardized Rating Scales in Clinical Practice. In D. S. Goldbloom (Ed.), *Psychiatric Clinical*

Skills. PA: Elsivier Mosby. https://doi.org/10.1016/B978-0-323-03123-3.50007-7

Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2019). Is Man the Measure of All Things? A Social Cognitive Account of Androcentrism. *Personality and Social Psychology Review*, *23*(4), 307–331. https://doi.org/10.1177/1088868318782848

Baltes, B. B., & Parker, C. P. (2000). Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior and Human Decision Processes*, *82*(2), 237–267. https://doi.org/10.1006/obhd.2000.2897

Barbour, R. S. (2014). Quality of Data Analysis. In U. Flick (Ed.), *The SAGE Handbook of Qualitative Data Analysis* (pp. 496–509). London: Sage. https://doi.org/10.4135/9781446282243.n34

Barreto, M., & Ellemers, N. (2005). The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities. *European Journal of Social Psychology*, *35*(5), 633–642. https://doi.org/10.1002/ejsp.270

Barthes, R. (1967). *Elements of Semiology*. New York: Hill & Wang.

Basow, S. A. (1995). Student Evaluations of College Professors: When Gender Matters. *Journal of Educational Psychology*, *87*(4), 656–665. https://doi.org/10.1037/0022-0663.87.4.656

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles*, *43*(5–6), 407–417. https://doi.org/10.1023/a:1026655528055

Basow, S. A., & Martin, J. L. (2012). Bias in Student Evaluations. In M. E. Kite (Ed.), *Effective Evaluation of Teaching: A Guide for Faculty and Administrators* (pp. 40–49). Washington, DC: Society for the Teaching of Psychology. http://teachpsych.org/ebooks/evals2012/index.php

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, *30*(1), 25–35. https://doi.org/10.1111/j.1471-6402.2006.00259.x

Basow, S. A., & Silberg, N. T. (1987). Student Evaluations of College Professors: Are Female and Male Professors Rated Differently? *Journal of Educational Psychology*, *79*(3), 308–314. https://doi.org/10.1037/0022-0663.79.3.308

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior?

Perspectives on Psychological Science, *2*(4). https://doi.org/10.1111/j.1745-6916.2007.00051.x

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education*, *3*(4), 245–256. https://doi.org/10.1037/a0020763

Bedggood, R. E., & Pollard, R. J. (1999). Uses and misuses of student opinion surveys in eight Australian universities. *Australian Journal of Education*, *43*(2), 129–141. https://doi.org/10.1177/000494419904300203

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, *88*(4), 354–364. https://doi.org/10.1037/0033-295X.88.4.354

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bennett, R., & Kane, S. (2014a). Students' interpretations of the meanings of questionnaire items in the National Student Survey. *Quality in Higher Education*, *20*(2), 129–164. https://doi.org/10.1080/13538322.2014.924786

Bennett, R., & Kane, S. (2014b). Students' interpretations of the meanings of questionnaire items in the National Student Survey. *Quality in Higher Education*, *20*(2), 129–164. https://doi.org/10.1080/13538322.2014.924786

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, *74*(2), 170–179. https://doi.org/10.1037/0022-0663.74.2.170

Bennett, T. (2021). The effects of student-consumerism on discipline specific teaching practices: a comparison of education and law. *Journal of Further and Higher Education*, *45*(3), 417–432. https://doi.org/10.1080/0309877X.2020.1774050

Benton, S. L., & Cashin, W. E. (2014). Student Ratings of Instruction in College and University Courses. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 279–326). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-017-8005-6_7

Benz, C. R., & Blatt, S. J. (1996). Meanings underlying student ratings of faculty. *Review of Higher Education*, *19*(4), 411–433. https://doi.org/10.1353/rhe.1996.0016

Berger, R. (2015). Now I see it, now I don't: researcher's position and reflexivity in qualitative research. *Qualitative Research*, *15*(2), 219–234. https://doi.org/10.1177/1468794112468475

Berk, R. A. (2011). Top 20 Strategies to Increase the Online Response Rates of Student Rating Scales. *International Journal of Technology in Teaching and Learning*, *8*(2), 98–107.

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205–213. https://doi.org/10.1177/0022487105275904

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391. https://doi.org/10.1126/SCIENCE.AAH6524

Binderkrantz, A. S., Bisgaard, M., & Lassesen, B. (2022). Contradicting findings of gender bias in teaching evaluations: evidence from two experiments in Denmark. *Assessment and Evaluation in Higher Education*, *47*(8), 1345–1357. https://doi.org/10.1080/02602938.2022.2048355

Binning, J. F., Zaba, A. J., & Whattam, J. C. (1986). Explaining the Biasing Effects of Performance Cues in Terms of Cognitive Categorization. *Academy of Management Journal*, *29*(3), 521–535. https://doi.org/10.5465/256222

BIPM. (2008). JCGM 200:2008 International vocabulary of metrology - Basic and general concepts and associated terms (VIM). In *Bipm* (3rd Ed.). http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf

Block, D. (1998). Exploring interpretations of questionnaire items. *System*, *26*(3), 403–425. https://doi.org/10.1016/S0346-251X(98)00022-0

Bocher, M., Ulvrova, M., Arnould, M., Coltice, N., Mallard, C., Gerault, M., & Adenis, A. (2020). Drawing everyday sexism in academia: Observations and analysis of a community-based initiative. *Advances in Geosciences*, *53*, 15–31. https://doi.org/10.5194/adgeo-53-15-2020

Bono, F., De Craene, V., & Kenis, A. (2019). My best geographer's dress: bodies, emotions and care in early-career academia. *Geografiska Annaler, Series B: Human*

Geography, *101*(1), 21–32. https://doi.org/10.1080/04353684.2019.1568200

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., & Stark, P. (2016a). Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. In *ScienceOpen Research*. https://doi.org/10.14293/s2199-1006.1.sor-edu.aetbzc.v1

Boring, A., Ottoboni, K., & Stark, P. B. (2016b). *Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors.* http://blogs.lse.ac.uk/impactofsocialsciences/2016/02/04/student-evaluations-of-teaching-gender-bias/

Bourabain, D. (2021). Everyday sexism and racism in the ivory tower: The experiences of early career researchers on the intersection of gender and ethnicity in the academic workplace. *Gender, Work and Organization*, *28*(1), 248–267. https://doi.org/10.1111/gwao.12549

Bowen, G. A. (2008). Naturalistic inquiry and the saturation concept: A research note. *Qualitative Research*, *8*(1), 137–152. https://doi.org/10.1177/1468794107085301

Boysen, G. A. (2009). A review of experimental studies of explicit and implicit bias among counselors. *Journal of Multicultural Counseling and Development*, *37*(4), 240–249. https://doi.org/10.1002/j.2161-1912.2009.tb00106.x

Bradford, K., Pendergast, D., & Grootenboer, P. (2021). What is Meant By "Teacher Quality" in Research and Policy: A Systematic, Quantitative Literature Review. *Education Thinking*, *1*(1), 57–76. https://www.analytrics.org

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71–88. https://doi.org/10.1016/j.econedurev.2014.04.002

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. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Bray, J. H., & Howard, G. S. (1980). Interaction of teacher and student sex and sex role orientations and student evaluations of college instruction. *Contemporary Educational Psychology*, *5*(3), 241–248. https://doi.org/10.1016/0361-476X(80)90047-8

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of Personality and Social Psychology*, *47*(6), 1191–1205. https://doi.org/10.1037/0022-3514.47.6.1191

British Psychological Society. (1940). The Lancet. https://doi.org/10.1016/S0140-6736(00)72789-9

British Psychological Society. (2014). *Code of Human Research Ethics*.

Brown, A. L., Lee, J., & Collins, D. (2015). Does student teaching matter? Investigating pre-service teachers' sense of efficacy and preparedness. *Teaching Education*, *26*(1), 77–93. https://doi.org/10.1080/10476210.2014.957666

Bunce, L., Baird, A., & Jones, S. E. (2017). The student-as-consumer approach in higher education and its effects on academic performance. *Studies in Higher Education*, *42*(11), 1958–1978. https://doi.org/10.1080/03075079.2015.1127908

Burr, V. (2003). *Social constructionism: Second edition* (2nd Ed.). Routledge, Taylor & Francis Group.

Burr, V., & Dick, P. (2017). Social constructionism. In *The Palgrave Handbook of Critical Social Psychology* (pp. 59–80). Palgrave Macmillan, London, UK. https://doi.org/10.1057/978-1-137-51018-1_4

Cameron, R. (2011). An analysis of quality criteria for qualitative research. *25th ANZAM Conference*, *December*, 1–16.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carminati, L. (2018). Generalizability in Qualitative Research: A Tale of Two Traditions. *Qualitative Health Research*, *28*(13), 2094–2101. https://doi.org/10.1177/1049732318788379

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without

increasing actual learning. *Psychonomic Bulletin and Review*, *20*(6), 1350–1356. https://doi.org/10.3758/s13423-013-0442-z

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3), 409–432. https://doi.org/10.1086/653808

Carson, L. (2001). Gender relations in higher education: Exploring lecturers perceptions of student evaluations of teaching. *Research Papers in Education*, *16*(4), 337–358. https://doi.org/10.1080/02671520152731990

Carter, P. L., Skiba, R., Arredondo, M. I., & Pollock, M. (2017). You Can't Fix What You Don't Look At: Acknowledging Race in Addressing Racial Discipline Disparities. *Urban Education*, *52*(2), 207–235. https://doi.org/10.1177/0042085916660350

Carter, S. M., & Little, M. (2007). Justifying knowledge, justifying method, taking action: Epistemologies, methodologies, and methods in qualitative research. *Qualitative Health Research*, *17*(10), 1316–1328. https://doi.org/10.1177/1049732307306927

Cashin, W. E. (1995). Student Ratings of Teaching: The Research Revisited. No. In *IDEA Paper* (Issue 32). http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED402338%5Cnhttp://eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED402338%5Cnhttp://www.eric.ed.gov/ERICWebPortal/search/recordDetails.jsp?ERICExtSearch_S

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *71*(1), 17–33. https://doi.org/10.1080/00221546.2000.11780814

Chambers, D. W. (1983). Stereotypic images of the scientist: The draw-a-scientist test. *Science Education*, *67*(2), 255–265. https://doi.org/10.1002/sce.3730670213

Chapman, D. D., & Joines, J. A. (2017). Strategies for increasing response rates for online end-of-course evaluations. *International Journal of Teaching*, *29*(1), 47–60. http://www.isetl.org/ijtlhe/

Chávez, K., & Mitchell, K. M. W. (2020). Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity. *PS - Political Science and Politics*, *53*(2), 270–274. https://doi.org/10.1017/S1049096519001744

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education*, *28*(1), 71–88. https://doi.org/10.1080/02602930301683

Chen, Y. Y., Shek, D. T. L., & Bu, F. F. (2011). Applications of interpretive and constructionist research methods in adolescent research: Philosophy, principles and examples. In *International Journal of Adolescent Medicine and Health* (Vol. 23, Issue 2, pp. 129–139). https://doi.org/10.1515/IJAMH.2011.022

Ching, G. (2019). A literature review on the student evaluation of teaching. *Higher Education Evaluation and Development*, *12*(2), 63–84. https://doi.org/10.1108/heed-04-2018-0009

Cicourel, A. V. (1964). *Method and measurement in sociology*. New York: Free Press.

Clayson, D. E. (2005). Within-Class Variability in Student-Teacher Evaluations: Examples and Problems. *Decision Sciences Journal of Innovative Education*, *3*(1), 109–124. https://doi.org/10.1111/j.1540-4609.2005.00055.x

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn?: A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16–30. https://doi.org/10.1177/0273475308324086

Clayson, D. E. (2018). Student evaluation of teaching and matters of reliability. *Assessment and Evaluation in Higher Education*, *43*(4), 666–681. https://doi.org/10.1080/02602938.2017.1393495

Clayson, D. E., & Haley, D. A. (2011). Are Students Telling Us the Truth? A Critical Look at the Student Evaluation of Teaching. *Marketing Education Review*, *21*(2), 101–112. https://doi.org/10.2753/mer1052-8008210201

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*(2), 149–160. https://doi.org/10.1177/0273475306288402

Coaley, K. (2012). *An Introduction to Psychological Assessment and Psychometrics*. SAGE Publications Ltd. https://doi.org/10.4135/9781446221556

Coats, W. D., Swierenga, L., & Wickert, J. (1972). Student perceptions of teachers- a

factor analytic study. *Journal of Educational Research*, *65*(8), 357–360. https://doi.org/10.1080/00220671.1972.10884347

Cochran-Smith, M. (2021). Exploring teacher quality: international perspectives. *European Journal of Teacher Education*, *44*(3), 415–428. https://doi.org/10.1080/02619768.2021.1915276

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Science* (2nd ed.). Hillsdale, NJ: Erlbaum. http://repositorio.unan.edu.ni/2986/1/5624.pdf

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. *Review of Educational Research*, *51*(3), 281–309. https://doi.org/10.3102/00346543051003281

Collingridge, D. S., & Gantt, E. E. (2008). The quality of qualitative research. *American Journal of Medical Quality*, *23*(5), 389–395. https://doi.org/10.1177/1062860608320646

Collins, J., & Hebert, T. (2008). Race and gender images in psychology textbooks. *Race, Gender & Class*, *15*, 300–307.

Corcoran, K., & Mussweiler, T. (2010). The cognitive miser's perspective: Social comparison as a heuristic in self-judgements. *European Review of Social Psychology*, *21*(1), 78–113. https://doi.org/10.1080/10463283.2010.508674

Crano, W. D., & Prislin, R. (2006). Attitudes and persuasion. *Annual Review of Psychology*, *57*, 345–374. https://doi.org/10.1146/annurev.psych.57.102904.190034

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Crump, M. J. C., Price, P. C., Jhangiani, R., Chiang, I. A., & Leighton, D. C. (2018). *Research Methods for Psychology: Brooklyn College Edition* (3rd Ed.).

Curley, L. J., Munro, J., & Dror, I. E. (2022). Cognitive and human factors in legal layperson decision making: Sources of bias in juror decision making. *Medicine, Science and the Law*, *62*(3), 206–215. https://doi.org/10.1177/00258024221080655

Dardenne, B., Dumont, M., & Bollier, T. (2007). Insidious Dangers of Benevolent Sexism: Consequences for Women's Performance. *Journal of Personality and Social Psychology*, *93*(5), 764–779. https://doi.org/10.1037/0022-3514.93.5.764

Darling-Hammond, L. (2009). Recognizing and Enhancing Teacher Effectiveness. *The International Journal of Educational and Psychological Assessment*, *3*, 1–25.

Darling-Hammond, L. (2021). Defining teaching quality around the world. *European Journal of Teacher Education*, *44*(3), 295–308. https://doi.org/10.1080/02619768.2021.1919080

Davison, E., & Price, J. (2009). How do we rate? an evaluation of online student evaluations. *Assessment and Evaluation in Higher Education*, *34*(1), 51–65. https://doi.org/10.1080/02602930801895695

De la Fuente, L., Inmaculada de la Fuente, E., & García, J. (2003). Effects of pretrial Juror Bias, strength of evidence and deliberation process on Juror decisions: New validity evidence of the Juror Bias scale scores. *Psychology, Crime and Law*, *9*(2), 197–209. https://doi.org/10.1080/1068316031000116283

Dearden, H. T. (2014). How long is a metre? *Measurement and Control (United Kingdom)*, *47*(1), 26–27. https://doi.org/10.1177/0020294013517449

DeMarrais, K., & Lapan, S. D. (2003). *Foundations for research: Methods of inquiry in education and the social sciences*. Mahwah, NJ: L. Erlbaum Associates. https://doi.org/10.4324/9781410609373

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dill, D. (2009). Quality Assurance in Higher Education – Practices and Issues. In *International Encyclopedia of Education, Third Edition* (pp. 377–383). https://doi.org/10.1016/B978-0-08-044894-7.00833-2

Dobbin, F., & Kalev, A. (2018). Why Doesn't Diversity Training Work? The Challenge for Industry and Academia. *Anthropology Now*, *10*(2), 48–55. https://doi.org/10.1080/19428200.2018.1493182

Dodson, C. S., Darragh, J., & Williams, A. (2008). Stereotypes and Retrieval-Provoked Illusory Source Recollections. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(3), 460–477. https://doi.org/10.1037/0278-7393.34.3.460

Domanski, C. W. (2004). A biographical note on Max Friedrich (1856-1887), Wundt's first PhD student in experimental psychology. *Journal of the History of the Behavioral Sciences*, *40*(3), 311–317. https://doi.org/10.1002/jhbs.20022

DSouza, M. J. (2017). The Practice of Qualitative Research. *Qualitative Research in Organizations and Management: An International Journal*, *12*(3), 247–248. https://doi.org/10.1108/qrom-09-2016-1416

Dunegan, K. J., & Hrivnak, M. W. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education*, *27*(3), 280–303. https://doi.org/10.1177/1052562903027003002

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. England: Lawrence Erlbaum Associates, Inc.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: Professors' Biased Evaluations of Physics and Biology Post-Doctoral Candidates. *Sex Roles*, *82*(3–4), 127–141. https://doi.org/10.1007/s11199-019-01052-w

Eitzen, D. S., & Zinn, M. B. (2016). The De-Athleticization of Women: The Naming and Gender Marking of Collegiate Sport Teams. *Sociology of Sport Journal*, *6*(4), 362–370. https://doi.org/10.1123/ssj.6.4.362

El-Alayli, A., Hansen-Brown, A. A., & Ceynar, M. (2018). Dancing Backwards in High Heels: Female Professors Experience More Work Demands and Special Favor Requests, Particularly from Academically Entitled Students. *Sex Roles*, *79*(3–4), 136–150. https://doi.org/10.1007/s11199-017-0872-6

Elder-Vass, D. (2012). Towards a realist social constructionism. *Sociologia, Problemas e Praticas*, *70*, 9–24. https://doi.org/10.7458/SPP2012701208

Ellemers, N. (2018). Gender Stereotypes. *Annual Review of Psychology*, *69*(1), 275–298.

https://doi.org/10.1146/annurev-psych-122216-011719

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, *11*(1), 37–46. https://doi.org/10.1108/09684880310462074

Eppley, K. (2006). Defying insider-outsider categorization: One researcher's fluid complicated positioning on the insider-outsider continuum. *Forum: Qualitative Social Research*, *7*(3), 16.

Ewing, V. L., Sheehan, E. P., & Stukas, A. A. (2003). Student prejudice against gay male and lesbian lecturers. *Journal of Social Psychology*, *143*(5), 569–579. https://doi.org/10.1080/00224540309598464

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS ONE*, *14*(2), e0209749. https://doi.org/10.1371/journal.pone.0209749

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, *24*(2), 139–213. https://doi.org/10.1007/BF00991885

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*(6), 583–645. https://doi.org/10.1007/BF00992392

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II-Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*(2), 151–211. https://doi.org/10.1007/BF00992161

Finlay, L. (2002). "Outing" the researcher: The provenance, process, and practice of reflexivity. *Qualitative Health Research*, *12*(4), 531–545. https://doi.org/10.1177/104973202129120052

Fisher, A. N., Stinson, D. A., & Kalajdzic, A. (2019). Unpacking Backlash: Individual and Contextual Moderators of Bias against Female Professors. *Basic and Applied Social Psychology*, *41*(5), 305–325. https://doi.org/10.1080/01973533.2019.1652178

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, *20*(2), 303. https://doi.org/10.1086/209351

Fisk, S., Stolee, K. T., & Battestilli, L. (2020). A Lightweight Intervention to Decrease Gender Bias in Student Evaluations of Teaching. *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology, RESPECT 2020 - Proceedings*. https://doi.org/10.1109/RESPECT49803.2020.9272454

Fiske, S. T. (2000). Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology*, *30*(3), 299–322. https://doi.org/10.1002/(SICI)1099-0992(200005/06)30:3<299::AID-EJSP2>3.0.CO;2-F

Fiske, S. T. (2012). Warmth and competence: Stereotype content issues for clinicians and researchers. *Canadian Psychology*, *53*(1), 14–20. https://doi.org/10.1037/a0026054

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002a). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002b). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Fiske, S. T., & Taylor, S. E. (1991). Social cognition (2nd ed.). In *McGrawHill Series in Social Psychology*.

Fiske, S. T., & Taylor, S. E. (2013). *Social Cognition: From Brains to Culture* (2nd ed.). Sage Publications Ltd.

Fopp, R. (2007). Can social constructionism go too far? *Reshaping Australasian Housing Research: Refereed Papers and Presentations from the 2nd Australasian Housing Researchers' Conference 2007, AHRC 2007*, 20–22.

Frambach, J. M., van der Vleuten, C. P. M., & Durning, S. J. (2013). AM last page. Quality criteria in qualitative and quantitative research. *Academic Medicine : Journal*

of the Association of American Medical Colleges, *88*(4), 552. https://doi.org/10.1097/ACM.0b013e31828abf7f

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning*, *2001*(87), 85–100. https://doi.org/10.1002/tl.10001

Friedman, H. H., & Amoo, T. (1999). SSRN-id2333648. *Journal of Marketing Management*, *9*(3), 114–123.

Gair, S. (2012). Feeling their stories: Contemplating empathy, insider/outsider positionings, and enriching qualitative research. *Qualitative Health Research*, *22*(1), 134–143. https://doi.org/10.1177/1049732311420580

Gaiseanu, F. (2020). Attitude as an Expressible Info-Operational Reaction to a Perceived/Purposed Object/Objective. *International Journal on Neuropsychology and Behavioural Sciences (IJNBS)*, *1*(1), 12–16. https://doi.org/10.51626/ijnbs.2020.01.00002

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Research in Higher Education*, *53*(3), 353–374. https://doi.org/10.1007/s11162-011-9229-0

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708–724. https://doi.org/10.1037/0022-3514.78.4.708

Gelber, K., Brennan, K., Duriesmith, D., & Fenton, E. (2022). Gendered mundanities: gender bias in student evaluations of teaching in political science. *Australian Journal of Political Science*, *57*(2), 199–220. https://doi.org/10.1080/10361146.2022.2043241

Gelo, O., Braakmann, D., & Benetka, G. (2008). Quantitative and qualitative research: Beyond the debate. *Integrative Psychological and Behavioral Science*, *42*(4), 266–290. https://doi.org/10.1007/s12124-009-9107-x

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78.

https://doi.org/10.1016/j.paid.2016.06.069

Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology*, *70*(3), 491–512. https://doi.org/10.1037/0022-3514.70.3.491

Golafshani, N. (2003). The Qualitative Report Understanding Reliability and Validity in Qualitative Research Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, *8*(4), 597–606. http://nsuworks.nova.edu/tqr%5Cnhttp://nsuworks.nova.edu/tqr/vol8/iss4/6

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341–364. https://doi.org/10.1007/s11162-016-9429-8

Graves, A. L., Hoshino-Browne, E., & Lui, K. P. H. (2017). Swimming against the tide: Gender bias in the physics classroom. *Journal of Women and Minorities in Science and Engineering*, *23*(1), 15–36. https://doi.org/10.1615/JWomenMinorScienEng.2017013584

Guilford, J. P. (1954). *Psychometric methods* (2nd Ed.). New York, NY, US: McGraw-Hill.

Hagtvet, B. E., & Wold, A. H. (2003). On the dialogical basis of meaning: Inquiries into ragnar Rommetveit's writings on language, thought, and communication. *Mind, Culture, and Activity*, *10*(3), 186–204. https://doi.org/10.1207/s15327884mca1003_2

Hamilton, D. L. (1968). Personality Attributes Associated With Extreme Response Style. *Psychological Bulletin*, *69*(3), 192–203. https://doi.org/10.1037/h0025606

Hammarberg, K., Kirkman, M., & De Lacey, S. (2016). Qualitative research methods: When to use them and how to judge them. *Human Reproduction*, *31*(3), 498–501. https://doi.org/10.1093/humrep/dev334

Hammond, C. (2005). The wider benefits of adult learning: An illustration of the advantages of multi-method research. *International Journal of Social Research Methodology: Theory and Practice*, *8*(3), 239–255. https://doi.org/10.1080/13645570500155037

Hardy, I., Jakhelln, R., & Smit, B. (2021). The policies and politics of teachers' initial learning: the complexity of national initial teacher education policies. *Teaching*

Education, *32*(3), 286–308. https://doi.org/10.1080/10476210.2020.1729115

Hartmann, N. (1964). *Der Aufbau der realen Welt. Grundriss der allgemeinen Kategorienlehre [The Structure of the Real World. Outline of the General Theory of Categories]*. Berlin: Walter de Gruyter.

Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *Journal of Abnormal and Social Psychology*, *49*(1), 129. https://doi.org/10.1037/h0057880

Hatch, M. J., & Yanow, D. (2008). Methodology by metaphor: Ways of seeing in painting and research. *Organization Studies*, *29*(1), 23–44. https://doi.org/10.1177/0170840607086635

Heath, J. K., Weissman, G. E., Clancy, C. B., Shou, H., Farrar, J. T., & Dine, C. J. (2019). Assessment of gender-based linguistic differences in physician trainee evaluations of medical faculty using automated text mining. *JAMA Network Open*, *2*(5), e193520–e193520. https://doi.org/10.1001/jamanetworkopen.2019.3520

Heffernan, T. (2023). Abusive comments in student evaluations of courses and teaching: the attacks women and marginalised academics endure. *Higher Education*, *85*(1), 225–239. https://doi.org/10.1007/s10734-022-00831-x

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior*, *32*, 113–135. https://doi.org/10.1016/j.riob.2012.11.003

HESA. (2023). *No Title*. https://www.hesa.ac.uk/data-and-analysis/students/what-study

Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., Seidel, L. M., Zarbock, A., & Wenk, M. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education*, *52*(10), 1064–1072. https://doi.org/10.1111/medu.13627

Hideg, I., & Ferris, D. L. (2016). The compassionate sexist? How benevolent sexism promotes and undermines gender equality in the workplace. *Journal of Personality and Social Psychology*, *111*(5), 706–727. https://doi.org/10.1037/pspi0000072

Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Hinton, P. (2017). Implicit stereotypes and the predictive brain: cognition and culture in

"biased" person perception. *Palgrave Communications*, *3*(1), 1–9. https://doi.org/10.1057/palcomms.2017.86

Hoel, A., & Dahl, T. I. (2019). Why bother? Student motivation to participate in student evaluations of teaching. *Assessment and Evaluation in Higher Education*, *44*(3), 361–378. https://doi.org/10.1080/02602938.2018.1511969

Holland, E. P. (2019). Making sense of module feedback: accounting for individual behaviours in student evaluations of teaching. *Assessment and Evaluation in Higher Education*, *44*(6), 961–972. https://doi.org/10.1080/02602938.2018.1556777

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

HowManyofMe. (n.d.). *No Title*. Retrieved March 29, 2021, from https://web.archive.org/web/20220516130515/http://howmanyofme.com

Husu, L. (2019). Nordic Countries and the Nordic Region: Gender Research and Gender Studies in Northern Europe. In B. Kortendiek, B. Riegraf, & K. Sabisch (Eds.), *Handbuch Interdisziplinäre Geschlechterforschung. Geschlecht und Gesellschaft, Vol. 65*. Springer VS, Wiesbaden. https://doi.org/10.1007/978-3-658-12500-4_150-1

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1080/09332480.2005.10722754

Iqbal, I., Lee, J. D., Pearson, M. L., & Albon, S. P. (2016). Student and faculty perceptions of student evaluations of teaching in a Canadian pharmacy school. *Currents in Pharmacy Teaching and Learning*, *8*(2), 191–199. https://doi.org/10.1016/j.cptl.2015.12.002

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action.

Administrative Science Quarterly, *24*(4), 602. https://doi.org/10.2307/2392366

Johnson, M. A., Stevenson, R. M., & Letwin, C. R. (2018). A woman's place is in the… startup! Crowdfunder judgments, implicit bias, and the stereotype content model. *Journal of Business Venturing*, *33*(6), 813–831. https://doi.org/10.1016/j.jbusvent.2018.04.003

Johnson, S. K., Murphy, S. E., Zewdie, S., & Reichard, R. J. (2008). The strong, sensitive type: Effects of gender stereotypes and leadership prototypes on the evaluation of male and female leaders. *Organizational Behavior and Human Decision Processes*, *106*(1), 39–60. https://doi.org/10.1016/j.obhdp.2007.12.002

Kahlke, R. M. (2014). Generic qualitative approaches: Pitfalls and benefits of methodological mixology. *International Journal of Qualitative Methods*, *13*(1), 37–52. https://doi.org/10.1177/160940691401300119

Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kemmelmeier, M. (2016). Cultural differences in survey responding: Issues and insights in the study of response biases. *International Journal of Psychology*, *51*(6). https://doi.org/10.1002/ijop.12386

Kezar, A., & Talburt, S. (2004). Introduction: Questions of Research and Methodology. *Journal of Higher Education*, *75*(1), 1–6. https://doi.org/10.1080/00221546.2004.11778892

Khazan, E. S., Greenhaw, L., & Borden, J. B. (2019). Examining Gender Bias in Student Evaluations of Teaching for Graduate Teaching Assistants. *NACTA Journal*, *64*(2), 422–427. https://www.researchgate.net/publication/345178456

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex Role Stereotyping of College Professors: Bias in Students' Ratings of Instructors. *Journal of Educational Psychology*, *80*(3), 342–344. https://doi.org/10.1037/0022-0663.80.3.342

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide

No. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Kite, M. E., Subedi, P. C., & Bryant-Lees, K. B. (2015). Students' Perceptions of the Teaching Evaluation Process. *Teaching of Psychology*, *42*(4), 307–314. https://doi.org/10.1177/0098628315603062

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda Effect in Science Communication: An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest. *Science Communication*, *35*(5), 603–625. https://doi.org/10.1177/1075547012472684

Koenig, A. M., & Eagly, A. H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*(3), 371–392. https://doi.org/10.1037/a0037215

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2010). Student evaluations of teaching: Perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*(6), 623–636. https://doi.org/10.1080/13562517.2010.491911

Koocher, G. P., Norcross, J. C., & Greene, B. A. (2015). Ethical Principles of Psychologists and Code of Conduct. *Psychologists' Desk Reference*, *57*(12), 529–550. https://doi.org/10.1093/med:psych/9780199845491.003.0103

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). Foundations of Measurement. Volume 1. *Academic Press*, *New York. Lindley*,.

Lamarche, V. M., Seery, M. D., Kondrak, C. L., Saltsman, T. L., & Streamer, L. (2020). Clever girl: Benevolent sexism and cardiovascular threat. *Biological Psychology*, 107781. https://doi.org/10.1016/j.biopsycho.2019.107781

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The impact of gender on the evaluation of teaching: What we know and what we can do. *NWSA Journal*, *19*, 87–104. https://doi.org/Article

Lefever, S., Dal, M., & Matthíasdóttir, Á. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574–582. https://doi.org/10.1111/j.1467-8535.2006.00638.x

Lenton, A. P., Blair, I. V., & Hastie, R. (2001). Illusions of Gender: Stereotypes Evoke False Memories. *Journal of Experimental Social Psychology*, *37*(1), 3–14. https://doi.org/10.1006/jesp.2000.1426

Levine, R. V., West, L. J., & Reis, H. T. (1980). Perceptions of time and punctuality in the United States and Brazil. *Journal of Personality and Social Psychology*, *38*(4). https://doi.org/10.1037/0022-3514.38.4.541

Lincoln, Y., Lynham, S. ., & Guba, E. (2018). Paradigmatic Controversies, Contradictions, and Emerging Confluences, revisited. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (5th Ed.). Thousand Oaks, CA: Sage.

Lindahl, M. W., & Unger, M. L. (2010). Cruelty in Student Teaching Evaluations. *College Teaching*, *58*(3), 71–76. https://doi.org/10.1080/87567550903253643

Llorens, A., Tzovara, A., Bellier, L., Bhaya-Grossman, I., Bidet-Caulet, A., Chang, W. K., Cross, Z. R., Dominguez-Faus, R., Flinker, A., Fonken, Y., Gorenstein, M. A., Holdgraf, C., Hoy, C. W., Ivanova, M. V., Jimenez, R. T., Jun, S., Kam, J. W. Y., Kidd, C., Marcelle, E., … Dronkers, N. F. (2021). Gender bias in academia: A lifetime problem that needs solutions. *Neuron*, *109*(13), 2047–2074. https://doi.org/10.1016/j.neuron.2021.06.002

Lowenthal, P., Bauer, C., & Chen, K. Z. (2015). Student Perceptions of Online Learning: An Analysis of Online Course Evaluations. *American Journal of Distance Education*, *29*(2), 85–97. https://doi.org/10.1080/08923647.2015.1023621

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1989). Foundations of Measurement - Geometrical, Threshold, and Probabilistic Representations. In *Dover* (Vol. 2). Academic Press.

Lundmann, L., & Villadsen, J. W. (2016). Qualitative variations in personality inventories: subjective understandings of items in a personality inventory. *Qualitative Research in Psychology*, *13*(2), 166–187. https://doi.org/10.1080/14780887.2015.1134737

MacDougall, M., Riley, S. C., Cameron, H. S., & McKinstry, B. (2008). Halos and Horns in the Assessment of Undergraduate Medical Students: A Consistency-Based Approach. *Journal of Applied Quantitative Methods*, *3*(2), 116–128.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly: Management Information Systems*, *35*(2), 293–334. https://doi.org/10.2307/23044045

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education*, *40*(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, *92*(1), 239–255. https://doi.org/10.1348/000712601162059

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of Mind but Back in Sight: Stereotypes on the Rebound. *Journal of Personality and Social Psychology*, *67*(5), 808–817. https://doi.org/10.1037/0022-3514.67.5.808

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and Letters of Recommendation for Academia: Agentic and Communal Differences. *Journal of Applied Psychology*, *94*(6), 1591–1599. https://doi.org/10.1037/a0016539

Maltby, J., Day, L., & Macaskill, A. (2010). *Personality, Individual Differences and Intelligence* (4th Ed.). Pearson. http://books.google.com/books?id=65IJkMkO2XMC&pgis=1

Mamica, Ł., & Mazur, B. (2020). Expectations versus reality: What matters to students of economics vs. what they receive from universities? *Education Sciences*, *10*(1), 2. https://doi.org/10.3390/educsci10010002

Marecek, J., Crawford, M., & Popp, D. (2004). On the Construction of Gender, Sex, and Sexualities. In A.H. Eagly, A. E. Beall, & R. J. Sternberg (Eds.), *The psychology of gender* (2nd e., pp. 192–216). NY: Guilford.

Mari, L. (2000). Beyond the representational viewpoint: A new formalization of measurement. *Measurement: Journal of the International Measurement Confederation*, *27*(2), 71–84. https://doi.org/10.1016/S0263-2241(99)00055-X

Mari, L., Carbone, P., Giordani, A., & Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Studies in History and Philosophy*

of Science Part A, *65–66*, 46–56. https://doi.org/10.1016/j.shpsa.2017.08.001

Mari, L., Maul, A., Irribarra, D. T., & Wilson, M. (2016). A meta-structural understanding of measurement. *Journal of Physics: Conference Series*, *772*(1), 012009. https://doi.org/10.1088/1742-6596/772/1/012009

Marimon, F., Mas-Machuca, M., & Berbegal-Mirabent, J. (2020). Fulfilment of expectations on students' perceived quality in the Catalan higher education system. *Total Quality Management and Business Excellence*, *31*(5–6), 483–502. https://doi.org/10.1080/14783363.2018.1433027

Marks, R. B. (2000). Determinants of Student Evaluations of Global Measures of Instructor and Course Value. *Journal of Marketing Education*, *22*(2), 108–119. https://doi.org/10.1177/0273475300222005

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology*, *71*(2), 149–160. https://doi.org/10.1037/0022-0663.71.2.149

Martin, L. L. (2016). Gender, teaching evaluations, and professional success in political science. *PS - Political Science and Politics*, *49*(2), 313–319. https://doi.org/10.1017/S1049096516000275

Maruli, S. (2014). Quality in Teaching : A review of literature. *International Journal of Education and Research*, *2*(12), 193–200.

Mays, N., & Pope, C. (2000). Qualitative research in health care: Assessing quality in qualitative research. *British Medical Journal*, *320*(7226), 50–52. https://doi.org/10.1136/bmj.320.7226.50

McCullough, B. D., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation and Research in Education*, *24*(3), 183–202. https://doi.org/10.1080/09500790.2011.603411

McGrath, J. E. (1981). Dilemmatics: The Study of Research Choices and Dilemmas. *American Behavioral Scientist*, *25*(2), 179–210. https://doi.org/10.1177/000276428102500205

McNatt, D. B. (2010). Negative reputation and biased student evaluations of teaching:

Longitudinal results from a naturally occurring experiment. *Academy of Management Learning and Education*, *9*(2), 225–242. https://doi.org/10.5465/AMLE.2010.51428545

McNatt, D. B. (2022). The biasing impact of positive instructor reputation on student evaluations of teaching. *International Journal of Management Education*, *20*(1), 100607. https://doi.org/10.1016/j.ijme.2022.100607

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Menegatti, M., & Rubini, M. (2017). Gender Bias and Sexism in Language. In H. Giles & J. Harwood (Eds.), *The Oxford encyclopedia of intergroup communication (Volume 1)* (pp. 1–25). Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.470

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535–566. https://doi.org/10.1093/jeea/jvx057

Menyhart, O., Weltz, B., & Gyorffy, B. (2021). Multipletesting.com: A tool for life science researchers for multiple hypothesis testing correction. *PLoS ONE*, *16*(6), e0245824. https://doi.org/10.1371/journal.pone.0245824

Mero, N. P., Anna, A. L., & Motowidlo, S. J. (2003). Effects of Accountability on Rating Behavior and Rater Accuracy. *Journal of Applied Social Psychology*, *33*(12), 2493–2514. https://doi.org/10.1111/j.1559-1816.2003.tb02777.x

Merritt, D. J. (2008). Bias, the Brain, and Student Evaluations of Teaching. *St John's Law Review*, *82*, 235–287. https://doi.org/10.2139/ssrn.963196

Migiro, S. O., & Magangi, B. a. (2011). Mixed methods : A review of literature and the future of the new research paradigm. *African Journal of Business Management*, *5*(10), 3757–3764. https://doi.org/10.5897/AJBM09.082

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. 2nd ed. Thousand Oaks, CA: Sage.

Miller, J. A., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, *28*(4), 283–298. https://doi.org/10.2307/1318580

Miller, V. A., Reynolds, W. W., Ittenbach, R. F., Luce, M. F., Beauchamp, T. L., & Nelson, R. M. (2009). Challenges in measuring a new construct: Perception of voluntariness for research and treatment decision making. *Journal of Empirical Research on Human Research Ethics*, *4*(3), 21–31. https://doi.org/10.1525/jer.2009.4.3.21

Mitchell, K. M. W., & Martin, J. (2018). Gender Bias in Student Evaluations. *PS - Political Science and Politics*, *51*(3), 648–652. https://doi.org/10.1017/S104909651800001X

Molinari, C., Lundhal, S., & Shanderson, L. (2019). The culturally competent and inclusive leader. In G. L. Rubino, J. S. Esparza, & Y. Chassiakos (Eds.), *New leadership for today's health care professionals* (2nd ed., pp. 49–67). Jones & Bartlett.

Moore, R. (1990). Student Evaluations of Teaching. *American Biology Teacher*, *52*(5), 260–262. https://doi.org/10.2307/4449104

Morgenroth, T., & Ryan, M. K. (2018). Gender in a Social Psychology Context. In O. Braddick (Ed.), *Oxford Research Encyclopedia of Psychology*. New York: Oxford University Press. https://doi.org/10.1093/acrefore/9780190236557.013.309

Morgenroth, T., & Ryan, M. K. (2021). The Effects of Gender Trouble: An Integrative Theoretical Framework of the Perpetuation and Disruption of the Gender/Sex Binary. *Perspectives on Psychological Science*, *16*(6), 1113–1142. https://doi.org/10.1177/1745691620902442

Morgenroth, T., Sendén, M. G., Lindqvist, A., Renström, E. A., Ryan, M. K., & Morton, T. A. (2021). Defending the Sex/Gender Binary: The Role of Gender Identification and Need for Closure. *Social Psychological and Personality Science*, *12*(5), 731–740. https://doi.org/10.1177/1948550620937188

Morley, D. (2014). Assessing the reliability of student evaluations of teaching: Choosing the right coefficient. *Assessment and Evaluation in Higher Education*, *39*(2), 127–139. https://doi.org/10.1080/02602938.2013.796508

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation*, *38*(1), 15–20. https://doi.org/10.1016/j.stueduc.2012.01.001

Morris, C., Hinton-Smith, T., Marvell, R., & Brayson, K. (2022). Gender back on the

agenda in higher education: perspectives of academic staff in a contemporary UK case study. *Journal of Gender Studies*, *31*(1), 101–113. https://doi.org/10.1080/09589236.2021.1952064

Morse, J. M. (1991). Approaches to Qualitative-Quantitative Methodological Triagulation. *Nursing Research*, *40*, 120–125.

Morse, J. M. (2010). Procedures and Practice of Mixed Method Design. Maintaining control, rigor, and complexity. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social & Behavioral Research* (2nd ed., pp. 339–352). Thousand Oaks, CA: Sage.

Mortenson, K. G., & Sathe, R. S. (2017). A case study of group processes and student evaluation of teaching. *Accounting Education*, *26*(1), 28–53. https://doi.org/10.1080/09639284.2016.1274908

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and Consequences of Halo Error: A Critical Analysis. *Journal of Applied Psychology*, *78*(2). https://doi.org/10.1037/0021-9010.78.2.218

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox Lecture. *Academic Medicine*, *48*(7), 630–635. https://doi.org/10.1097/00001888-197307000-00003

National Student Survey. (n.d.). *What is the National Student Survey (NSS)?* Retrieved June 5, 2023, from https://thestudentsurvey.com/faqs

National Student Survey. (2019). *About the NSS.* https://web.archive.org/web/20200214053053/https://www.thestudentsurvey.com/institutions.php

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports*, *78*(3), 1363–1372. https://doi.org/10.2466/pr0.1996.78.3c.1363

Nelson, T. E., Acker, M., & Manis, M. (1996). Irrepressible stereotypes. *Journal of Experimental Social Psychology*, *32*(1), 13–38. https://doi.org/10.1006/jesp.1996.0002

Newell, D. B., & Tiesinga, E. (2019). The International System of Units (SI). *NIST Special Publication*, *330*, 1–138. https://doi.org/10.6028/NIST.SP.330-2019

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education*, *33*(3), 301–314. https://doi.org/10.1080/02602930701293231

O'Meara, K. A., Kuvaeva, A., Nyunt, G., Waugaman, C., & Jackson, R. (2017). Asked More Often: Gender Differences in Faculty Workload in Research Universities and the Work Interactions That Shape Them. *American Educational Research Journal*, *54*(6), 1154–1186. https://doi.org/10.3102/0002831217716767

Oakes, P. J., & Turner, J. C. (1990). Is limited information processing capacity the cause of social stereotyping? *European Review of Social Psychology*, *1*(1), 111–135. https://doi.org/10.1080/14792779108401859

Office for Students. (2019). *What Is the TEF?* www.officeforstudents.org.uk/advice-and-guidance/teaching/what-is-the-tef

Omi, Y. (2012). Tension Between the Theoretical Thinking and the Empirical Method: Is it an Inevitable Fate for Psychology? *Integrative Psychological and Behavioral Science*, *46*(1). https://doi.org/10.1007/s12124-011-9185-4

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*(1), 113–160. https://doi.org/10.3102/0002831206298169

Özgümüs, A., Rau, H. A., Trautmann, S. T., & König-Kersting, C. (2020). Gender Bias in the Evaluation of Teaching Materials. *Frontiers in Psychology*, *11*, 1074. https://doi.org/10.3389/fpsyg.2020.01074

Peimani, N., & Kamalipour, H. (2021). Online education and the covid-19 outbreak: A case study of online teaching during lockdown. *Education Sciences*, *11*(2), 1–16. https://doi.org/10.3390/educsci11020072

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots

in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. https://doi.org/10.1002/cae.22326

Perloff, R. M. (2020). The Dynamics of Persuasion: In *Communication and Attitudes in the Twenty-First Century* (7th Ed.). New York: Routledge. https://doi.org/10.4324/9780429196959

Peterson, D. A. M., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PLoS ONE*, *14*(5), e0216241-. https://doi.org/10.1371/journal.pone.0216241

Peterson, S. B., & Kroner, T. (1992). Gender Biases in Textbooks for Introductory Psychology and Human Development. *Psychology of Women Quarterly*, *16*(1), 17–36. https://doi.org/10.1111/j.1471-6402.1992.tb00237.x

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Phillips, D. E. (1932). What is scientific? *Journal of Educational Psychology*, *23*(4), 299–308. https://doi.org/10.1037/h0074212

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Podsakoff Mackenzie Lee Podzakoff JAP 2003 Common method biases. *Journal of Applied Psychology*, *88*(5), 879–903.

Polancos, D. T., Cinches, M. F. C., & Ortiz, R. O. (2013). The Reliability and Validity of Student Assessment of Teacher Performance (SATP) Scale: A Confirmatory Factor Analysis. *Liceo Journal of Higher Education Research*, *9*(1), 20–44. https://doi.org/10.7828/ljher.v9i1.638

Potvin, G., Hazari, Z., Tai, R. H., & Sadler, P. M. (2009). Unraveling bias from student evaluations of their high school science teachers. *Science Education*, *93*(5), 827–845. https://doi.org/10.1002/sce.20332

Poucher, Z. A., Tamminen, K. A., Caron, J. G., & Sweet, S. N. (2020). Thinking through and designing qualitative research studies: a focused mapping review of 30 years of qualitative research in sport psychology. *International Review of Sport and Exercise*

Psychology, *13*(1), 163–186. https://doi.org/10.1080/1750984X.2019.1656276

Priede, C., & Farrall, S. (2011). Comparing results from different styles of cognitive interviewing: "verbal probing" vs. "thinking aloud." *International Journal of Social Research Methodology*, *14*(4), 271–287. https://doi.org/10.1080/13645579.2010.523187

Protogerou, C., & Hagger, M. S. (2020). A checklist to assess the quality of survey studies in psychology. *Methods in Psychology*, *3*, 100031. https://doi.org/10.1016/j.metip.2020.100031

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Reinsch, R. W., Goltz, S. M., & Hietapelto, A. B. (2020). Student Evaluations and the Problem of Implicit Bias. *Journal of College and University Law*, *45*(1), 114.

Renström, E. A., Gustafsson Sendén, M., & Lindqvist, A. (2021). Gender Stereotypes in Student Evaluations of Teaching. *Frontiers in Education*, *5*, 280. https://doi.org/10.3389/feduc.2020.571287

Reupert, A., Maybery, D., Patrick, K., & Chittleborough, P. (2009). The Importance of Being Human: Instructor's Personal Presence in Distance Programs. *International Journal of Teaching and Learning in Higher Education*, *21*(1), 47–56. http://www.isetl.org/ijtlhe/

Rivera, L. A. (2017). When Two Bodies Are (Not) a Problem: Gender and Relationship Status Discrimination in Academic Hiring. *American Sociological Review*, *82*(6), 1111–1138. https://doi.org/10.1177/0003122417739294

Roberts, L. D., & Allen, P. J. (2015). Exploring ethical issues associated with using online surveys in educational research. *Educational Research and Evaluation*, *21*(2), 95–108. https://doi.org/10.1080/13803611.2015.1024421

Robertson, S. I. (2004). Student perceptions of student perception of module questionnaires: Questionnaire completion as problem solving. *Assessment and Evaluation in Higher Education*, *29*(6), 663–679. https://doi.org/10.1080/0260293042000227218

Robinson, O. C. (2022). Conducting Thematic Analysis on Brief Texts: The Structured Tabular Approach. *Qualitative Psychology*, *9*(2), 194–208. https://doi.org/10.1037/qup0000189

Robinson, O. C., & Smith, J. A. (2010). Investigating the form and dynamics of crisis episodes in early adulthood: The application of a composite qualitative method. *Qualitative Research in Psychology*, *7*(2), 170–191. https://doi.org/10.1080/14780880802699084

Rose, V. (2005). Assessing consumer ratings of quality in general practice needs more than just rating scales. *Health Issues*, *83*, 18–21.

Rosenbaum, P. J., & Valsiner, J. (2011). The un-making of a method: From rating scales to the study of psychological processes. *Theory & Psychology*, *21*(1), 47–65. https://doi.org/10.1177/0959354309352913

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, *48*(1), 165–179. https://doi.org/10.1016/j.jesp.2011.10.008

Sabri, D. (2013). Student Evaluations of Teaching as "Fact-Totems": The Case of the UK National Student Survey. *Sociological Research Online*, *18*(4), 148–157.

Saini, A. (2017). *Inferior: How science got women wrong and the new research that's rewriting the story*. Boston, MA: Beacon Press.

Sargeant, J. (2012). Qualitative Research Part II: Participants, Analysis, and Quality Assurance. *Journal of Graduate Medical Education*, *4*(1), 1–3. https://doi.org/10.4300/jgme-d-11-00307.1

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*(5), 1248–1264. https://doi.org/10.5465/1556348

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, *57*(7–8), 509–514. https://doi.org/10.1007/s11199-007-9291-4

Schwarz, N. (2008). Attitudes and attitude change. In W. D. Crano & R. Prislin (Eds.), *Attitudes and Attitude Change* (pp. 1–439). https://doi.org/10.4324/9780203838068

Shadreck, M., & Isaac, M. (2012). Science teacher quality and effectiveness: Gweru Urban junior secondary school students' points of view. *Asian Social Science*, *8*(8),

160–165. https://doi.org/10.5539/ass.v8n8p160

Sherman, J. W., Groom, C. J., Ehrenberg, K., & Klauer, K. C. (2003). Bearing false witness under pressure: Implicit and explicit components of stereotype-driven memory distortions. *Social Cognition*, *21*(3), 213–246. https://doi.org/10.1521/soco.21.3.213.25340

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment and Evaluation in Higher Education*, *25*(4), 397–405. https://doi.org/10.1080/713611436

Shweder, R. A. (1975). How relevant is an individual difference theory of personality? *Journal of Personality*, *43*(3), 455–484. https://doi.org/10.1111/j.1467-6494.1975.tb00716.x

Shweder, R. A., Casagrande, J. B., Fiske, D. W., Greenstone, J. D., Heelas, P., & Lancy, D. F. (1977). Likeness and Likelihood in Everyday Thought: Magical Thinking in Judgments About Personality [and Comments and Reply]. *Current Anthropology*, *18*(4), 637–658. https://doi.org/10.1086/201974

Sigurdardottir, M. S., Rafnsdottir, G. L., Jónsdóttir, A. H., & Kristofersson, D. M. (2023). Student evaluation of teaching: gender bias in a country at the forefront of gender equality. *Higher Education Research and Development*, *42*(4), 954–967. https://doi.org/10.1080/07294360.2022.2087604

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simpson, P. M., & Siguaw, J. A. (2000). Student Evaluations of Teaching: An Exploratory Study of the Faculty Response. *Journal of Marketing Education*, *22*(3), 199–213. https://doi.org/10.1177/0273475300223004

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin*, *26*(11), 1329–1342. https://doi.org/10.1177/0146167200263002

Sinha, J. K., & Sundaram, S. (1962). A Technique for the Measurement of Q. *IETE*

Journal of Research, *8*(2), 79–82. https://doi.org/10.1080/03772063.1962.11486337

Slaney, K. L., & Garcia, D. A. (2015). Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical and Philosophical Psychology*, *35*(4), 244–259. https://doi.org/10.1037/teo0000025

Slate, J. R., LaPrairie, K., Schulte, D. P., & Onwuegbuzie, A. J. (2009). A mixed analysis of college students' best and poorest college professors. *Issues in Educational Research*, *19*(1), 61–78.

Smith, D. G., Rosenstein, J. E., & Nikolov, M. C. (2018). The Different Words We Use to Describe Male and Female Leaders. *Harvard Business Review*, 1–8. https://hbr.org/2018/05/the-different-words-we-use-to-describe-male-and-female-leaders

Smith, J. S., & Wertlieb, E. C. (2005). Do First-Year College Students' Expectations Align with their First-Year Experiences? *NASPA Journal*, *42*(2), 153–174. https://doi.org/10.2202/1949-6605.1470

Snoek, M. (2021). Educating quality teachers: how teacher quality is understood in the Netherlands and its implications for teacher education. *European Journal of Teacher Education*, *44*(3), 309–327. https://doi.org/10.1080/02619768.2021.1931111

Social Security Administration. (n.d.). *Popular Baby Names by Decade*. Retrieved March 29, 2021, from https://www.ssa.gov/oact/babynames/decades

Spector, P. E. (1992). *Summated rating scale construction: An Introduction. Quantitative Applications in the Social Sciences*. Newbury Park: Sage Publications.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education*, *27*(5), 397–409. https://doi.org/10.1080/0260293022000009285

Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation*, *54*, 43–49. https://doi.org/10.1016/j.stueduc.2016.12.003

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles*, *53*(11–12), 779–793.

https://doi.org/10.1007/s11199-005-8292-4

Sproule, R. (2000). Student Evaluation of Teaching: Methodological Critique. *Education Policy Analysis Archives*, *8*, 50. https://doi.org/10.14507/epaa.v8n50.2000

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside Bias, Rational Thinking, and Intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264. https://doi.org/10.1177/0963721413480174

Stark, P., & Freishtat, R. (2014a). An Evaluation of Course Evaluations. *ScienceOpen Research*. https://doi.org/10.14293/s2199-1006.1.sor-edu.aofrqa.v1

Stark, P., & Freishtat, R. (2014b). An Evaluation of Course Evaluations. *ScienceOpen Research*. https://doi.org/10.14293/s2199-1006.1.sor-edu.aofrqa.v1

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*(7–8), 509–528. https://doi.org/10.1023/A:1018839203698

Stepan-Norris, J., & Kerrissey, J. (2016). Enhancing Gender Equity in Academia. *Sociological Perspectives*, *59*(2), 225–245. https://doi.org/10.1177/0731121415582103

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Strage, A. (2008). Traditional and Non-Traditional College Students' Descriptions of the "Ideal" Professor and the "Ideal" Course and Perceived Strengths and Limitations. *College Student Journal*, *42*(1), 225.

Stroebe, W. (2016). Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. *Perspectives on Psychological Science*, *11*(6), 800–816. https://doi.org/10.1177/1745691616650284

Su, F., & Wood, M. (2012). What makes a good university lecturer? Students' perceptions of teaching excellence. *Journal of Applied Research in Higher Education*, *4*(2), 142–155. https://doi.org/10.1108/17581181211273110

Šula, T. (2018). Thematic analysis. In C. Willig & W. S. Rogers (Eds.), *Ambient media in advertising : importance of design in ambient media creation* (2nd ed., pp. 66–78).

London: Sage. https://doi.org/10.7441/978-80-7454-682-2_4

Surratt, C. K., & Desselle, S. P. (2007). Pharmacy students' perceptions of a teaching evaluation process. *American Journal of Pharmaceutical Education, 71*(1), 6. https://doi.org/10.5688/aj710106

Swindells, B. (1975). Centenary of the Convention of the Metre. *Platinum Metals Review*, *19*(3), 110–113.

Tagomori & Bishop. (1994). Content analysis of evaluation instruments used for student evaluation of classroom teaching performhnce in higher education. *American Educational Research Association*, *22*(637), 36.

Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass*, *8*(12), 1159–1173. https://doi.org/10.1111/phc3.12089

Tam, K. Y., Heng, M. A., & Jiang, G. (2009). What undergraduate students in China say about their professors' teaching. *Teaching in Higher Education*, *14*(2), 147–159. https://doi.org/10.1080/13562510902757179

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research & Development in Education*, *28*(3), 169–173.

Taylor, S. E. (1981). The Interface of Cognitive and Social Psychology. In H. J. Harvey (Ed.), *Cognition, Social Behavior, and the Environment* (pp. 189–211).

Taylor, S. P. (2021). Assessing Critical Realism Vs Social Constructionism & Social Constructivism for a Social Housing Research Study. In Hus & Vlasta (Eds.), *Selected Topics in Humanities and Social Sciences Vol. 3*. BP International, London, UK. https://doi.org/10.9734/bpi/sthss/v3/1736c

TEF. (2016). *Teaching Excellence Framework: Technical Consultation for Year Two* (Issue May). https://www.gov.uk/government/consultations/teaching-excellence-framework-year-2-technical-consultation

Terkik, A., Prud'hommeaux, E., Alm, C. O., Homan, C. M., & Franklin, S. (2016). Analyzing gender bias in student evaluations. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers.*

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic Analysis. In C. Willig

& W. S. Rogers (Eds.), *The SAGE Handbook of Qualitative Research in Psychology* (2nd ed., pp. 17–36). London: Sage. https://doi.org/10.4135/9781526405555.n2

Thiel, J. (2019). The UK National Student Survey: An amalgam of discipline and neo-liberal governmentality. *British Educational Research Journal*, *45*(3), 538–553. https://doi.org/10.1002/berj.3512

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1). https://doi.org/10.1037/h0071663

Tikrity, M. A. (2023). *Defining and Measuring Teaching Quality. Article. Smart Teaching System*. https://doi.org/DOI: 10.13140/RG.2.2.20516.76162

Toftness, A. R., Carpenter, S. K., Geller, J., Lauber, S., Johnson, M., & Armstrong, P. I. (2018). Instructor fluency leads to higher confidence in learning, but not better learning. *Metacognition and Learning*, *13*(1), 1–14. https://doi.org/10.1007/s11409-017-9175-0

Tok, Ş. (2010). The problems of teacher candidate's about teaching skills during teaching practice. *Procedia - Social and Behavioral Sciences*, *2*(2), 4142–4146. https://doi.org/10.1016/j.sbspro.2010.03.654

Tomas, C., & Jessop, T. (2019). Struggling and juggling: a comparison of student assessment loads across research and teaching-intensive universities. *Assessment and Evaluation in Higher Education*, *44*(1), 1–10. https://doi.org/10.1080/02602938.2018.1463355

Towers, E., Rushton, E. A. C., Gibbons, S., Steadman, S., Brock, R., Cao, Y., Finesilver, C., Jones, J., Manning, A., Marshall, B., & Richardson, C. (2023). The "problem" of teacher quality: exploring challenges and opportunities in developing teacher quality during the Covid-19 global pandemic in England. *Educational Review*, 1–17. https://doi.org/10.1080/00131911.2023.2184771

Turner, J. C. (1991). *Social Influence*. Open University Press.

Turner, S. F., Cardinal, L. B., & Burton, R. M. (2017). Research Design for Mixed Methods: A Triangulation-based Framework and Roadmap. *Organizational Research Methods*, *20*(2), 243–267. https://doi.org/10.1177/1094428115610808

Tzanakou, C. (2017). Dual career couples in academia, international mobility and dual career services in Europe. *European Educational Research Journal*, *16*(2–3), 298–

312. https://doi.org/10.1177/1474904116683185

Uher, J. (2013). Personality Psychology: Lexical Approaches, Assessment Methods, and Trait Concepts Reveal Only Half of the Story-Why it is Time for a Paradigm Shift. *Integrative Psychological and Behavioral Science*, *47*(1), 1–55. https://doi.org/10.1007/s12124-013-9230-6

Uher, J. (2014a). Developing "Personality" Taxonomies: Metatheoretical and Methodological Rationales Underlying Selection Approaches, Methods of Data Generation and Reduction Principles. *Integrative Psychological and Behavioral Science*, *49*(4), 531–589. https://doi.org/10.1007/s12124-014-9280-4

Uher, J. (2014b). Fundamental challenges of contemporary "personality" research: Comment on "Personality from a cognitive-biological perspective" by Y. Neuman. *Physics of Life Reviews*, *11*(4), 695–696. https://doi.org/10.1016/j.plrev.2014.10.005

Uher, J. (2015a). Exploring the workings of the psyche: Metatheoretical and methodological foundations. In J. Valsiner, G. Marsico, N. Chaudhary, T. Sato, & V. Dazzani (Eds.), *Psychology as the Science of Human Being: The Yokohama Manifesto* (pp. 299–324). Cham, Springer International. https://doi.org/10.1007/978-3-319-21094-0_18

Uher, J. (2015b). Agency enabled by the psyche: Explorations using the transdisciplinary philosophy-of-science paradigm for research on individuals. In *Constraints of Agency: Explorations of Theory in Everyday Life* (pp. 175–228). Springer International Publishing. https://doi.org/10.1007/978-3-319-10130-9_13

Uher, J. (2015c). Conceiving "personality": Psychologist's challenges and basic fundamentals of the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals. *Integrative Psychological and Behavioral Science*, *49*(3), 398–458. https://doi.org/10.1007/s12124-014-9283-1

Uher, J. (2017). *How We Judge Others' Personality: gender, ethnicity and questionnaires.* http://www.lse.ac.uk/Events/2017/06/20170608t1830vOT/How-We-Judge-Others

Uher, J. (2018a). Taxonomic models of individual differences: A guide to transdisciplinary approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1744). https://doi.org/10.1098/rstb.2017.0171

Uher, J. (2018b). Quantitative data from rating scales: An epistemological and methodological enquiry. *Frontiers in Psychology*, *9*(2599), 1–27. https://doi.org/10.3389/fpsyg.2018.02599

Uher, J. (2018c). The Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals: Foundations for the Science of Personality and Individual Differences. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE Handbook of Personality and Individual Differences: Volume I: The Science of Personality and Individual Differences* (pp. 84–109). London, UK: Sage. https://doi.org/10.4135/9781526451163.n4

Uher, J. (2019). Data generation methods across the empirical sciences: differences in the study phenomena's accessibility and the processes of data encoding. *Quality and Quantity*, *53*(1), 221–246. https://doi.org/10.1007/s11135-018-0744-3

Uher, J. (2020). Measurement in metrology, psychology and social sciences: data generation traceability and numerical traceability as basic methodological principles applicable across sciences. *Quality and Quantity*, *54*(3), 975–1004. https://doi.org/10.1007/s11135-020-00970-2

Uher, J. (2021a). Psychology's Status as a Science: Peculiarities and Intrinsic Challenges. Moving Beyond its Current Deadlock Towards Conceptual Integration. *Integrative Psychological and Behavioral Science*, *55*(1), 212–224. https://doi.org/10.1007/s12124-020-09545-0

Uher, J. (2021b). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical and Philosophical Psychology*, *41*(1), 58–84. https://doi.org/10.1037/teo0000176

Uher, J. (2021c). Quantitative psychology under scrutiny: Measurement requires not result-dependent but traceable data generation. *Personality and Individual Differences*, *170*(110205). https://doi.org/10.1016/j.paid.2020.110205

Uher, J. (2022a). Functions of units, scales and quantitative data: Fundamental differences in numerical traceability between sciences. *Quality and Quantity*, *56*(4), 2519–2548. https://doi.org/10.1007/s11135-021-01215-6

Uher, J. (2022b). Rating scales institutionalise a network of logical errors and conceptual problems in research practices: A rigorous analysis showing ways to tackle psychology's crises. *Frontiers in Psychology*, *13*, 1009893. https://doi.org/10.3389/fpsyg.2022.1009893

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass*, *17*(5), e12740. https://doi.org/10.1111/spc3.12740

Uher, J., & Visalberghi, E. (2016). Observations versus assessments of personality: A five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments. *Journal of Research in Personality*, *61*, 61–79. https://doi.org/10.1016/j.jrp.2016.02.003

Uher, J., Werner, C. S., & Gosselt, K. (2013). From observations of individual behaviour to social representations of personality: Developmental pathways, attribution biases, and limitations of questionnaire methods. *Journal of Research in Personality*, *47*(5), 647–667. https://doi.org/10.1016/j.jrp.2013.03.006

Uijtdehaage, S., & O'Neal, C. (2015). A curious case of the phantom professor: Mindless teaching evaluations by medical students. *Medical Education*, *49*(9), 928–932. https://doi.org/10.1111/medu.12647

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, *2017*(5). https://doi.org/10.7717/peerj.3299

Uttl, B., & Violo, V. C. (2021). Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students? *ScienceOpen Research*. https://doi.org/10.14293/s2199-1006.1.sor.2021.0001.v1

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

van Eerde, W., & Azar, S. (2020). Too Late? What Do You Mean? Cultural Norms

Regarding Lateness for Meetings and Appointments. *Cross-Cultural Research*, *54*(2–3), 111–129. https://doi.org/10.1177/1069397119866132

VIM. (2004). International vocabulary of basic and general terms in metrology (VIM). In *International Organization* (Vol. 2004). International Organization for Standardization.

Vinet, L., & Zhedanov, A. (2011). A "missing" family of classical orthogonal polynomials. *Journal of Physics A: Mathematical and Theoretical*, *44*(8), 120–123. https://doi.org/10.1088/1751-8113/44/8/085201

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education*, *23*(2), 191–212. https://doi.org/10.1080/0260293980230207

Wambsganss, T., Winkler, R., Schmid, P., & Söllner, M. (2020). Unleashing the Potential of Conversational Agents for Course Evaluations: Empirical Insights from a Comparison with Web Surveys. *Twenty-Eighth European Conference on Information Systems (ECIS2020)*, *May*, 1–18. https://aisel.aisnet.org/ecis2020_rp/50

Wambsganss, T., Winkler, R., Söllner, M., & Leimeister, J. M. (2020). A conversational agent to improve response quality in course evaluations. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3334480.3382805

Weng, C., Weng, A., & Tsai, K. (2014). Online teaching evaluation for higher quality education: Strategies to increase university students' participation. *Turkish Online Journal of Educational Technology*, *13*(4), 105–114.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The Role of Gender in Scholarly Authorship. *PLoS ONE*, *8*(7), e66212. https://doi.org/10.1371/journal.pone.0066212

Westra, E. (2019). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, *196*(7), 2821–2846. https://doi.org/10.1007/s11229-017-1575-9

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response Biases. In *The ITC International Handbook of Testing and Assessment* (pp. 349–363). New York: Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0024

White, L. T., Valk, R., & Dialmy, A. (2011). What is the meaning of "on time"? the

sociocultural nature of punctuality. *Journal of Cross-Cultural Psychology*, *42*(3), 482–493. https://doi.org/10.1177/0022022110362746

Whitehead, A. N. (1929). *Process and Reality*. NY: Harper. https://doi.org/10.1038/126754a0

Wiggins, B. J. (2011). Confronting the Dilemma of Mixed Methods. *Journal of Theoretical and Philosophical Psychology*, *31*(1), 44–60. https://doi.org/10.1037/a0022612

Williams, W. M., & Ceci, S. J. (1997). "How'm I Doing?" Problems with Student Ratings of Instructors and Courses. *Change: The Magazine of Higher Learning*, *29*(5), 12–23. https://doi.org/10.1080/00091389709602331

Willig, C. (2008). *Introducing Qualitative Research in Psychology* (2nd ed.). Open University Press, UK.

Wong, B., & Chiu, Y. L. T. (2019). Let me entertain you: the ambivalent role of university lecturers as educators and performers. *Educational Review*, *71*(2), 218–233. https://doi.org/10.1080/00131911.2017.1363718

Wood, A. M., Brown, G. D. A., Maltby, J., & Watkinson, P. (2012). How Are Personality Judgments Made? A Cognitive Model of Reference Group Effects, Personality Scale Responses, and Behavioral Reactions. *Journal of Personality*, *80*(5), 1275–1311. https://doi.org/10.1111/j.1467-6494.2012.00763.x

Wu, A. H. (2017). Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum. In *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3051462

Wu, M. J., Zhao, K., & Fils-Aime, F. (2022). Response rates of online surveys in published research: A meta-analysis. *Computers in Human Behavior Reports*, *7*, 100206.

Youmans, R. J., & Jee, B. D. (2007). Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course. *Teaching of Psychology*, *34*(4), 245–247. https://doi.org/10.1080/00986280701700318

Zagaria, A., Ando', A., & Zennaro, A. (2020). Psychology: a Giant with Feet of Clay. *Integrative Psychological and Behavioral Science*, *54*(3), 521–562. https://doi.org/10.1007/s12124-020-09524-5

Zimbardo, P. G. (1969). *The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. Nebraska Symposium on Motivation* (W. J.

Arnold & D. Levine (Eds.); pp. 237–309). Lincoln: University of Nebraska Press.

Appendix A

Findings from Study 1

Table A1

The Overview of the Main themes into which I categorised Information Participants Considered when Rating the "Organisation", "Engagement", "Effective Teaching" and "Support with Assessment" Items.

	"Best" lecturers	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**
THEMES FOR Item 1: Their module was well-organised			**THEMES FOR Item 2: The lecturer made the subject interesting**			**THEMES FOR Item 3: The way their module was taught has helped me to learn.**			**THEMES FOR I4: I have received good support to manage my assessment workload.**		
Structure and organisation	**63.9% (1% negative)**	**67.9% (11.9% positive)**	Content of the lectures	**27.8%**	**25% (4.8% positive)**	Structure and organisation	**32%**	**34.5% (6% positive)**	Structure and organisation	**35.1% (1% negative)**	**21.4% (8.3% positive)**
Teaching skills/ methods	**35.1%**	**34.5 (1.2% positive)**	Teaching skills/ methods	**14.4%**	**25% (1.2% positive)**	Teaching skills/ methods	**14.4%**	**21.4% (4.8% positive)**	N/A	N/A	N/A

Interaction with students	**22.7%**	**20.2%**	Interaction with students	**7.2%**	**9.5%**	Interaction with students (“best” support only)	**8.2%**	**15.5%**	Interaction with students	**57.7% (5.2% negative)**	**61.9% (8.3% positive)**
Students’ experience	**11.3%**	**7.1%**	Students’ experience	**21.6%**	**10.7% (1.2% positive)**	Students’ experience	**24.7%**	**40.5% (7.1% positive)**	Students’ experience	**5.2%**	**9.5%**
Lecturer’s passion	**4.1%**	**Not mentioned**	Lecturer’s passion-enthusiasm	**13.4%**	**7.1%**	Lecturer’s passion	**4.1%**	**Not mentioned**	N/A	N/A	N/A
N/A	N/A	N/A	Engagement with the class	**42.3% (1% negative)**	**38.1% (1.2% positive)**	Engagement with class	**30.9%**	**7.1%**	N/A	N/A	N/A

Note. This table depicts percentages of how many participants mentioned content related to the particular themes and sub-themes. A participant’s comment could be categorised into several sub-themes. The percentages therefore do not always add to 100. The table portrays how participants interpreted the SET items and shows that a large field of meaning was constructed for the same items.

Table A2

The Overview of the Main themes into which I categorised Information Participants Considered when Rating the "Feedback", "Challenge" and "Overall satisfaction" items.

	"Best" lecturers	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**
THEMES FOR I5: Feedback has helped me develop and improve my performance.			**THEMES FOR I6: This module has challenged me to do my best work.**			**THEMES FOR I7: Overall, I was satisfied with the teaching on this module.**		
Quality of feedback	**29.9%**	**38.1%**	Structure and organisation	**25.8%**	**26.2% (7.1% positive)**	Structure and organisation	**17.5%**	**28.6% (3.6% positive)**
Problems with feedback or Useful-interesting-organised feedback	**10.3% (5.2% negative, 3.1% neutral)**	**16.7% (11.9% positive, 4.8% neutral)**	Teaching methods/skills	**13.4%**	**15.5% (2.4% positive)**	Teaching methods/skills	**41.2% (1% negative)**	**32.1% (2.4% positive)**
Interaction with students	**13.4%**	**7.1%**	The lecturer's approach towards students	**15.5% (1% negative)**	**20.2% (1.2% positive)**	The lecturer's approach towards students	**20.6%**	**16.7% (2.4% positive)**
Students' experience	**4.1%**	**9.5%**	Students' experience	**37.1% (2.1% negative)**	**20.2% (1.2% positive)**	Students' experience	**34% (1% negative)**	**21.4%**
Provision-amount of feedback	**12.4%**	**16.7%**	Level of challenge	**13.4%**	**17.9% (2.4% positive)**	Module quality/comparison	**14.4%**	**20.2% (1.2% positive)**

Helpful -unhelpful feedback	**23.7%**	**14.3%**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**

Table A3

The Overview of the Main Themes into which I Categorised Participants' General Interpretations of SET items, the "Organisation", "Engagement", "Effective Teaching" and "Support with Assessments"

	"Best" lecturers	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**
THEMES FOR Item 1: Their module was well-organised			**THEMES FOR Item 2: The lecturer made the subject interesting**			**THEMES FOR Item 3: The way their module was taught has helped me to learn.**			**THEMES FOR Item 4: I have received good support to manage my assessment workload.**		
Structure and organisation	**90.7%**	**91.7%**	Content of the lectures	**42.3%**	**38.1%**	Structure and organisation	**35.1%**	**45.2%**	Structure and organisation	**41.2%**	**46.4%**
Teaching methods/skills	**23.7%**	**36.9%**	Teaching methods/skills	**16.5%**	**20.2%**	Teaching methods/skills	**15.5%**	**23.8%**	Teaching methods/skills	**7.2%**	**9.5%**
Interaction with students	**12.4%**	**10.7%**	The lecturer's attitude towards students	**Not mentioned**	**3.6%**	Interaction with students	**24.7%**	**35.7%**	Interaction with students	**57.7%**	**71.4%**
Students' learning-understanding	—	**4.8%**	Students' experience	**23.7%**	**22.6%**	Students' experience	**30.9%**	**22.6%**	Students' experience	**3.1%**	**3.6%**
N/A	—	—	Lecturer's passion-enthusiasm	**11.3%**	**17.9%**	Lecturer's passion	—	**1.2%**	N/A	—	—
N/A	—	—	Engagement with the class	**57.7%**	**59.5%**	Engagement with class	**19.6%**	**19%**	N/A	—	—

Table A4

The Overview of the Main Themes into which I Categorised Participants' General Interpretations of SET items, the "Feedback", "Level of Challenge" and "Overall Satisfaction"

	"Best" lecturers	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**		**"Best" lecturers**	**"Worst" lecturers**
THEMES FOR Item 5: Feedback has helped me develop and improve my performance.			**THEMES FOR Item 6: This module has challenged me to do my best work.**			**THEMES FOR Item 7: Overall, I was satisfied with the teaching on this module.**		
Quality of feedback	**54.6%**	**54.8%**	Structure and organisation	**23.7%**	**33.3%**	Structure and organisation	**18.6%**	**17.9%**
Helpfulness of feedback	**49.5%**	**47.6%**	Teaching methods/skills	**15.5%**	**17.9%**	Teaching methods/skills	**48.5%**	**47.6%**
Interaction with students	**9.3%**	**14.3%**	The lecturer's approach towards students	**20.6%**	**25%**	Interaction with students	**29.9%**	**23.8%**
Students' experience	**2.1%**	**3.6%**	Students' experience	**29.9%**	**32.1%**	Students' experience	**43.3%**	**33.3%**
Provision-amount of feedback	**6.2%**	**7.1%**	Level of challenge	**24.7%**	**27.4%**	Previously mentioned/ overall	**9.3%**	**6%**

Table A5

Chi-square Analyses for Considered Main Themes for Male and Female Lecturers Across All Items During Ratings

	"BEST" LECTURERS				**"WORST" LECTURERS**			
	Male	**Female**			**Male**	**Female**		
THEMES FOR ALL ITEMS	*N*	*N*	$X^2(1)$	*p*	*N*	*N*	$X^2(1)$	*p*
Structure and organisation	49	29	0.018	.893	36	32	0.045	.832
Teaching skills/methods	43	25	0.054	.817	33	26	1.002	.317
Interaction with students	48	29	0.004	.987	32	28	0.076	.782
Students' experience	42	24	0.116	.733	29	30	0.828	.363
Engagement with the class	38	18	1.646	.200	19	20	0.392	.531
Quality of feedback	15	13	1.345	.246	11	20	**5.624**	**.018***
Total number of lecturers	60	36			44	40		

Note. One case was excluded from analysis because of an unspecified lecturer's gender.

Table A6

Example of a Coding Frame from Study 1

CODING FRAME FOR WORST LECTURERS Q4L: I have received good support to manage my assessment workload. What did you consider in the rating of this lecturer?

THEMES	SUB-THEMES	CODES - EXAMPLES	DESCRIPTION	STATEMENTS (EXAMPLES)
Interaction with students **52 (42 unique, 10 double)** **61.9%**	Support from the lecturer (or lack of) (33) 4 positive 39.3%	Lack of availability Lack of help outside lectures Lack of support Lack of support in class	Refers to support from the lecturer in terms of general support, general availability, response to students when they require help. Includes positive comments.	"The professor was rarely available…" (W2) "… didn't seem open to meeting outside lectures to discuss concerns. Also didn't seem to appreciate that our results were important to us and took a "good enough" approach" (W7) "Lecturer failed to notice that I was struggling with heavy workload" (W1) "I was ignored when asked for help with assessment in class but attended every other student. I was only attended to last even thought I was sitting closer to her desk and she was going row by row." (W21)
7 positive comments (6 unique, 1 double) **8.3% positive**	Question-related comments-general communication (12) 2 positive 14.3%	Could not answer questions well Hard communication Opportunity for questions Unwilling to answer the question	Refers to the answers to students' questions or opportunity for questions. Also refers to comments about general communication with students (other than emails). Includes positive comments.	"The lecturer couldn't provide clear answers to questions…" (W7) "…was slightly had to communicate with them so I would have to ask another lecturer via email or drop-in sessions." (W32) "Whether they offered a time to ask questions about the assignment" (W14) "She didn't answer my question in detail though she knew the answer." (W35)
	Email communication (11) 2 positive 13.1%	Email communication was bad Emails ignored Email response (not replying on time)	Comments about email responses, ignoring emails, and email contact in general. Includes positive comments.	"I have many times contacted my tutor and have gotten zero response" (W33) "When I would have doubts she would ignore emails" (W28) "I found that my lecturer did not respond to emails on assignments in a timely manner" (W72)

	Hostile behaviour from the lecturer (6) 7.1%	Hostile behaviour from the lecturer Lecturer was hostile Lecturer was rude	Describing lecturer's behaviour as rude, aggressive or defensive.	"… then during the help sessions, this lecturer was hostile and defensive about any questions asked…" (W5) "When asked for support he became aggressive" (W55) "The lecturer was quite rude in response to questions and asked us to refer to what was on the course website." (W34)
Structure and organisation of the module 18 (15 unique, 3 double) **21.4%** **7 positive (8.3%)**	Assessment-related comments (14) 4 positive 16.7%	Assignment requirements and guidelines unclear Assignment requirements unclear Assignment specifications confusing	Comments referring specifically to assessments, clear structure or explanation of the assessment, assignment specifications and guidelines. Includes positive comments.	"Didn't provide clear advice about the assessment, how it would be marked etc" (W68) "Assessments were unclear about what was required from students, the lecturer was not particularly good at explaining what was needed to be successful in the assessment." (W56) "There was a major confusion about the assessment guidelines and the coursework itself was vague and it seemed like even the person who designed it did not understand it" (W36)
	Time management (7) 3 positive 8.3%	Timeline unclear Time to complete assignments Timing insufficient	Comments referring to general time management, clear deadlines or timelines for completing the assignment. Includes positive comments.	"He wasn't sure when the assessment would be released." (W68) "The time that given to assignments…" [sic] (W70) "The time windows for the completion of the assigments were too short." [sic] (W62)
Students' experience (8 unique) **9.5%**	Students did not need support (6) 7.1%	Student did not need help Student did not seek support from the lecturer The lecturer had no relevance to workload	Students did not need help or support, preferred to manage assessment themselves or felt that assessment support has no relevance to the lecturer.	"There was no need for this" (W66) "I am always supported by myself or classmates" (W4) "Nothing to do with the lecturer" (W54)
	Students felt stressed-unprepared (2) 2.4%	Student felt unprepared Student was stressed out	Students lacked confidence, felt unprepared or stressed out.	"I feel we are extremely unprepared." (W31) "…It caused me a major distress as it undermined my belief in my academic abilities, which affected my further assignments." (W36)

Further examples of findings from qualitative analyses (figures, coding frames, tables) are available upon request.

Appendix B

Findings from Study 2

Table B1

Means and Standard Deviations for Different Directions of Scale

	Scale direction SA (N=168)	**Scale direction SD N= (165)**
ITEM	**M (SD)**	**M (SD)**
Q1: The lecturer makes the subject interesting.	3.16 (1.03)	3.02 (1.06)
Q2: The way their module is taught helps me to learn.	3.02 (1.04)	2.98 (1.09)
Q3: This lecturer provides good support to manage my assessment workload.	2.73 (1.07)	2.78 (1.12)
Q4: Their module is well-organised.	3.14 (1.11)	3.07 (1.14)
Q5: Their feedback helps me develop and improve my performance.	3.59 (1.13)	3.84 (1.04)

Table B2

Means and Standard Deviations for Male and Female Lecturers

	Female lecturers (N=168)	**Male lecturers N= (165)**
ITEM	**M (SD)**	**M (SD)**
Q1: The lecturer makes the subject interesting.	3.04 (1.06)	3.13 (1.03)
Q2: The way their module is taught helps me to learn.	2.98 (1.08)	3.02 (1.06)
Q3: This lecturer provides good support to manage my assessment workload.	2.70 (1.05)	2.81 (1.14)
Q4: Their module is well-organised.	3.05 (1.09)	3.16 (1.17)
Q5: Their feedback helps me develop and improve my performance.	3.70 (1.08)	3.73 (1.11)