Selection Effects in Students' Evaluation of Teaching

Methodological Pitfalls for the Measurement of Teaching Quality

Edgar J. Treischl

May 19, 2019

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Graduate School of Economic and Social Sciences at the University of Mannheim

Dean of the School of Social Sciences:	Prof. Dr. Michael Diehl
Supervisor:	Prof. Dr. Tobias Wolbring
First reviewer:	Prof. Dr. Thomas Gautschi
Second reviewer:	Prof. Dr. Oliver Dickhäuser
Date of defence:	October 9, 2019

You never walk alone ...

To all my friends I am not sure whether I would have finished the dissertation without your support. I have many people in mind, but I am thinking especially of *Marta Alvarez-Alonso, Bert Brandt, Anna-Lena Hönig, and Laila Schmitt*. Thanks for your support, your advice, and your patience.

To the team at the University of Munich I studied in Munich, I worked for two years at the University of Munich, and for a very long time Munich was my safe haven. I am very grateful for that time. Thank you *Katrin Auspurg, Felix Bader, Johannes Bauer, Christiane Bozoyan, Josef Brüderl, Werner Fröhlich, Christian Ganser, Marc Keuschnigg, Volker Ludwig, Sonja Pointner, Patrick Riordan, and Fabian Thiel.*.

To the Graduate School of Economic and Social Sciences at the University of Mannheim Time had come to change. I left Munich and headed to Mannheim in 2015. Being part of the graduate school was a stressful period in the beginning, but most of it was a game changer. I want to thank all people who made that possible: *Thomas Bräuninger, Henning Hillmann, and Florian Keusch* as instructors of the doctoral colloquium, as well as all other members of the CDSS family, especially *Milanka Stojkovic* for her support from the very start. Looking back at my first year at the graduate school keeps me thinking of *Thomas Gschwend* and his advice for all (first year) PhD candidates: *Don't get it right, get it written.* In the meantime I caught myself giving my students the same advice. *I hope you take this as a compliment for both courses in the first year.*

To all team members at the University of Mannheim Special attention deserve *Florian Keusch* and *Frauke Kreuter* for giving me the chance to work at their chair at the University of Mannheim. Of course, I also want to thank my former colleagues *Vlad Achimescu, Ursula Eckle, Christoph Kern, Alex Scherf, Malte Schierholz,* and the whole chair of statistics and methodology for being a part of the journey.

To the team at the University Erlangen-Nuremberg All beginnings are difficult, but this does certainly not apply to my personal start at the University of Erlangen-Nuremberg. Thus, I'd like to thank many people who welcomed me and made that smooth transition possible. Foremost, I like to thank *Tobias Wolbring and everybody at the chair of empirical economic sociology* as well as *Martin Abraham and his research team.*

To those who made the dissertation possible Certainly, I could not write a dissertation about students' evaluation of teaching without data. Therefore, I want to thank all students and all instructors who participate in each study. Relying on data from three different institutions, my special thanks goes to all persons who are (were) involved at each institution. Especially *Josef Brüderl, Maria Mühlbauer, Carolin Simon, Jonas Weigert, and Karl Wilbers* for their *evaluation manpower.*

To Tobias Wolbring When I first met Tobias, I could hardly imagine where my journey would go to. At that time Tobias was a PhD student and I was one of his students at the University of Munich learning SPSS. I am very grateful that Tobias offered me the opportunity to work for Norman Braun as a student assistant. Since that days, Tobias had different roles in my life. He was my supervisor at the Chair in Munich, he is a co-author who offered me plenty of opportunities – and struggles – to develop academically, and he supervised this PhD thesis. *I hope you know I couldn't walk my way without your guidance.*

To my reviewers Ultimately, I want to thank *Oliver Dickhäuser* and *Thomas Gautschi* for their guidance and support which helped me finalize this dissertation. I have never taken it for granted that you offered to review this dissertation without any hesitation.

Contents

List of Tables

List of Figures

“Selection bias as a fundamental social process [is] worthy of study in its own right rather than a statistical nuisance [...]”

Robert J. Sampson (2008: 189) [Note: E.T.]

Part I.

Dissertation Overview

1. Executive Summary

Student evaluation of teaching (SET) is an empirically well-established instrument for quality assurance in higher education. Most research shows that SET is a reliable and valid instrument to measure teaching quality from students' perception and in accordance with Seldin's statement one may come to the conclusion, that: *"the opinions of those who eat the dinner should be considered if we want to know how it tastes"* (Seldin 1993: 40). However, in many instances students can select courses they want to attend before the course starts, and even in mandatory courses students can often decide how regularly they attend. Fairness concern remain if those *who eat the dinner can decide which cuisine they prefer* before the dinner starts. And in terms of attendance, one may argue that SET results will be distorted if those *who did not like the dinner decided to go home* before they were asked about their opinion. Against this background this dissertation asks about the consequences of students' course selection and attendance for SET and provides four research papers about students' selection effects.

First, from a survey-methodological perspective students in class may deviate in central covariates from the target population due to a systematic sorting-process which may lead to a selective sample of students participating in the SET. For this reason, Treischl (2019a) asks from a theoretical point of view how students come to a decision and what explains students' course selection. Moreover, I stress the risk that SET are distorted if known bias variables from SET have a causal impact on SET and students' course selection. I conducted a factorial survey experiment which contained three potential SET bias variables – course interest, difficulty, and workload – to examine the impact on students' course selection. Main results in terms of selection effects indicate that course interest has a very strong causal impact on students' decision, while workload and difficulty have a lower but significant effect on students' course selection. These findings make clear that students' course selection may have a substantial effect on SET in case the sorting process is related to the measurement of teaching quality. SET will be distorted if a course attracts students with a higher interest and the same applies if course difficulty or workload dis- or encourage students to attend a course.

Second, Treischl (2019b) asks about the causal impact of the three potential bias variables on SET. Previous research investigates the impact of many bias variables, but for most bias variables mixed results can be found in the literature. Most research about bias variables relies on observational SET data which is not well suited to identify a causal effect and previous research does not account for student self-selection. Therefore, Treischl (2019b) introduces a factorial survey experiment in SET and estimates the causal impact of course interest, course difficulty and workload on student's assessment of teaching quality. Main findings indicate that course interest and workload have a significant, but moderate impact on students' perception of teaching quality, while difficulty has a significant but low effect. Furthermore, the factorial survey experiment contains the perceived learning outcome, which gives me the opportunity to examine the mechanism behind the considered bias variables and to estimate the effect independently of the perceived learning outcome and other dimensions of teaching quality. Overall, main findings indicate that course interest and workload are bias variables that distort SET results and that an excellent instructor may reduce or induce the effect.

Third, students can also decide how often or how regularly they attend a course. From a missing values perspective students attendance decisions (and course cancellation) lead to a loss of data, but may equally induce a non-response bias in case the loss is selective. To this end Wolbring & Treischl (2016) ask whether *one-shot* paper-and-pencil SET are affected by students' absenteeism from class. Results based on a research design with two measurement points in time indicate that approximately one third of the students do not attend class at the second time of measurement, by which time the regular paper-and-pencil SET would take place. We test several hypotheses regarding absenteeism and show that course quality significantly and strongly increases the chance that students will participate in the regular SET, which underlines that the loss of data is selective and not missing at random. Subsequently we examine the consequences of the data loss regarding course ratings and rankings.

Forth, Treischl & Wolbring (2017) highlight methodological pitfalls of online and paper-and-pencil surveys in SET and demonstrate that methodological research about SET needs to account for students' self-selection in order to assess the causal effect of a survey mode in SET. We examine the causal effect of a change in survey mode on SET based on a series of three consecutive field experiments: A split-half design, a twin course design, and pre-post design. In conclusion, we find systematic, but small differences in the overall course ratings. On average, online SET reveal a slightly less optimistic picture of teaching quality in students' perception. Albeit these findings, email-based methods account for selection effects and attendance decisions. As a consequence, students can provide feedback in online SET even if they regularly attend the course, but are not present in class at the time of the SET. Based on the experiments we show that it is crucial to include non-attending students when evaluating teaching: their course assessments clearly deviates in many cases from the overall rating.

1.1. Introduction

Student evaluation of teaching (SET) is the standard tool to measure teaching quality from students' perception (see Spooren et al. 2013). From an empirical point of view SET is a well-established instrument for quality assurance in higher education, with at least two different aims (see Marsh 2007 for an overview). First, SET is used as a *formative assessment instrument* which gives teachers a systematic feedback in order to improve their own teaching skills. Second, SET is also considered as a *summative assessment instrument*. SET results are increasingly used for decisions about allocation (hiring, remuneration, etc.) of performance-related resources (Avery et al. 2006; Dommeyer et al. 2004; Kherfi 2011), or for the assortment of teaching prize nominees (see Gravestock & Gregor-Greenleaf 2008: 16). SET results also play an important part in accreditation procedures, are accounted in higher education rankings, and are taken into account in the evaluation of career trajectories (e.g. tenure track). Using SET results in such instances stress the assumption that course ratings and rankings are an appropriate way to measure and to compare course quality. To this end SET must reliably and validly identify *exemplary good or bad* teachers or courses. The usage of SET as a summative assessment instrument is unacceptable and unfair if concerns about the reliability and validity of SET exist.

There is an extensive amount of research focussing on reliability and validity concerns of SET; many research contributions ask especially about the discriminant validity of SET and call into question whether some *bias variables* – e.g. prior course interest, sympathy of the instructor, or workload – distort SET (for an overview see Feldman 2007; Marsh 2007; Spooren et al. 2013). Moreover, previous research examines from a substantial point of view, which and how many dimensions need to be considered in a SET instrument (Burdsal & Harrison 2008, Mortelmans & Spooren 2009), how these dimensions can be measured (see Marsh 1982; Marsh et al. 2009), or which methodological pitfalls come along with the choice of a survey instrument (see Adams & Umbach 2012; Carini et al. 2003). Regarding the amount of research it is surprising that students' self-selection is almost no concern in higher education and most research does not address the possibility and the consequences of a self-selection bias due to students' decision (except for some early contributions, see Esser 1997; Romer 1993).

Self-selection may work through two selection mechanisms. First, in many instances students decide which course they want to attend, especially in elective- and multi-section courses it is up to students to decide. Students' course selection points to the question whether a (non-) selective sample of students will select, and ultimately, assesses the course. One may expect that SET will be distorted if a selective sample of students assess the course. Second, students also often decide *whether and/or how regularly* they attend a course. Selection in form of *missing class, absenteeism and course attendance* is an acknowledged research topic in economics and higher education. Most research deal with the consequences of absenteeism for students and universities (see Bahr 2009 for an overview). Previous research even identified attendance as an important bias variable for SET (see Beran & Violato 2005, Wolbring 2012), but the methodological pitfalls and possible consequences of a self-selection bias are rarely considered, and this lack of knowledge represents a research gap with implications for higher education and research about SET. In consequence, both selection mechanisms point to an incomplete and probably selective coverage of course participants during the survey which challenges the reliability and validity of SET. One may fear that in some instance only those who are interest in a course will *show up* in the first place. The other way around, student may vote with their feet and do *not regularly show up* in case of a mandatory course.

This dissertation focusses therefore exclusively on students' self-selection effects and examines possible consequences for SET and research about SET[1]. First, I briefly describe the psychometric properties that SET needs to fulfil in order to measure teaching quality. This point underlines that selection effects may have major consequences for the measurement of teaching quality and research about SET. I subsequently deduce four research questions as main contribution of the dissertation. These research questions are fully discussed in the four papers of the dissertation, but main findings are highlighted in this overview. Finally, I un-

[1] This dissertation examines only the effects of students' self-selection, even though active forms of selection may affect SET as well. For example, in terms of restriction to study programmes or specific requirements for course admission.

derline implications for higher education and emphasize the research contribution in broader terms.

1.2. Psychometric Properties of SET

Research about SET often stresses psychometric properties of evaluation instruments to measure teaching quality, but the possible consequences of a selection bias are far less considered when it comes to the question whether SET is a reliable and valid instrument to measure teaching quality. This subsection gives an overview about the psychometric properties for SET based on Spooren et al. (2013) and adds concerns regarding a selection bias in SET[2].

1.2.1. Reliability of SET

Reliability refers to the consistency of a measurement (see King et al. 1994). A reliable instrument comes to the same result if the measurement is repeated under the same conditions. For instance, will students come to the same result if they rate the course twice (intrarater reliability). A vast majority of research shows that SET is a reliable instrument, at least if the assessment is based on several participants and average scores are reported. Marsh (2007) indicates, that SET results strongly overlap in terms of variance when students evaluate the same course twice. The same applies to interrater reliability: SET results are still quite stable if different course participants rate the course over time (see Marsh & Hocevar 1991, Rindermann & Schofield 2001) or if reliability measures rely on observations from actual students compared to former students in retrospective (Feldman 1989).

A strong overlap in terms of variance speaks for the reliability of SET, but this does certainly not imply that a SET instrument has a precision of a *calibrated personal scale*, especially if the measurement relies on a volatile course audience over time. As Romer (1993) noticed: “In short, on a typical day at a typical elite American university, roughly one-third of the students in economics courses are not attending class” (Romer 1993: 168). Conducting a SET at such a typical day implies that a substantial part of the audience is excluded from the survey.

[2] Spooren et al. (2013) provide an overview article to summarize the recent literature about SET. Overall, they found 542 peer-reviewed articles and selected studies with a focus on reliability and validity concerns.

Certainly, we could assume that this non-response is random, but even in this case the loss of data reduces the reliability of a SET instrument by increasing the variance of the measurement. However, these *no-shows* may differ in terms of perceived teaching quality which implies that loosing information from these students may have severe consequences in terms of validity.

1.2.2. Validity of SET

Concerns regarding the *validity* of the instrument refer to the question whether SET measures what it is supposed to measure: teaching quality. Previous research examines several types of validity[3]. Most researchers in the SET field acknowledge that teaching quality is a multidimensional construct, which is why SET need to cover different dimensions of teaching quality (see Rindermann & Schofield 2001). Research in higher education has developed a variety of SET instruments, which vary with respect to the amount and the considered dimensions of teaching quality (see Spooren et al. 2013: 5). Some researchers also criticize the lack of a precise *theory of teaching effectiveness* (see Ory & Ryan 2001, Penny 2003), even though some consensus exist (Rindermann & Schofield 2001). The lack of a precise theory of teaching effectiveness clearly underlines that an instrument may not capture teaching quality if the questionnaire omits important dimensions. Therefore Spooren et al. (2013) demands that all involved stakeholder (institutions, teacher, and students) should be included when it comes to the development of SET instruments.

However, including all important dimensions in a SET instrument does not guarantee that SET is not distorted by a selective sample of course participants; and students may assess an excellent or bad teacher based on criteria which does not reflect teaching quality. The latter concern stresses the construct validity (convergent- and discriminant) of SET (see Spooren et al. 2013 for more details and additional criteria of validity).

[3] For example, from a *content-related perspective* a valid SET instrument points to the questionnaire as a whole (sampling validity) and to single indicators (item validity) that represent the content (see Spooren et al. 2013).

Construct Validity of SET: Convergent Validity

From a convergent validity perspective, a valid SET should substantially be associated with other measurements of teaching quality; and *similar constructs* like teaching quality. Therefore, one may expect that SET are correlated with learning outcomes like students' achievements ("objective measure" like grades) or students' perceptions of learning (subjective measure).

Most previous research reports that SET are positively and moderately correlated to *students grades* (Cohen 1981; Feldman 2007), even though contradictory findings can be found in the literature (e.g. Clayson 2009). Nevertheless, the association seems even stronger with respect to subjective measurements and single dimensions of teaching quality compared with an overall teaching indicator. Based on a meta analysis Feldman (2007) reports that, for example, instructors' preparation explains 33% variance of students' grades and various studies highlight that SET results are significantly correlated with external ratings (expert ratings, alumni ratings, or self-assessment of instructors). Overall, these findings illustrate why most authors in higher education perceive SET as a sufficiently valid instrument, at least in terms of convergent validity (see Marsh & Roche 1997: 1190, Spooren et al. 2013).

Construct Validity of SET: Discriminant Validity

On the contrary, a second branch of research is concerned about the discriminant validity of SET and refers to potential *bias variables*, which by definition should not be associated with SET results (Aleamoni 1999). Students' pre-course motivation (see Griffin 2004, Spooren 2010), instructors' grading behaviour (e.g. Centra 2003, Olivares 2001), class size (Bedard & Kuhn 2008, Ting 2000), and many more variables on a student-, instructor-, and course level are significantly correlated to SET and may substantially distort SET results.

The current state of research is inconclusive for many bias variables and in many instances, it is unclear whether a bias variable distorts SET results, causally influences teaching quality, or is caused by methodological pitfalls of previous research[4]. For example, the association

[4] As Marsh (1987) notes, the quest for bias variables is itself biased for several methodological reasons and can be perceived as a *witch hunt* (Marsh 1987: 309).

between *prior course interest* and SET might be an artefact based on a spurious correlation (e.g. students motivation), the class size effect might reflect unobserved heterogeneity (e.g. instructor's ability to teach larger classes), and estimating a grading leniency effect with cross sectional data is a challenge due to concerns about reverse causality[5].

In summary, this short discussion about the validity of SET shows that both research areas have strong arguments for or against the validity of SET. Previous research also points to moderate or small impact of most student's bias variables (except for course interest), which is why a rather broad consensus exists that many bias variables can be neglected (see Spooren et al. 2013)

1.3. Selection Effects in SET

SET results are mostly affected by dimensions of effective teaching (Barth 2008, Greimel-Fuhrmann & Geyer 2003) which is why many researchers perceive SET as a valid instrument to measure teaching quality. This judgement does not take the consequences of student self-selection into account. In case of a self-selection bias one may assume that prior course interest has an effect on SET, because in many instances student in class have selected the class based on their personal *taste*, which points to the discriminant validity of the instrument. Or in terms of attendance one may hypothesize that students in class attend on a regular base because they have learned a great deal. In consequence SET may have a lower convergent validity than previous research reports.

A causal inference paradigm highlights why student self-selection are of major importance and may distort SET results: Under *ideal methodological conditions* SET would rely on a random sample and randomly assign students to courses and control groups. In such an experimental setup it is possible to estimate an *average teaching effect* and we could observe the causal impact of a teacher, for example in terms of learning or students perceived teaching quality. However, in many instances students can select which course – or in terms of

[5] Cross sectional data is not well equipped to assess the association between students' grades and SET results since both directions of causality are plausible. Students can consider expected grades in the assessment process, but an effective instructor can increase student learning and grades.

causal inference which *treatment* – they want to receive in the first place. Under *ideal* survey-methodological conditions all enrolled students would have the same chance to attend a course and participate in SET; but students may omit early or late courses due to motivational aspects (Brown & Kosovich 2015; Devadoss & Foltz 1996), which illustrates that students' course selection has an influence on the question who will assess the course. In terms of causal inference students can also decide how often or how regular they *receive the treatment*, foremost in classes without compulsory attendance. It is also not unlikely to assume that these *no-shows* are on average less satisfied with teaching quality and give worse ratings in SET.

Against this background this dissertation focusses on four research questions in terms of selection effects. The first two papers focus on students' course selection and Treischl (2019a, unpublished manuscript) asks first whether students systematically sort into classes and what does students' course selection explain? Subsequently, Treischl (2019b, unpublished manuscript) examines the consequences of students' course selection for SET in detail by asking about the causal impact of three bias variables – course interest, difficulty, and workload – which seem have an impact on students' course selection and to be related to SET.

The second part of this dissertation focusses on students' attendance decisions and the consequences for SET based on two papers. Wolbring & Treischl (2016, published in Research in Higher Education, 57(1), 51–71) ask whether students' attendance decisions lead to a missing data problem or a non-response error. Especially paper-and-pencil SET at the end of the term may exclude irregularly attending students by design. Ultimately, Treischl & Wolbring (2017, published in Research in Higher Education, 58 (8), 904–921) ask about the causal effect of the survey mode in SET and examine methodological advantages and pitfalls of paper-and-pencil and online SET. Moreover, we highlight the opportunity of email-based survey mode to adjust for the consequences of students' attendance decisions, even though online SET come along with additional drawbacks (e.g. low response-rates).

The following sections provide an overview of the dissertation and summarize the main *research question*, *contribution* and *results* of each paper, starting with students course selection (Treischl 2019a).

1.3.1. Course Selection and Selective Courses? Students' Course Selection and its Consequences for SET

Students' course selection and cancellation is an important topic with little attention in higher education (see Babad & Tayeb 2003, Wilhelm 2004), even though research can provide key insights for all involved stakeholders: Course selection (CS) is obviously important for students' academic integration and success (see Tinto 1975, 1993); and it may even operate as a crucial signal (e.g. acquired skills) for the transition into the labour market (see Spence 1973). Students' CS also points to class composition effects and an instructor is faced with additional challenges in case students' CS leads to a heterogeneous background of students in class (e.g. with respect to prior knowledge). Understanding students' considerations that lie beneath the decision-making process would also be extremely helpful to improve study conditions and to identify students with a high stake to drop-out of courses and universities. From this perspective more systematic research is clearly desirable.

Students' CS points to the question whether a (non-) selective sample of students has access to participate in SET. For example, previous research underlines that *course interest* explains students' CS (see Babad 2001, Babad & Tayeb 2003) and many researcher report that prior course interest is a bias variable for SET (see Griffin 2004, Olivares 2001). Course selection implies that some students have a lower (higher) probability to be observed in SET, which can either directly be related to (prior beliefs about) teaching quality; or indirectly via course interest and further bias variables that affect teaching quality. In both cases, a selective sub-sample may (not) be observed in SET and this loss of data is probably not *neglectable*[6]. Accordingly one may argue that SET will probably overestimate (underestimate) teaching quality if a course sample consists of (un-) interested students who will subsequently give a better (worse) course assessment. And the same applies to further bias variables (e.g. workload, difficulty,

[6] In terms of measurement, the consequences of students' CS can be perceived from a nonresponse and missing values perspective (for more information about the problem of missing data, see Little & Rubin 2002). Selective samples and missing data are not problematic for SET if students' CS and in consequence the nonresponse of students is unrelated to students' actual, but unobserved course assessment (completely at random); and as long as the missing data is unrelated to any other observed or unobserved variable, which has an impact on students' course assessment. For example, students may sort into a class with a convenient time slot, but SET results are not distorted as long as the course timing is unrelated to students' course assessment.

etc.) which may affect students' CS (Babad & Tayeb 2003, Wilhelm 2004) and therefore SET (Centra 2003, Remedios & Lieberman 2008). For these reasons Treischl (2019a) focuses on students' course decision to understand the selection process in greater detail and possible consequences for SET.

Research Contribution

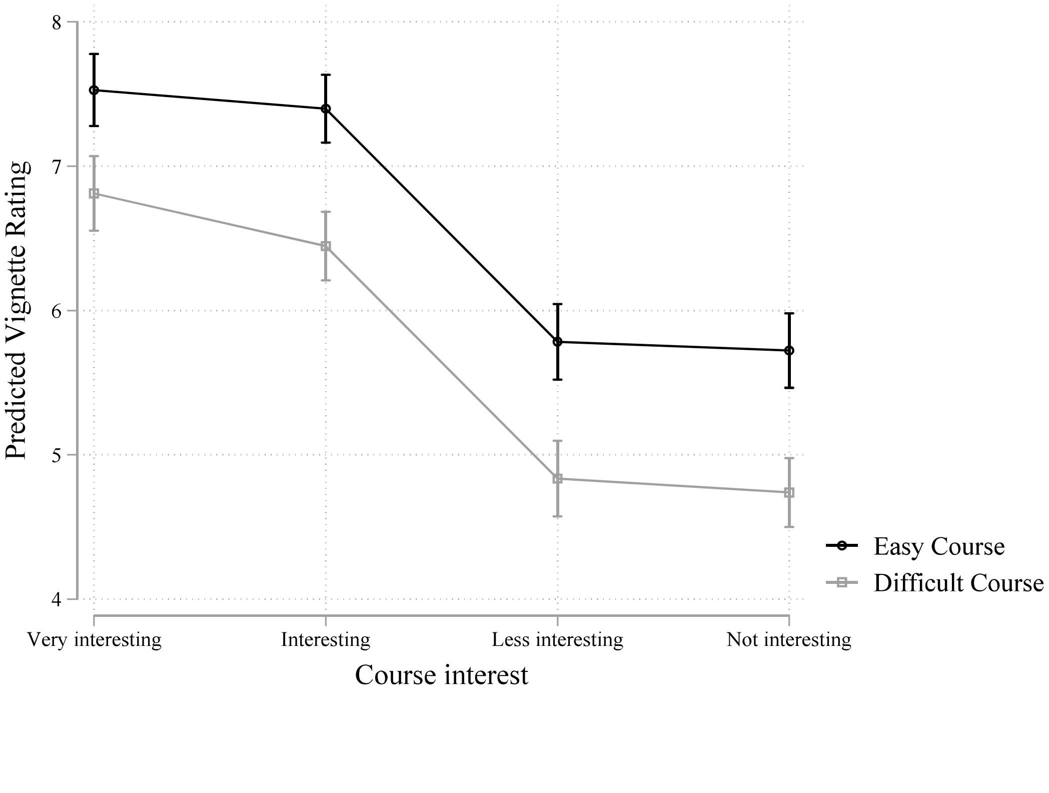
This paper highlights that a factorial survey experiment (FSE) or vignette study offers the opportunity to observe and dissect students' decision-making process. So far most previous research perceives students course selection as a rational choice but does not observe the simultaneous consideration of anticipated benefits and costs. Moreover, I highlight that students are often faced with a trade-off situation between different diverging course and teaching dimensions. For instance, students may maximize the benefit by selecting a course according to their interests, but how do they solve a trade-off situation between different diverging dimensions? Do students select the most interesting course if the latter comes along with a significant higher workload? Ultimately, I examine the consequences of students CS for the measurement of teaching quality by estimating the causal impact of three bias variables on students' CS. In order to assess whether students' CS has consequences for SET, knowledge about the causal effect of course interest and further bias variables is needed. However, previous research about students' CS relies almost exclusively on observational data which is not well suited to identify a causal effect (except for Babad & Tayeb 2003, Wilhelm 2004). To this end I included course interest, workload, and difficulty in the factorial survey experiment and I estimate the causal impact on students' CS.

Main Findings

Overall, two course benefits have a very strong impact on students' CS. Teaching quality and course interest significantly increase students' willingness to select a course. Workload and course difficulty – on the other side – significantly decrease students' willingness to select a course as well, but especially workload has a lower impact on students CS, while the effect size of course difficulty lies in between.

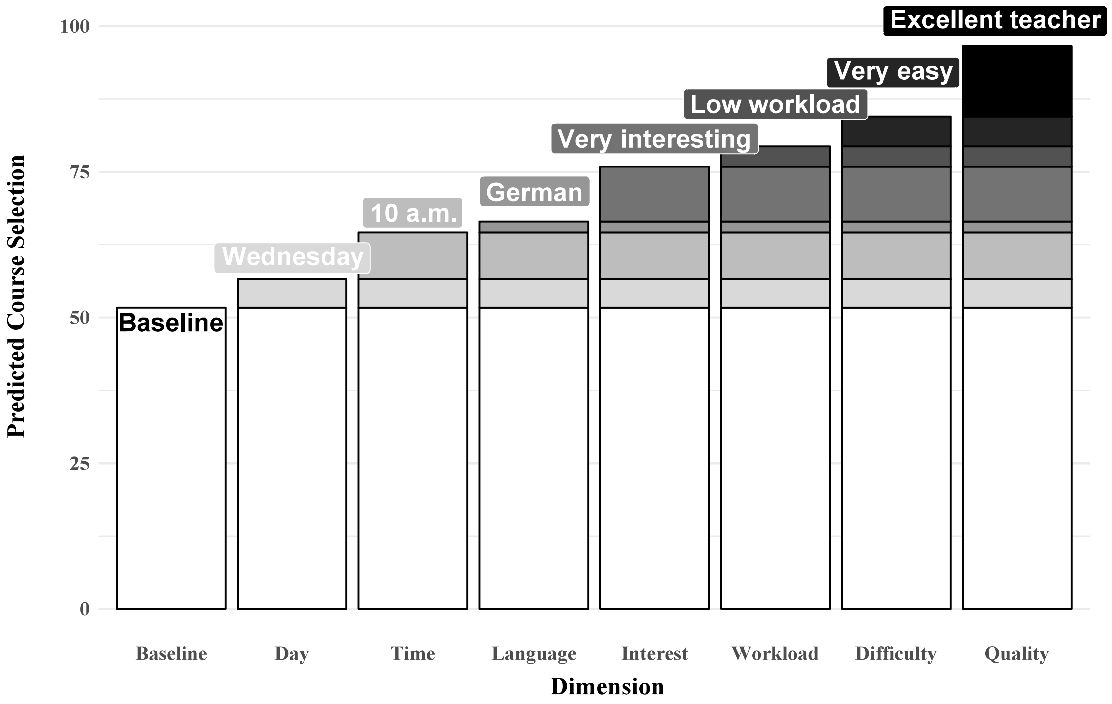
These findings do not imply that students solely select courses based on anticipated benefits. Considering students' CS in trade-off situation emphasizes that students weigh several costs and benefits simultaneously. I demonstrate this point by hypothesizing that high course benefits may compensate costs. I expected that the impact of workload and further cost aspects decrease with a higher course interest. Main results indicate that higher workload reduces students' willingness to select a course regardless of high or low course interest. The same applies for further cost and benefit aspects. For example, Figure 1.1 shows the interaction between course interest and course difficulty. An easy course increases students' willingness to attend a course, irrespective whether the course sounds interesting or not. All main effects have approximately the same impact on students' CS, regardless of the expected benefit. These findings are in clear contrast to expectation regarding the current state of research and the proposed hypothesis. In most instances the expected higher benefit does not compensate for the higher perceived cost which is in line with rational choice assumptions.

Figure 1.1.: Interaction between course interest and difficulty



Thus, findings from the FSE underline that students consider and weigh several course and teaching attributes, which makes clear why in some instances only a sub-sample of students may select the course. This point can be illustrated based on the main results. Figure 1.2 uses the result of the multi-level regression analysis and depicts predicted values for the most favourable level of each dimension, from students' point of view. The bar plot in Figure 1.2 depicts the average willingness to select a course and shows how each dimension (and corresponding levels) increase students' willingness to select a course. The bar on the left side contains the baseline (51.7%) only, which means the predicted value for students' willingness to select a course while keeping all FSE dimensions at the mean. To see how each dimension increases the baseline, I calculated the predicted values by adding each dimension stepwise. Each bar adds another dimension to illustrate how the willingness to select the course increases (or decrease in case of the most unfavourable scenario, see Treischl 2019a).

Figure 1.2.: Predicted willingness for the most favourable scenario



For example, in the first two steps I calculate the predicted values for course days and time. Courses on Wednesday at 12 a.m. increases the baseline by 12 %, while a late course on Friday decreases the baseline or the willingness to select a course. Courses come with several anticipated cost and benefits and students' willingness to select a course is substantially increased (decreased) in case a course has several positive (negative) aspects. Most obviously in extreme FSE scenarios which are on the right side of the bar plot: For example, the last bar depicts students' willingness to select a very easy, very interesting course, with low workload from an excellent teacher. In such a scenario the willingness to select a course is approximately increased by 44% and it becomes clear which course students pick in case a course dominates on all, or at least, several dimensions. The same applies to the unfavourable decision scenario (see Treischl 2019a). The predicted willingness to select a course *hits rock bottom* in case all negative aspects come together and students' willingness to select a course drops to 6%.

The combination of positive or negative course attributes appears artificially, since most courses come not with (un-) favourable attributes only. Nevertheless, these findings from the FSE underline that students consider and weigh several course and teaching attributes, which makes clear that in some instances only a sub-sample of students may select the course, especially if a course combines several expected benefits or costs. From this perspective, student course selection may come with severe consequences for SET in terms of validity if course interest, workload or course difficulty are bias variables that shape students' perception of teaching quality. Therefore Treischl (2019b) asks about the causal impact of the considered bias variables.

1.3.2. It's a Bias, isn't it? The Causal Influence of Course Interest, Workload and Difficulty on SET

For decades SET research points out that many bias variables are correlated with SET, which might impair the measurement of teaching quality. The literature about validity concerns and bias variables in SET addresses heterogeneous aspects about the teaching – learning environment. Students may be responsive for attributes which are clearly not desirably in terms of measurement and fairness. For instance, students may be influenced by personality traits and attributes like charisma (Clayson & Sheffet 2006, Shevlin et al. 2000), physical attractiveness (Wolbring & Riordan 2016), or sexual orientation of the instructor (Ewing et al. 2003).

However, the current state of research is inconclusive for many bias variables. Some report that *workload* has a significant effect on SET (e.g. Centra 2003, Marsh & Roche 2000), while others do not (e.g. Dee 2007) and the same applies to several bias variable like *course difficulty* (Remedios & Lieberman 2008, Ting 2000), *physical attractiveness* (Campbell et al. 2005, Wolbring & Riordan 2016) or *prior course interest* (Feistauer & Richter 2018, Olivares 2001). From an analytical point of view, it is not even clear whether some bias variable like prior course interest shall be perceived as a bias variable or whether the association reflects teaching quality and students' perceived learning outcome (see Marsh 2007).

Research Contribution

In line with this argumentation Treischl (2019b) contributes to the current state of research, at least with two aspects. First, I make a methodological contribution by introducing an FSE in SET research (see Wallander 2009 for an overview of different research areas). I conducted an FSE and elaborate the methodological strength of the approach in comparison to non-experimental research designs. A factorial survey experiment combines advantages of survey research with the methodological rigour of an experiment in terms of causal inference (see Auspurg & Hinz 2015, Mutz 2011). And an FSE reduces concerns regarding students' self-selection, because participants (students) of an FSE are randomly assigned to varying stimuli. Thus, findings from the FSE are robust against concerns about students' self-selection and using an FSE instead of the SET survey precludes some methodological drawbacks of cross-sectional surveys[7]. Overall, I explore the causal impact of course interest, difficulty, and workload on students' perception of teaching quality.

Moreover, I make a theoretical contribution by disentangling the assumptions behind the considered bias variables. Student may consider prior course interest and accordingly give better grades (bias hypothesis). However, from a theoretical point of view an excellent instructor may increase both students' ratings and further educational outcomes, which may

[7] For example, we can rule out concerns of previous research that an association between a bias variable and SET is a spurious correlation caused by unobserved heterogeneity (see Marsh 2007: 352).

affect students rating and facilitates the teaching process. An excellent teacher may increase students' perceived learning and course interest (learning value assumption), which is why course interest and other bias variables may not be perceived as bias at all (see Greenwald & Gillmore 1997). All teaching dimensions in the FSE are uncorrelated due to the experimental design. This advantage gives me the opportunity to examine two mechanisms and to estimate the causal impact of a bias variable, independently from students learning and other dimensions of teaching quality.

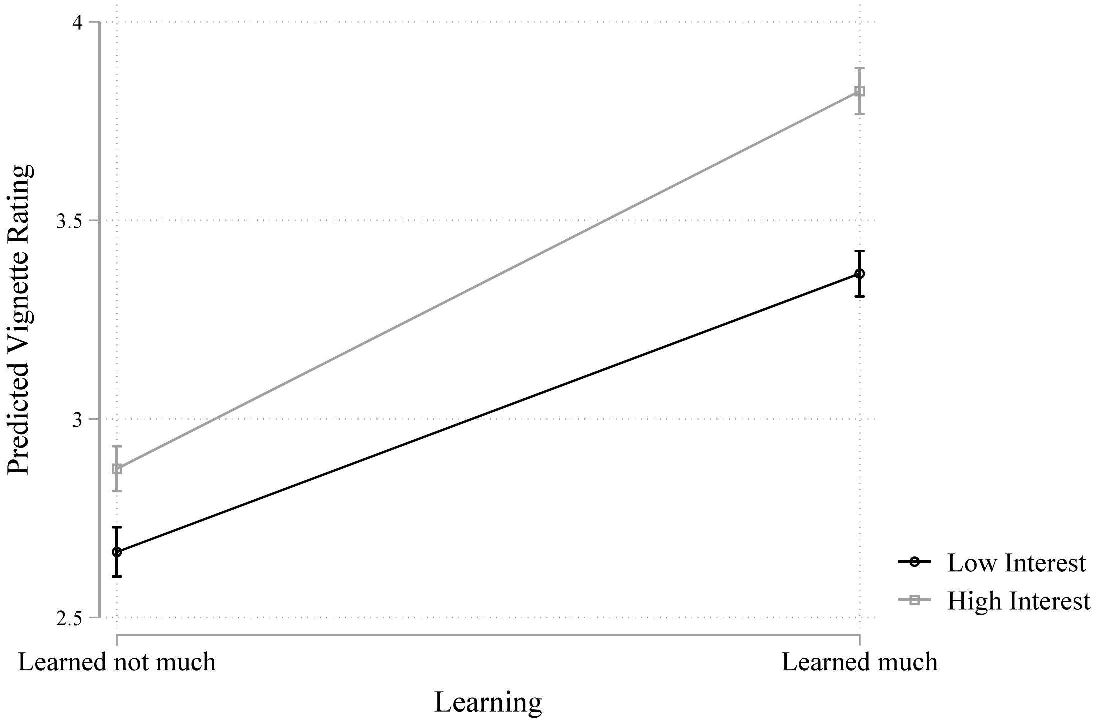
Main Findings

Course interest as well as workload significantly increases students' hypothetical course rating ([1.0] "insufficient"; [5.0] "excellent"). For example, a course description with the lowest level of course interest is rated with 3.03, while vignettes with the highest level of interest were rated with 3.43, all else equal. An interesting course or a course with lower workload gets on average a better course rating, but the effect sizes are moderate in both instances. On the contrary, course difficulty only marginally affected students' rating, even though the effect is significant. The experimental research design made it possible to estimate the effect independently of teaching aspects and the causal path of perceived *learning* outcome was blocked. Therefore, I concluded that *course interest* as well as *workload* are bias variables which shape students' perception of course quality.

Based on these finding one may conclude that the effect of the discussed bias variables are neglectable in terms of effect size. These findings do not imply that a moderate bias variable cannot distort the assessment to a substantial extend, in particular since several bias variables may affect students' rating simultaneously. Predicted vignette rating illustrates how the assessment is distorted by the three bias variables: An uninteresting course with high workload and a difficult content is rated on average 2.83, while the counterpart scenario (with high interest, etc.) gets - from a substantial point of view - significant better rating (3.52). The difference between these two rating scenarios is approximately 66% of a standard deviation.

In addition, I included the perceived learning outcome in the FSE to shed light on the association between students learning and the perception of the bias variables. For example, the effect of course interest increases in a high learning scenario, which is in line with the bias hypothesis. Figure 1.3 depicts the point estimates of the predicted values for the interaction between course interest and learning to provide a better intuition of the effect. Students give a significant better course rating and the effect of course interest significantly increases in case students learned a great deal in interesting course. An interesting course gets on average a 0.23 better rating, while the effect is twice as large (0.47) in a high learning scenario. This finding is in contrast with assumptions of the teaching effectiveness theory. Compared to observational SET data these results are not affected by an excellent teacher who did a better job due to students' course interest. Thus, these findings speak for the bias-hypothesis. It seems that students give better ratings if they learned a great deal about a topic which match their interest.

Figure 1.3.: Interaction between learning and course interest



1.3.3. Selection Bias in SET: Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings

In comparison to students' course selection, research and higher education must take into account that students often decide how regularly they attend a course. Wolbring & Treischl (2016) aim to understand students' absenteeism from class and concentrate on the question if there is selection bias in paper-and-pencil SET[8]. The aim of this paper is to raise awareness for the possibility of an attendance related selection bias in research on higher education and its consequences for measuring students' course satisfaction. To this end we ask what might explain students' absenteeism and what consequences has absenteeism for the measurement of teaching quality in terms of course ratings and rankings?

Research Contribution

Wolbring and Treischl (2016) empirically illustrate the problem of absenteeism by conducting a SET survey with two measurement points in time. In principal, we assume that *no-shows* are on average less satisfied with teaching quality and give worse grades, but the reverse causal mechanism may apply as well. Low attendance may reduce students' chance to catch up with the course content, which subsequently reduce students' course satisfaction. By conducting a SET with two measurement points in time we can circumvent concerns regarding reverse causality. The first SET were carried out at the beginning of the term and the second SET was part of the regular evaluation period in the last quarter of the semester.

Furthermore, the additional first SET gives us the chance to collect many information that might explain students' absenteeism at the regular evaluation. Both surveys contained a self-generated identification code. The identification code makes it possible to match cases and to identify observations which were filled out by the same student at each measurement, while preserving the anonymity of the participants (see Yurek et al. 2008). Therefore, we can test

[8] Wolbring & Treischl (2016) rely on SET data conducted at the University of Munich in the summer term 2012. This data set was used as basis for my thesis (Diplomarbeit) in 2013. However, Wolbring & Treischl (2016) substantially extend the current state of research on theoretical and empirical grounds, especially in terms of missing data and multiple imputation techniques.

which aspects (e.g. teaching quality) at the first measurement explain students' absenteeism at the regular evaluation.

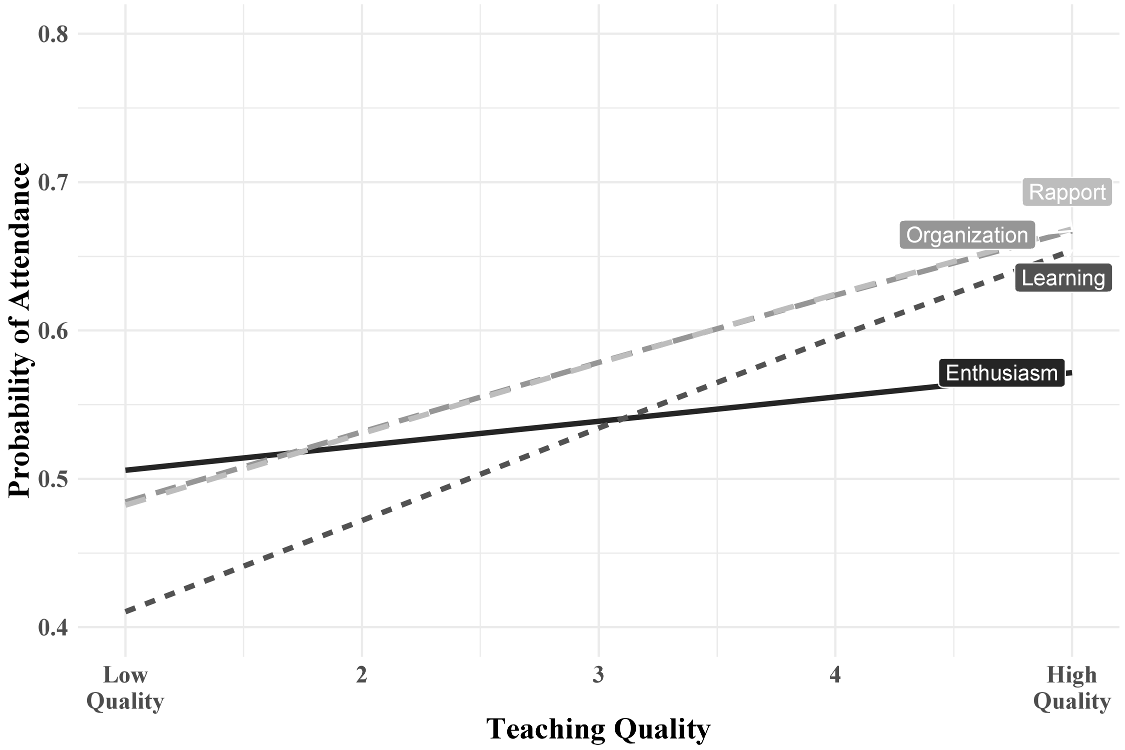
Finally, we use multiple imputation techniques to adjust for missing values. Based on the identified causes of students' absenteeism we use multiple imputation techniques to predict the SET outcome of absent students. We use information from several causes of absenteeism from the first measurement of time and predictors for teaching quality to calculate several imputed values, which were combined to a single course rating for absent students (for more information about multiple imputation see Enders 2010). Thus, we adjust regular SET results in case student did not show up.

Main findings

The results indicate that approximately one third of the students do not attend class at the second time of measurement, by which time the regular paper-and-pencil SET would take place. Based on the research design we test several hypotheses regarding absenteeism and show that course quality has a substantial effect on absenteeism. A high course quality significantly and strongly increases the chance that students will participate in the regular SET.

These findings can be illustrated with four independent dimensions of teaching quality (from the students' evaluation of educational instrument, see Marsh 1982). In addition to the reported main effect of teaching quality, Figure 1.4 depicts the predicted probabilities for students' absenteeism based on the result of a logistic regression with the assessment of four teaching dimensions as independent variables. In line with theoretical considerations regarding the multi-dimensionality of SET, teaching indicators for learning, organization and (individual) rapport significantly increase the chance that students will participate in the regular SET. For instance, the likelihood of course attendance is about 20-25% increased if an instructor is well organized or students have the impression that they learned a great deal. The only exception is enthusiasm, which does not significantly predict students' attendance. Nevertheless, these results underline the consequences of students' absenteeism for SET, but

Figure 1.4.: Influence of several teaching dimensions on course attendance



Note: This figure displays the predicted probabilities of course quality dimensions on students' absenteeism. Results are based on the logistic regression with the binary dependent variable attendance: 0 = absent; 1 = present at the second measurement time. Higher values indicate a better assessment (or higher agreement) in each dimension of teaching quality.

also in terms of convergent (and discriminant) validity since learning (as well as workload) substantially explains students' absenteeism.

Subsequently Wolbring & Treischl (2016) provide information about the consequences of data loss with regard to course ratings and rankings. Based on our sample we conclude that SET are a quite reliable instrument to assess quality of teaching at the individual level of single courses, but that course ratings are ill-suited to determine teaching quality in terms of rankings. This last point can be illustrated graphically with a histogram: The left panel of Figure 1.5 depicts the distribution of the course rankings based on SET results but displays how the course ranking changes ($|t_1 - t_2|$) due to students' absenteeism at the second time of measurement. The figure displays how many courses did not changed in the ranking (no bias), the number of courses that improve in the ranking due to absenteeism (positive bias), and courses with a worse ranking (negative bias). As Figure 1.5 indicates approximately 12% of the courses (3 of 26) receive the same course ranking (no bias), while the majority of courses improves, or gets a worse placement in the ranking due to absenteeism.

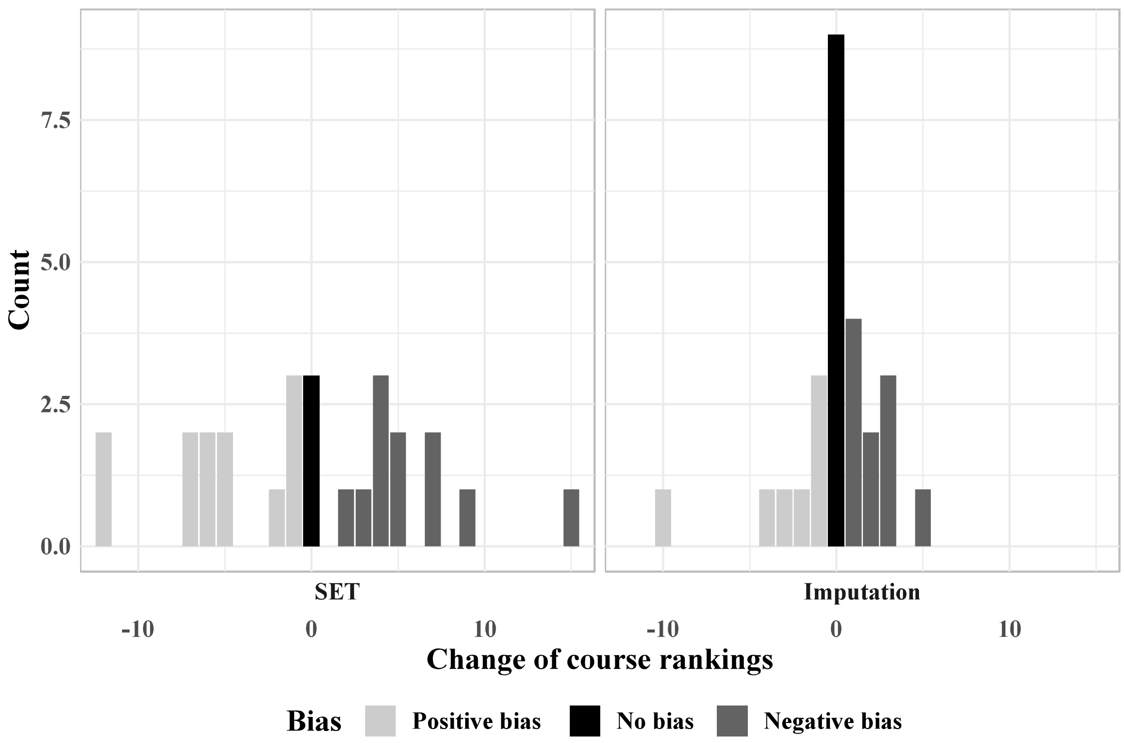
Last, we use multiple imputation techniques to adjust for students' absenteeism at the second point in time. Graphical results of the adjusted rankings are summarized in the right panel of Figure 1.5, which contains again the change of course rankings, but this time it is based on the adjusted course ratings from the imputed values ($|t_1 - t_{imputed}|$). The adjustment for absenteeism substantially reduces the bias in comparison to the unadjusted SET ranking. Approximately 31% of the courses (8 of 26) receive the same course ranking. However, the induced bias cannot be reduced for all courses and from a practical point of view an adjustment is costly. In particular with respect to an online SET survey mode, which takes into account that not all students are present at the day of survey. Online SET bring about additional methodological pitfalls, which is why Treischl & Wolbring (2017) ask about the causal effect of the survey mode in SET.

1.3.4. The Causal Effect of Survey Mode on SET: Empirical Evidence from Three Field Experiments

In recent years many universities switched from paper- to online-based SET without knowing the consequences for data quality. The current state of research is inconclusive, but most researcher find no differences between paper-based and online SET in terms of reliability and validity (for example Risquez et al. 2014, Stowell et al. 2012), while others report small (Morrison 2011, Nowell et al. 2010) or even pronounced differences (Meinefeld 2010, Morrison 2013). Previous research only agrees that online SET may *suffer* from low response rates, which is why many researcher and instructors fear that the survey mode has an impact on SET results and only a selective sub-sample of students will participate.

Treischl & Wolbring (2017) emphasize that this heterogeneous state of research needs to be considered at least with respect to three constraints: First, a large number of methodological studies are based on non- or quasi-experimental designs and/or small selective samples, which

Figure 1.5.: Distortion of course rankings due to absenteeism



impedes the question of a causal effect due to a change in survey mode. Consequently, large parts of the past and current literature on the topic only permit limited conclusions about the causal effect of survey mode on SET (see also Risquez et al. 2014: 130). Second, mixed findings in the literature might arise from minor but important technical differences in the actual implementation of online SET and could be a reason for the heterogeneous state of research[9]. Different online modes (TAN vs. email-based SET) may have quite different effects on course ratings – a fact which further aggravates the comparison of different studies, especially in the light of self-selection bias.

Ultimately, we stress that the existing research on SET mode effects do not adequately address the possibility of self-selection of students in the form of absenteeism from class which might affect classroom SET. It is not, as sometimes assumed, the case that in class paper-and-pencil SET cover the whole population while online SET do only represent a fraction of that

[9] Students are provided with a transaction number (TAN) in order to give them access to the SET; while email-based SET relies on course lists and each listed student gets an *invitation* to evaluate per email.

group. Results from paper-and-pencil surveys therefore cannot serve as a gold standard against which one can assess the quality of online SET. Given an all-encompassing email list, online surveys even offer a decisive advantage: all students, irrespective of whether they are present in class or not when the survey takes place, can be contacted by email and have an equal opportunity to participate in the SET.

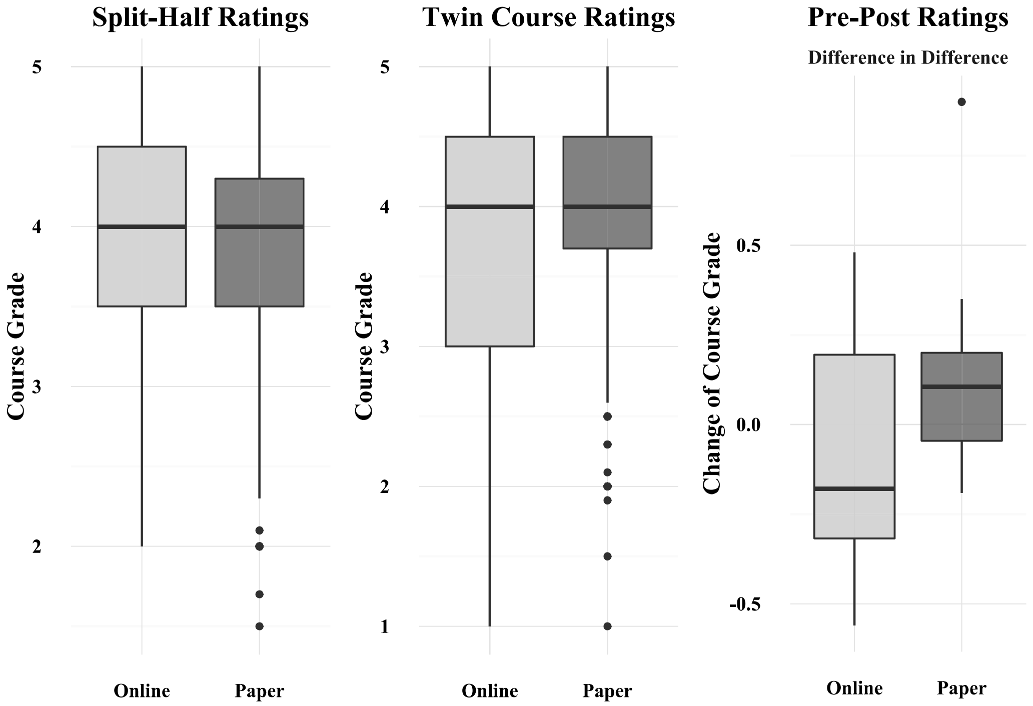
Research Contribution

This paper aims to contribute to the literature on mode effects on SET in three ways. First, we extend the current state of knowledge using three different field experiments – a split-half design, twin courses, and pre-post-measurements – with a randomization at the student and course level, which allows firmer conclusions about the causal effects of survey mode. Besides investigating effects on non-response rate and overall course rating, we also examine consequences for reliability and variance of student ratings. Second, we highlight the ample importance of distinguishing between TAN- and email-based SET when studying mode effects and show that the two variants of online mode have different consequences for non-response and teaching evaluation. This has so far not been adequately addressed in the literature. Third, inherently to the last contribution, we explicitly take into account the possibility of self-selection of students and show that email-based SET are helpful to include all enlisted students in the evaluation, regardless of the actual attendance. To show the consequences for methodological research, all instructors in the third field experiment were asked to provide students a *codeword*. Students were asked to fill in a course specific codeword in the SET, which gives us the chance to differentiate between students attending class and *no-shows* or irregularly attending students at the day of survey.

Main findings

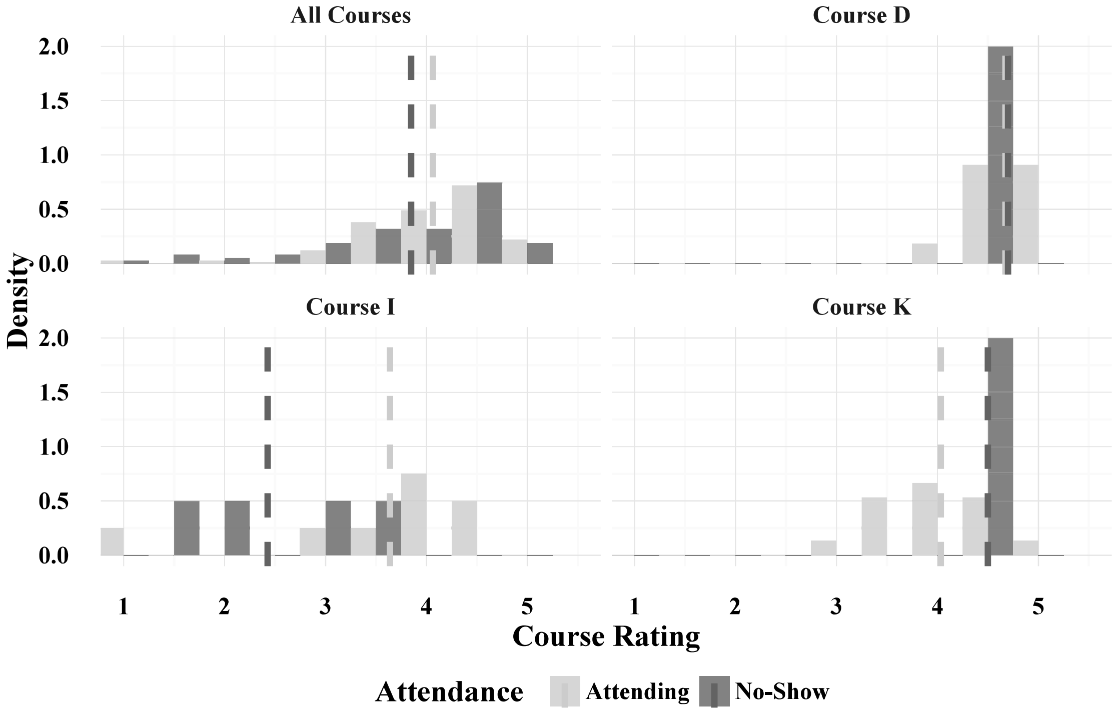
In conclusion, all three studies revealed marked differences in non-response between online- and paper-based SET and systematic, but small differences in overall course ratings. Figure 1.6 depicts average course ratings ([1.0] “insufficient”; [5.0] “excellent”) of all three field experiments. Overall, online SET reveal a slightly less optimistic picture of teaching quality in students' perception. For example, courses in the split-half design are on average 0.12 grade points better than their paper-based counterparts of the same lecture. Overall, differences between online- and paper-mode are in most conditions of the three field experiments small and not significant. These findings are in line with most of the previous studies on the topic. However, in contrast to many previous investigations the experimental setup of our studies allows us to draw firmer conclusions about the *causal* effect of survey mode on SET than quasi-experimental and non-experimental pre-post designs. Thereby, besides investigating effects on non-response rate and overall course rating, we also found that survey mode does not impair the reliability of online SET.

Figure 1.6.: Main findings of the survey mode experiments



Albeit these main findings, we demonstrated empirically that email-based methods account for students' attendance decisions. Figure 1.7 depicts the course rating by attendance (attending vs. no-shows) for courses with a code word (study III) as well as three courses for illustration purposes. At least for the sample under investigation we show that it is crucial to include irregularly attending students when evaluating teaching. As Figure 1.7 highlights, in some cases the course assessments from no-shows clearly deviates from the overall rating. In general, irregularly attending students perceive teaching quality slightly more critical than students in class (all courses). This does not imply that a selection bias has a small effect on courses ratings. Single course ratings can substantially be influence by irregularly attending students (see course I), and in addition, the assessment can be consistent (course D), better (course K) or worse (course I) than the assessment from attending students in class.

Figure 1.7.: Course rating by attendance



1.4. Implications for Higher Education

This dissertation highlights the scarcity of empirical research about students' self-selection effects in SET. This applies to the introduced literature about the reliability and validity of SET, but equally to research about bias variables and to a wide range of methodological research. Most research relies on SET data conducted via the regular SET procedure and most research does not take students' selection effects into account. The same applies to further research areas and quality assessment instruments in higher education (e.g. graduate surveys, programme evaluation). Besides the discussion about selection effects, this dissertation provides some practical recommendations and implications for higher education as well.

This dissertation focussed first on students' course selection. Irrespective of the methodological consequences for SET, course selection precedes course completion and withdrawal (e.g. see Babad et al. 2008). This dissertation examines several reasons why students may not attend a course on a regular base, but in conclusion lower attendance leads to poorer performance (Arulampalam et al. 2012) and in the long run reduces the chances to attain a degree (McKinney et al. 2019). Thus, how can students be supported to select a course that fits them? Often students have to select courses based on a vague idea about the course and its content, but in many instances, it would be easy to provide students with concrete information about the course and the specific requirements. For example, in a central online repository the syllabus and material from former course years can be made accessible to help students finding a suitable course. Certainly, this source of information is not useful in case a new course is taught or if the course instructor has changed. However, in many instances students get a precise picture about the specific course content from former years which may support students.

Next, this dissertation underlines from several perspectives that SET shall not be perceived as a numerically exact instrument to measure teaching quality. The first part of this dissertation emphasizes the role of course selection and the consequences of selective samples in SET. So far previous research asks whether students assess a course on fair grounds and many studies show that SET reflects teaching quality to a great extent. Or as Seldin mentions: *"the opinions of those who eat the dinner should be considered if we want to know how it tastes"* (Seldin

1993: 40). In principle one may agree with this statement, but in terms of fairness concerns remain if those who eat the dinner *can decide which cuisine they prefer*. In accordance with this thought this dissertation highlights that in some courses only a selective sample of students may show up and SET results can be distorted if the course selection is directly (e.g. teaching reputation) or indirectly (e.g. course interest) related to the measurement of teaching quality.

One may also argue that SET results may be distorted if those *who did not like the dinner decided to go home* before they were asked about their opinion. Therefore, the second part of the dissertation focuses on the consequences of students' absenteeism for paper-and-pencil SET and highlights that SET rankings are not well suited to identify *exemplary good or bad* teachers or courses. However, this does not imply that SET results are in general too optimistic about the teaching quality. In some classes students may no longer show up for reason which are unrelated to the measurement of teaching quality. In the latter case the SET results may even be more critical without the voice of the *uncritical, but unobserved* mass. In summary, SET captures teaching quality from students' perception, but in light of selection effects, bias variables, or low response rates it becomes clear that SET results are not numerically accurate. For this very reason we should reconsider the common practise to use SET as the basis for decisions about allocation of performance-related resources.

Ultimately, higher education may think about minimum standards in SET reports by providing information about the audience, survey mode and response rates in SET results. This can be underlined by the *trade-off* between paper-and-pencil and online SET regarding student selection and response rates. On average, a paper-and-pencil survey mode has a higher response rate but excludes all absent students at the day of survey. On the contrary, all participants have the same chance to participate in an email-based SET, but the latter may suffer from low response rates[10]. Having critical information about the course assessment certainly help *to evaluate* students' evaluation. Minimum standards in SET reports gives us at least an intuition

[10] This trade-off may not even be caused by the survey mode, but might be due to the survey procedure and asking students to evaluate in their leisure time. Response rates from online SET may align to paper-and-pencil SET if students have to fill out a paper-and-pencil SET at home. From this perspective online in-class SET might overcome this trade-off. Accordingly, Kuch & Roberts (2019) recently reported that response rates can significantly be increased via survey collection in class. Furthermore, previous research (e.g. Nowell et al. 2014) and this dissertation underlines that not a high response rates is necessary, but a non-selective sample of students participating in the survey.

about the goodness of the measurement to draw a clearer picture under which methodological conditions - and pitfalls - the assessment took place. This may also help instructors and other stakeholder to interpret SET results.

1.5. Research Contribution

Irrespective of selection effects, this dissertation also contributes by highlighting a causal inference paradigm in SET and higher education research. For example, Treischl & Wolbring (2017) use different experimental research designs in SET which illustrates the existence of harmless experiments without ethical concerns. Experimental research in (higher) education is often deemed as not applicable, unethical, or complicated to be implemented. In consequence, opportunities to conduct an experiment are frequently neglected. This dissertation emphasizes the advantage of experimental research designs, even though randomized trials are by no means a *panacea* for every research question[11]. In the same vein, SET research does not often rely on methodological and statistical techniques which point to causal inference. For example, longitudinal SET data helps to reduce concerns regarding unobserved heterogeneity (see Allison 2009); or the change of survey mode in different faculties may be perceived as a *natural experiment* if certain assumptions apply (e.g. as-if random assignment, see Dunning 2012). Thus, a causal inference paradigm in higher education and SET research adds and extends the current state of knowledge.

Besides the question about the causal impact of a bias variable, the results from Treischl (2019b) underline that factorial survey experiments can be used to estimate the impact of a *teaching dimension* (e.g. feedback) on students' course quality perception. Thus, as a by-product, the findings illustrate that factorial survey experiments can be used to extend our knowledge about students' perception of course quality. SET does not rely on a precise *theory of teaching effectiveness* and several different aspects need to be considered in order to measure teaching quality. For this reason, previous research considers SET as multidimensional

[11] It is beyond dispute that experimental research should also be assessed critically relating to possible methodological pitfalls, which might impair the validity and/or the generalizability of findings (see Shadish et al. 2001).

instrument, but there is no clear picture what defines teaching quality nor does a standard instrument exist that contains all important teaching dimensions for different disciplines or stakeholders (see Spooren et al. 2013: 7). Future research may extend a theory of teaching effectiveness by establishing a normative consensus about the impact of several teaching dimensions for different stakeholders. Hence, factorial survey experiments may provide an empirical basis to select indicators of teaching effectiveness, subsequently improve SET instruments by identifying (un-) important teaching dimensions, and deepen our theoretical understanding of teaching effectiveness (see Treischl & Hönig 2019).

Ultimately, Treischl (2019a) highlights that factorial survey experiments can be used to study a consideration process and to examine decision-making on hypothetical grounds. Research about (higher) education often deals with decision-making: Pupils are faced with the decision about their next educational step and the consideration of costs and benefits may explain why students from a low socio-economic background decide not to study (see e.g. Hout 2012). Furthermore, pupils may consider different university locations and programme to come to a decision. And students are faced with a new transition point due to the Bologna reform (e.g. Reimer & Pollak 2010). Thus, it is surprising that factorial survey experiments are rarely used in educational research (for recent exceptions see Di Stasio & Gërxhani 2015, Finger 2016), especially in comparison to other research areas and topic (see Treischl & Wolbring 2019, Wallander 2009). Certainly, a factorial survey experiment is not a *panacea* and observing actual behaviour is preferable then relying on hypothetical decisions. However, factorial survey experiments make it possible to examine the decision-making process on hypothetical grounds and are therefore a supplement for research with a behavioural benchmark, especially if an experimental research design is not feasible or applicable.