Research in Higher Education

Selection Bias in Students' Evaluation of Teaching. Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings

--Manuscript Draft--

Manuscript Number:	
Full Title:	Selection Bias in Students' Evaluation of Teaching. Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings
Article Type:	Original Research
Keywords:	Class Attendance; Missing Data; Students' Evaluations of Teaching (SET); Sample Selection Bias; Self-selection
Corresponding Author:	Tobias Wolbring ETH Zurich Zurich, SWITZERLAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	ETH Zurich
Corresponding Author's Secondary Institution:	
First Author:	Tobias Wolbring
First Author Secondary Information:	
Order of Authors:	Tobias Wolbring
	Edgar Treischl
Order of Authors Secondary Information:	
Abstract:	Systematic sampling error due to self-selection is a common topic in methodological research and a key challenge for every empirical study. Since selection bias is often not sufficiently considered as a potential flaw in research on and evaluations in higher education, the aim of this paper is to raise awareness for the topic. First, we describe student's selection decisions at different points of their studies and elaborate potential biases which they might cause. Then we illustrate the problem for the case of students' evaluations of teaching. We report findings from a design with two measurement points in time showing that approximately one third of the students do not attend class at the second time of measurement. Furthermore, the results indicate that the probablity of absenteeism is influenced by course quality, students' motivation, course topic, climate between instructor and class, course- and workload, and timing of the course. Although data are missing not at random, average ratings do not strongly change after adjusting for selection bias. However, we find substantial changes in rankings based on SET. We conclude from this that SET are a reliable instrument to assess quality of teaching at the individual level but are not suited for the comparison of courses.

Selection Bias in Students' Evaluation of Teaching: Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings

Tobias Wolbring and Edgar Treischl

Corresponding author:

Tobias Wolbring
ETH Zurich
Mühlegasse 21
8001 Zurich
Switzerland

Telephone: +41 44 632 4416
Email: tobias.wolbring@gess.ethz.ch

Edgar Treischl
University of Munich
Konradstr. 6
80801 München
Germany

+49 89 2180 2929
edgar.treischl@lmu.de

Noname manuscript No.
(will be inserted by the editor)

Selection Bias in Students' Evaluation of Teaching

Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings

Received: date / Accepted: date

Abstract Systematic sampling error due to self-selection is a common topic in methodological research and a key challenge for every empirical study. Since selection bias is often not sufficiently considered as a potential flaw in research on and evaluations in higher education, the aim of this paper is to raise awareness for the topic. First, we describe student's selection decisions at different points of their studies and elaborate potential biases which they might cause. Then we illustrate the problem for the case of students' evaluations of teaching. We report findings from a design with two measurement points in time showing that approximately one third of the students do not attend class at the second time of measurement. Furthermore, the results indicate that the probablity of absenteeism is influenced by course quality, students' motivation, course topic, climate between instructor and class, course- and workload, and timing of the course. Although data are missing not at random, average ratings do not strongly change after adjusting for selection bias. However, we find substantial changes in rankings based on SET. We conclude from this that SET are a reliable instrument to assess quality of teaching at the individual level but are not suited for the comparison of courses.

Keywords Class Attendance · Missing Data · Students' Evaluations of Teaching (SET) · Sample Selection Bias · Self-selection

1 Introduction

Systematic sampling errors due to self-selection of participants is one of the major methodological challenges of every empirical study. For example, endogeneity or selection bias is likely to occur in evaluations of career preparation courses. Students who didn't manage the transition into employment get support through such a program. It is possible that participants differ in

Address(es) of author(s) should be given

unobserved factors when compared to the group of non-participants who also didn't make the transition into the labor market: Motivated and ambitious students have a higher probability to visit these courses and a higher chance of finding a job. Without the randomization of the participants or other direct control over the selection process estimates are biased, imprecise, and cannot be generalized. In this particular case it is likely that the effect of the course is overestimated by simple comparisons between both groups and it is hardly possible to gauge the average causal effect of the intervention. Self-selection of participants is not just a methodological finesse, but an ubiquitous challenge for every empirical research which can dramatically change results.

In research on higher education different instruments to measure students' perceptions of teaching quality, such as student or graduate surveys and ratings and rankings from students' evaluations of teaching (SET), are well-established. Thereby, evaluation in higher education faces the same methodological problem: the possibility of systematic sorting and dropout. This is scarcely examined and discussed in research on higher education, not to mention applications of statistical procedures to reduce endogeneity problems or to estimate the range of selection-induced distortions.[1] Consequently, one has to fear that most evaluations of students' and graduates' perceptions of teaching quality are imprecise, biased, and not generalizable. For example, even though a myriad of studies (for reviews see Marsh 2007; Spooren et al. 2013) has investigated the reliability, validity, and fairness of student course ratings, SET results could be significantly influenced by sorting processes prior to and during the course: Students who are not interested (anymore) in the course are less likely to take the course and are more likely to be absent at the time of in-class evaluation if they are still enrolled for the course. Clearly, absenteeism leads to missing values in in-class SET. There are several reasons why students may not attend class at the time of SET. However, not all missing mechanisms are equally problematic for course ratings and rankings.

Nonresponse is neglectable if the withdrawal is random (*missing completely at random [MCAR]*; Little & Rubin 2002) and completely unrelated to course quality and factors associated with the latter. Such non-systematic fluctuations, like illness or the persual of other random occuring commitments, will not affect the measurement. If however the decision to miss class or withdraw from the course is based on characteristics of the student, instructor or course which are correlated with students' assessment of teaching quality, then the missings are not completely random *(missing at random [MAR]*; Little & Rubin 2002) and SET will be distorted if no statistical countermeasures are applied. In econometrics the latter is also referred to as a sample selection bias (Heckman 1979) and in causal graph methodology the problem is known as a collider (Pearl 2009) or endogenous selection bias (Elwert & Winship 2014). In the case of *MAR*, measures of central tendency and correlations are distorted, but statistical remedies like weighting, Heckman selection correction,

[1] Exceptions are Weiler & Pierro (1988), Becker & Walstad (1990), and Titus (2007).

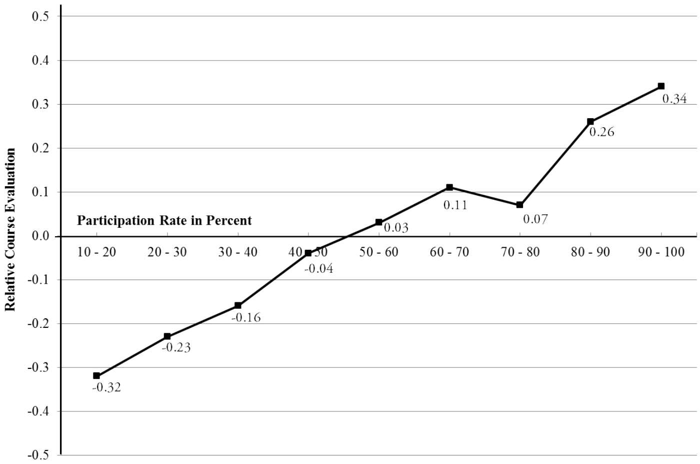

Fig. 1 Relationship between relative LVE evaluation and dropout rate
Note: SET data for 756 lectures over 23 semesters at a faculty of the University of Munich, which is kept anonymous on request. The participation rate is defined as the ratio of the number of participants in the survey and the number of participants at the beginning the lecture; the relative overall course rating is calculated as the difference between course evaluation (grade levels 1 "very good" to 5 "very bad") and average rating of all lectures at the Faculty over the evaluated period. Positive values for the relative evaluation imply ratings above and negative values ratings below average assessments.

propensity score matching, instrumental variables, event history modeling, and multiple imputation can correct for this bias if certain assumptions hold.

Finally, if quality of teaching itself or other unobserved variables, which correlate with teaching quality, influence the probability of absenteeim from class, data are *missing not at random [MNAR]* and, therefore, the missing mechansim is non-ignorable (Little & Rubin 2002). In this case of selection on unobservables statistical procedures become complex and are unlikely to provide unbiased estimates. This makes theoretical knowledge about the nature of the selection process essential and brings about the question of possible causes for student absenteeism.

Data from in-course SET at a faculty of the University of Munich, encompassing 23 semesters for 756 lectures, provide an initial insight whether SET data are missing at random or not at random. Figure 1 shows the relative overall evaluation of teaching quality plotted against the participation rate (% students which were present both at the beginning of the course and at the end-of-term SET). There is a very strong linear relationship between relative assessment and participation, where a higher participation rate is associated with a significantly better assessment. Classes with a dropout between 80 and 90 percent are rated poorer by .32 points on a 5-point-rating scale than average courses, while lectures with a dropout of 10 percent or less are .34 points better rated than the average. The maximum difference is more than half a grade point and, thus, also of substantial relevance.[2] Nonetheless, self-selection of students and the possible consequences for the validity of SET results are widely neglected in research on and evaluations in higher education, even though SET is gaining prominence in decision making (e.g., promotion, tenure, curriculum modification, teaching awards).

The aim of this paper is to raise awareness for the possibility of selection bias in research on higher education and its consequences for measuring students' course satisfaction. In Section 2 we give a short overview of common selection processes in higher education systems, including access to the university, choice of the field of study, and enrollment into courses. We then focus on the causes and consequences of students' class attendance in more detail in Section 3 and empirically ask for endogeneity bias in paper-based in-class evaluations in Section 4. The results of the SET study are summarized in Section 5, followed by a conclusion and a discussion of general implications for research and evaluation practices in higher education.

2 Selection Processes in Higher Education

Figure 2 provides an overview of relevant selection processes before, during and after the studies. It's straightforward to comprehend that pupils already specialize during high school, achieve a certain level of education, and then have to decide whether to study. Related to the latter decision the choices of subject and university constrain a set of available minors and areas of specializations. Of course, these decisions have to be made under restrictions such as admission criteria by the university or the state, available monetary resources, and geographical distance to one's family, friends or partner. However, future students are still free to choose between a variety of alternatives at this decision point. Furthermore, these decisions strongly influence, but do not fully determine which courses students take and how regularly they attend class. Thus, there is still much leeway for self-selection at each of these steps – assignment to universities, subjects, and courses is clearly not random. This becomes particularly clear if one keeps in mind that students might be dissatisfied or might fail examinations and reconsider their decision at each point in time. In Figure 2, this is indicated by arrows pointing back to previous decision nodes. In this section we depict the different selection steps before and after class attendance and their consequences for estimates of teaching quality in detail. In the next section we focus on the implications of class attendance and cancellation for course ratings and rankings.

Enrollment into courses at university constitutes a special form of decision-making. During their entire college life students have to make a series of in-

[2] These results are based on in-course evaluations and, thus, only encompass ratings of students which were present in class at the time of the evaluation. If our main hypothesis of selective, quality-induced migration holds than average course ratings reported in Figure 1 are positively biased and more strongly so, if dropout is high. Thus, these data very likely underestimate the actual relationship between both variables but at least give a lower bound of the strength of the association.

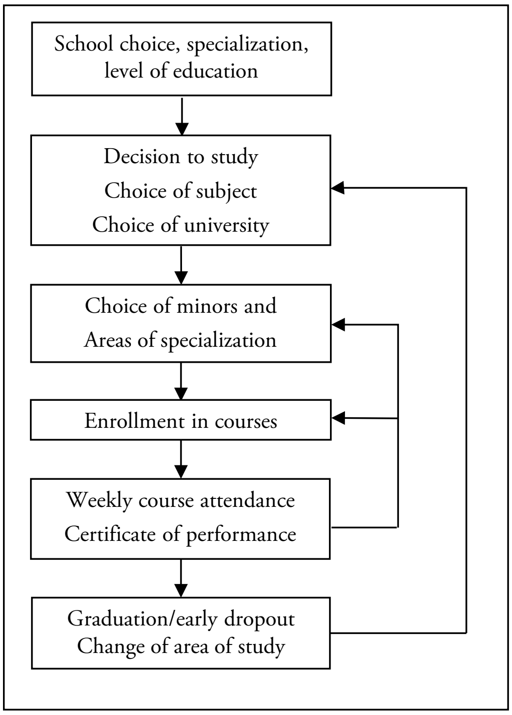

Fig. 2 Overview of important selection processes

terdependent decisions on which courses they will attend whereby they take their previous experiences into account (Babad 2001; Babad et al. 1999, 2008). The choice to enlist for a course is apparently restricted by the study program and the neccessity to pass the examination. Although this is a crucial factor in view of the intended academic progress, it is certainly not the only decision criterium. Students usually have certain degrees of freedom which elective courses they take and in which areas they want to specialize. Again, self-selection is likely and is partly driven by factors which might influence students' satisifaction with teaching.

Closely related to the topic of this paper Coleman and McKeachie (1981), Leventhal et al. (1975, 1976), and Wilhelm (2004) report that published SET influence course selection and Babad and Tayeb (2003) find that students tend to avoid difficult courses and take anticipated learning progress into account. Furthermore, Babad et al. (2008) show that students' interest in a topic is an important determinant of course choice. Dimensions of intrinsic motivation not only include "academic curiosity" for scientific topics, but also personal values and internalized social norms. For example, the consolidation of knowledge, mastering different perspectives of a phenomenon, or the practical relevance of a certain topic are related to students' course choice.

In addition, the timing of courses plays a central role for enrollment as a survey among undergraduate students by Babad (2001) illustrates: In planning their timetables students must integrate various major and minor courses, including those at different faculties. They also may need to coordinate work and study and consider private obligations. Examining time allocation effects Devadoss and Foltz (1996: 504) came to the conclusion that students prefer shorter courses (50 minutes) on a total of three days, rather than distribute the hours on the two remaining days with lectures of 75 minutes each. Moreover, they estimate a higher demand for courses on Monday, Wednesday, and Friday, although from a theoretical point of view, one would expect a preference for courses between Tuesday and Thursday. In line with this the authors report a "prime-time" effect: courses between 10 a.m. and 3 p.m. are attended more frequently than courses outside of this period. Thus, timing of a course influences the composition of the participants and the frequency of class attendance – both factors which determine whether a person completes an in-class SET and how she rates the course.

Not only while studying, but also when leaving a study program, selection processes in the form of graduation, retention, and change of subject/university are present. Especially the effects of motivation, satisfaction, and achievement on the likelihood to complete one's studies clearly illustrate why graduate surveys are an imperfect instrument for measuring the quality of teaching in students' perception and why one should be careful to generalize findings to whole student cohorts: If the group of graduates systematically differs from dropouts in terms of unobserved heterogenity (e.g. motivation, satisfaction, success) which correlates with the assessment of the study program and teaching quality, then the results of graduation surveys will be positively biased. Confounding is particularly likely if, inter alia, the social background, financial aspect, and exam performance increase the chance to dropout. Thus, a direct implication from these considerations is that one should not restrict the sample on graduates but better survey cohorts of first-year students and collect longitudinal data on their study path, attainment, and achievement. [3]

To sum up, students who are enrolled in a class are already a selective group. On the one hand, self-selection takes place ex ante in the form of choices of university, subject, and courses. On the other hand, course participants are also not representative for an initial study cohort because of student retention and early graduation.

[3] Apart from this "survivor bias", graduate surveys face further methodological challenges. First, institutional raw data usually provides information on students' addresses before or at the time of graduation. Obviously, as graduation is a decisive event in the life course, addresses are quickly becoming outdated. Since particularly successful graduates will, ceteris paribus, enter faster into the labor market than their less successful counterparts, graduate surveys are vulnerable for a "mobility bias". Second, on might also expect a 'bias by success": respondents who already achieved a successful launch into working life are more willing to participate and may be over-represented.

3 Course Attendance and Cancellation

There is a large number of empirical studies related to student absenteeism and course cancellation. However, the possible consequences for reseach on higher education are not sufficiently considered. A few weeks before the semester ended, Romer (1993) examined student attendance and came to the conclusion that: "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). Stanca (2003) reports similar attendance rates for Italian universities, while Wyatt (1992), Kirby & McElroy (2003), and Wolbring (2012) report more conservative, but still substantial estimates for the U.S., Ireland, and Germany.

In line with Figure 1, several studies document a significant association between students' perceptions of teaching quality and the frequency of individual absenteeism from class (Berger & Schleußner 2003; Romer 1993; Wolbring 2012). Although these are correlational studies which cannot finally rule out the possibility of reverse causation (attendance positively influences teaching quality and its perception, and not the other way around) the findings do at least not contradict the hypothesis that dissatisfied students withdraw early from the course or attend it less frequently. Hence, SET data are most likely missing not at random.

Findings by other empirical studies further corroborate this: Devadoss & Foltz (1996) find that attendance is significantly higher in courses taught by instructors who previously won a teaching award. And according to Greimel-Fuhrmann & Geyer (2003) and Reed (1981) students attend classes more regularly if they like the instructor or her teaching style. For all this, one must expect that SET are positively biased: Students present during the SET are more motivated, interested, and satisfied, and therefore, rate the quality of teaching significantly better than absent students would have, if they were still present.

Moreover, class attendance is often important to achieve academic progress. First, in some classes attendance is mandatory: Students will not receive credit points if they are absent too frequently. Thus, selection processes and resulting bias might be totally different in courses, in which presence is expected or even enforced, compared to non-mandatory courses. This does not only affect the validity of SET ratings, but also complicates comparisons and rankings based on SETs. Second, absenteeism increases the risk of missing relevant information and subsequently increases the chance of failing tests due to lacking knowledge about deadlines, assigned tasks, or exam contents. Based on previous experiences and information provided by the instructor, students will form subjective expectations about the importance of course attendance for passing the exam successfully. For example, previously failing an exam due to absenteeism might change beliefs about the importance of attendance and induce behavioral changes. In addition, students' grade aspirations as well as their interest in the course content and their study-specific values certainly play an important role and moderate the effect of examination relevance. Interest in the subject matter will obviously have a stronger effect on attendance if attendance is not mandatory or not important for a successful completion of the study program.

In line with the previous argument, Romer (1993) and Wolbring (2012) report higher attendance rates in courses which require a mathematical background. It goes without saying that benefits from class attendance depend on the nature of the course content. Students do therefore not only tend to elect and complete courses which they regard as interesting, but also attend them more frequently. However, the finding that classes, which require a formal background and which most students do not regard as rather inspiring, are attended more frequently begs for an explanation. At first glance, a math effect seems plausible only for those study programs that strongly rely on mathematics. For example, economic students in Romer's study are a selective group in need of profound mathematical knowledge. On the contrary: Students lacking profound math skills cannot efficiently compensate absenteeism by self-study and may attend these courses regularly to compensate this deficit. According to Bratti & Staffolani (2013) attendance increases students performance in exams, especially in quantitative subjects like Mathematics or Economics, while self-study time is essential for non-quantitative disciplines. Therefore, we propose the hypothesis that students especially attend quantitative courses more frequently if they lack knowledge and background on the subject matter (see also Douglas & Sulock 1995: 107).

Academic success is not solely an intellectual challenge, other skills are equally demanded in order to succeed the exams. Study-specific values, like perseverance and the attitude to work, can be subsumed as capabilities which influence both the success of a study and the willingness to attend a course. Students have their distinct ideals of study behavior, which reflect both socialisation and intrinsic motivation. The realization of these values in the daily routine can be understood as an idealization of a student's role which may include a highly positive picture of knowledge and learning, priority of university towards leisure activities and work, and the desire to perform well in exams. Hence, we assume that the more students have internalized these values, the greater will be their inclination to be present in class in general and at the time of the SET. We expect that this type of student, which is overrepresented in SET data, will also rate teaching more favorably.

Furthermore, according to Romer, course size has a substantial effect on attendance. He states that "absenteeism is considerably lower in the smallest third of classes than in the largest third" (see Romer 1993: 168). As well, Becker and Powers (2001), Marburger (2001) and Wolbring (2012) report significant associations between course size and average attendance rates. Two theoretical arguments help to explain these findings: First, large class sizes might directly influence teaching quality and facilitate withdrawal due to dissatisfaction and anonymity. Second, and more importantly: It is more difficult for an instructor to sanction individual absenteeism in larger groups. Hence, especially in larger classes and other courses in which mandatory attendance rules are not enforced students will not attend class if they are desinterested in the course content, dissatisfied with the teaching by the instructor, or anticipate a greater benefit from other activities (such as work, leisure, self-study, and learning for the exam). Thus, selection processes and biases likely differ between courses of different size as well as between courses with different content. Again, this makes comparisons of non-identical courses on the basis of SET problematic.

Last but not least, students also make their attendance decision in the light of given time and financial[4] ressources. Being employed requires the coordination of (possibly conflicting) demands from both study and work. As a result employment might negatively affect class attendance. Time investments, like the preparation for and the follow up of a course, are also crucial for attendance. First, in the face restricted resources we expect that students have a great interest to allocate their limited time efficiently (see Arulampalam et al. 2012; Dolton et al. 2003; Schmidt 1983). Consequently, the allocation of the available time aligns with student expectations about the requirements of a course. Second, time investment can also be considered as an indicator for student motivation and the subjectively anticipated benefits from participation. Students do therefore not only self-select by deciding which courses to take and attend, but also by the choice of the intensity of the "treatment". Strongly motivated and interested students will thus likely spend more time on the course material. We also expect this due to a third, related reason: Previous investments are lost if students do not pursue the course anymore. In contrast to many other investment situations, previous efforts should not be considered as sunk costs, since they increase the likelihood to succeed in the exams. Therefore, it is assumed that the larger the time investment, the higher the frequency of attendance and the chance to complete the course. In addition, we expect that the time of a specific course matters per se: being absent is more likely if the class is the first or the last of the day, or if there is a longer break before the course.

4 Empirical Application Using Students Evaluations of Teaching

After this theoretical underpinning of self-selection processes relevant for student and graduate surveys we corroborate our argument with a study on selection bias in SET. Before presenting empirical results on determinants of absenteeism and its consequences for SET ratings and rankings, we describe our dataset and explain the operationalizations of the theoretical constructs.

4.1 Data and Descriptive Results

We empirically examine the causes and consequences of self-selection for SET data at the Faculty of Social Sciences at the University of Munich. Using

[4] In 2009, 87% of the students in Germany were financially supported by their parents, 65% worked besides their studies, and another 29% received state funding (Isserstedt et al. 2010:193).

a design with two measurement points in time the attendance behavior of students at the regular time of evaluation as well as its determinants can be investigated. For this purpose we asked 29 lecturers to conduct an additional evaluation at the beginning of the semester.[5] In only 2 out of the 29 courses we were unable to collect data due to time scarcity or health problems. In addition, we have to exclude one course from the analysis due to incomplete data for the first SET. We do not expect this missing cases to be systematic. However, it is important to note that the selected courses do not represent a simple random sample, since we conciously chose them on the basis of the following three criteria: First, to ensure courses with a sufficiently large audience number, one third of the courses are lectures. Second, both a sufficient number of elective and compulsory courses was taken into account. Third, in order to avoid unit and item non-response due to survey fatigue (see Adams & Umbach 2012), we ensured that students had to complete the additional questionnaire for a maximum of three courses.

SET were carried out at the beginning of the summer term 2011 (t = 1: 3–4 week) and during the regular evaluation period in the last quarter of the semester (t = 2: 9–10 week). After the first evaluation instructors received no feedback in order to avoid reactivity. In addition to the measurement of teaching quality from regular SET, the questionnaire for the first survey included numerous items on a variety of student attributes which are theoretically expected to influence class attendance. In order to avoid missing important determinants of class attendance we additionally conducted eight semi-structured interviews with students from different backgrounds and with different progress in their studies. The results corroborated our theoretical analysis as outlined in Section 3. We include the following variables measured at t = 1 in the analysis (for descriptive statistics see Table A1):

- *overall course rating*: “All in all, how do you rate the overall quality of this course?” excellent [5.0] – insuffcient [1.0] (continuous),
- *prior interest in course topic*: “I chose the course, because I was interested in its content” totally agree [5] – totally disagree [1],
- *study-specific values*: index consisting of three items “To be present (at a course) is taken for granted”; “When the weather is fine I will not cancel the course for leisure activities” and “After a short night I will not stay in bed” totally agree [5] – totally disagree [1] (Cronbach’s $\alpha = .71$; interitem correlations: .43 – .52, $p < .05$),
- *course preparation*: “On average, how many minutes per week did you prepare for the course?”,
- *exam relevance of attendance*: “Course attendance is not important to pass the exam successfully” totally agree [5] – totally disagree [1],
- *bad exam performance due to absenteeism*: “I have previously performed bad in an exam due to absenteeism.” totally agree [5] – totally disagree [1],

[5] At this point we would like to thank both the lecturers as well as students who patricipated. This study could not have been realized without their support and cooperation.

- *recognition of absenteeism*: "Instructor will notice if I am absent." totally agree [5] – totally disagree [1],
- *course in quantitative methods*: non-quantitative course [0]; quantitative course [1],
- *student-instructor-climate*: "The climate between instructor and students is positive." totally agree [5] – totally disagree [1],
- *student-class-climate*: "The climate between the students in this class is positive." totally agree [5] – totally disagree [1],
- *first/last class of the day*: no [0]; yes [1],
- *break before course*: "If this is not the first class of the day: How many minutes break do you have before this course?",
- *courseload*: "How many credit hours per week do you take in this semester?", and
- *workload*: "How many hours per week do you work for payment in this semester?".

Both surveys contained a self-generated identification code.[6] This allows us to identify questionaires at t = 1 and t =2 which were completed by the same student, while preserving the anonymity of the participants. We manually matched the data from both surveys.[7] Using the idenfication code an unambiguous match was realized in 86.3% of the cases. Additionally, based on typeface analysis 13.7% of the observations were assigned to each other if there were only slight differences in the identification code (e.g., gender information was provided in the first, but not in the second survey). The subsequent calculations are based on this matched data set which includes 2.263 observations for 1.569 students.

The idenfication code allows us to determine whether students were present at both points in time or missed one of the two evaluations. About 53% of the students which participated in the first survey were absent during the regular evaluation and 28% of the students present at the t = 2 were absent at the first evaluation. Only 47% (72%) of the particpants at t = 1 (t = 2) were present at the second (first) evaluation ($r = .66$, $p < .001$ for ratings at t = 1 and t = 2). However, these estimates might be distorted by two types of matching error. The first error occurs if one does not recognize that two questionaires were completed by the same person. This leads to an overestimation of actual dropout. A second error occurs if one falsely assigns two SET to the same rater which results in an underestimate of actual dropout. Since participants may

[6] The self-generated identification code was based on person-specific and time constant information such as the first two letters of mother and father first name, number of older sisters and brothers, etc. For general information on identification codes see Kearney et al. 1984; Yurek, Vasey, and Havens 2008.

[7] An alternative approach to this matching task is the use of sequence analysis. Thereby, one first calculates similarity measures and then matches observations for which the similarity measure exceeds a certain threshold (for introductions into sequence analysis see Abbott/Tsay 2000; Taris 2000; for an application using identification codes see Schnell et al. 2010).

refuse to provide personal information[8] or provide inconsistent information, Type I errors are more likely in our application than Type II errors. Hence, in addition to the previously reported upper bound of dropout, an estimate for the lower bound of absenteeism is instructive, which is simply based on the number of completed questionaires at t = 1 and t = 2: According to this measure the dropout rate is at least 35.5% which closely corresponds with previously mentioned results for Germany and other countries.

Likely, the actual absenteeism at SET lies somewhere between 36% and 53%. We conclude from this that the number of participants significantly declined over time and that the composition of the courses substantially changed. We now ask for the determinants of class attendance on the student level and then study the consequences of absenteeism for ratings and rankings on the course level.

4.2 Causes of Absenteeism

In this section the binary outcome variable of interest is whether a student who was present at t = 1 completed a questionaire for the same course at the t = 2 (1 = yes; 0 = no). Since we are interested in the determinants measured at t = 1 for absenteeism at t = 2 we restrict the analysis to the 1335 students present during the first evaluation. Missing values at t = 1 are spread throughout the data with item non-response amounting .7% – 8.1% which would reduce the number of effective cases in the full model by 26.4% and therefore the power of our analysis if we applied listwise deletion. Hence, we relied on feasible information maximum likelihood to multiply impute missing values (see van Buuren 2012; Enders 2010).[9]

Table 1 summarize the results from logistic regressions with standard errors clustered around courses. In model 1 we look at the bivariate association between perceived course quality and absenteeism. Overall, we find a highly significant association. In line with the initial hypothesis, the poorer the course assessment is the smaller is the chance that students actually attend the course at t = 2. The predicted probability of attendance at t=2 declines by 31 percent points if the overall course rating goes from excellent (46.2%) to insufficient (69.4%). We conclude from this that in agreement with Devadoss & Foltz (1996: 504) "a good teacher can make a significant difference(...)", although further analyses on course level shows that this effect is not an equally decisive

[8] For the self-generated identification code the missing data problem seems neglectable. From an overall 2.263 single observations only 40 gave no personal information at all.

[9] We used the Stata ados *ice* (Royston 2005) and *mim* and included the outcome variable, all independent variables of our analyses, and some additional variables measured at t = 1 as predictors. Ordinal variables were estimated with ordinal logistic regressions, all other variables with simple linear regressions with 100 imputations. To test for the robustness of our approach, we additionally imputed the missing data using multivariate normal regression (Stata command *mi impute mvn*). Results did not change. Moreover, we compared our results with estimates based on listwise deletion (N = 1056). Regression coefficients and p-values changed, but the substantial findings remained untouched.

factor for all courses and is less important than other course characteristics. From a methodological standpoint this finding implies that SET data are missing not at random and are, on average, positively biased.

Controlling for some important determinants of overall course ratings in model 2 the latter variable looses explanatory power. This is mainly due to two important predictors of absenteeism: internalized study-specific values and time invested into the preparation for the course. Thereby, especially having internalized study-specific values strongly increases the likelihood to attend class at t = 2. Both effects are robust towards the inclusion of further controls in the following models.

In a next step, we include indicators for social control and exam relevance of class attendance into the model. Directly asking students for their subjectively perceived importance of attendance for successfully completing the exams proves of explanatory value. Surprisingly, having performed bad in previous exams due to absenteeism does not go along with a reduced, but with a significantly increased absenteeism rate in the last quarter of the semester. We do not interpret this increase (by up to almost 12 percentage points) as a causal effect of bad experiences, but rather as the result of a pretty stable trait: Students who skipped many classes in the past are more likely to do this in the future. Additionally, if social control is present ("Instructor will notice if I am absent.") the probability of absenteeism decreases by up to 17 percentage points, although the effect is not significant. Further analyses show that, as in previous studies, course size significantly affects attendance. However, this effect is mediated by perceived social control and disappears if we control for the latter.

Furthermore, our findings corroborate previous results on better attendance of courses which require a mathematical background. According to our results the probability that an average student is absent at t = 2 is 17 percentage points smaller in quantitative than in non-quantitative courses. Additional analyses show that this effect is mainly driven by perceived course difficulty: Students in quantitative courses significantly more often report that the content of the course is too complex for self-studies ($r_{Spearman}$ = .194; p < .001). As a result the effect of course type becomes markedly weaker if we include this indicator for course difficulty as a covariate.

Model 4 additionally contains indicators of social integration. A positive climate between the instructor and the class strongly reduces absenteeism: Students are by up to almost 20 percentage points less likely to miss the SET as compared to courses with a negative atmosphere between the students and the instructor. In contrast to that, the climate among the students does not significantly affect class attendance.

Finally, we adjust for students' time constraints in model 5. More courseload and workload negatively affect attendance, presumably reflecting the trade-off between time investments required for different courses and between studies and work. Students are also more likely to miss class if the course is the first or the last of the day. In contrast to that, the length of breaks before a course does not significantly influence absenteeism.

Table 1 Logistic Regression

Outcome: Absenteeism t=2	Model 1	Model 2	Model 3	Model 4	Model 5
Course rating at t = 1	0.784 **	0.850*	0.929	0.868^{+}	0.912
(1.0[–]—5.0[+])	(-3.026)	(-2.168)	(-0.943)	(-1.852)	(-1.004)
Prior interest in topic		1.111	0.993	0.988	0.987
(1.0[–]—5.0[+])		(1.373)	(-0.105)	(-0.184)	(-0.205)
Study-specific values		0.646***	0.688***	0.676***	0.677***
(1[–]—5[+])		(-5.262)	(-3.872)	(-3.921)	(-3.863)
Course preparation		0.838**	0.883**	0.888**	0.878**
(in log minutes)		(-2.805)	(-2.811)	(-2.666)	(-2.808)
Bad exam performance due			1.131*	1.116^{+}	1.108^{+}
to absenteeism (1[–]—5[+])			(2.157)	(1.894)	(1.739)
Exam relevance of			0.898*	0.894*	0.887*
attendance (1[–]—5[+])			(-2.166)	(-2.169)	(-2.398)
Course in quantitative			0.494**	0.470***	0.485**
methods (1 = yes)			(-2.919)	(-3.443)	(-3.232)
Recognition of Absenteeism			0.868	0.875	0.888
(1.0[–]—5.0[+])			(-1.558)	(-1.446)	(-1.322)
Student-instructor-climate				1.191	1.180
(1[–]—5[+])				(1.527)	(1.436)
Student-class-climate				0.810**	0.819**
(1[–]—5[+])				(-2.978)	(-2.821)
First/last class of the day					1.216^{+}
(1 = yes)					(1.925)
Break before course					1.002
(in minutes)					(0.978)
Courseload					1.026*
(in hours)					(2.058)
Workload					1.019**
(in hours)					(2.989)
Constant	2.891**	11.758***	17.101***	28.484***	10.369***
	(3.165)	(5.558)	(4.872)	(5.963)	(3.429)
N	1335	1335	1335	1335	1335
McFadden R^2	0.005	0.058	0.084	0.088	0.095
Log-likelihood	-917.6	-868.9	-845.2	-840.9	-834.7
AIC	1839.2	1747.9	1708.4	1703.8	1699.4

Note: Logistic Regression with the binary dependent variable absenteeism: 0 = present; 1= absent at the t = 2. Missing values multiply imputed (m = 100) with the Stata ado *ice* and results pooled with the Stata ado *mim*. McFadden R^2, log-likelihood, and AIC were calculated separately for each imputed datasets and averaged over all 100 models. Reported are *odds-ratios* and *z-values* in parentheses; $^{+}$p<.10, *p<.05, **p<.01, ***p<.001

4.3 Consequences of Absenteeism for Ratings and Rankings

Considering the uncertainty induced by the reduction of the audience and the change in its composition, it must be scrutinized whether SET results can be regarded as reliable and valid measurements of teaching quality in students' perception. To deepen this concern, we study the influence of student withdrawal on course ratings and rankings. Besides comparing ratings and rankings at t =1 and t = 2 we multiple imputed missing values at t = 2 for those students who were present at t = 1. To predict the overall course rating we used the following information: course rating at t = 1, change of individual course rating from t = 1 to t = 2 for those students present at regular SET, prior interest in course topic (t = 1), course choice because of instructor (t = 1), instructor highlights important points (t = 1), instructor teaches with enthusiasm (t = 1), course difficulty (t = 1), and self-assessed learning progress (t = 1) (R^2 = .57).

Table 2 shows course ratings and rankings for SET at t =1 and t = 2 as well as for imputed values. Contrary to our initiale suspicion, that especially dissatisfied students are no longer present and SET are overly positive at t = 2, the ratings do not indicate a clear trend. Although on average overall ratings slightly improved by .08 grade points from the first to the second measurement, at the course level all three possible patterns can be empirically observed: Ratings improve by more than .1 grade points for 41%, are stable for 23%, and worsen by more than .1 grade points in 36% of the cases. On average, the absolute value of the deviations between ratings at t = 1 and t = 2 (imputation and t = 2) is .23 (.11). We conclude from this that on average measurements are reliable and are not to substantially biased by self-selection. However, changes over time can be markedly stronger for individual cases as the results for course 5 and 21 underline.

The findings are different for SET-based course rankings. Due to changes in student ratings and due to self-selection the overall order of courses becomes unstable. The fact that direction and extent of selection bias is not constant over different courses makes the problem even worse. Only three classes (course 1, 25 and 26) obtain the same rank. The remaining 23 courses change substantially in the ranking. Since rankings are a zero-sum game, some courses deteriorate in their rank, while others ascend in the rank order. On average the position of a course changes by 5 ranks from t =1 to t = 2, with an observed maximum of up to 15 ranks (course 5). This average rank change might seem neglectable at first glance, but has to be put in context: the maximum number of rank changes that is mathematically possible for a ranking of 26 courses is 13. The observed average number of rank changes between t =1 and t =2 (t = 2 and imputed values) amounts to 38 (21) percent of maximum variation in the ranking and therefore is quite substantial. In addition, rank changes can be of practical relevance for individual courses: for example, according to the regular SET course 3 is among the top 3 and might possibly receive a teaching award whereas it only ranks in the upper midfield of rankings based on the other two indicators.

5 Conclusions and General Implications

The aim of this article is to raise awareness for self-selection in research on higher education and its practical consequences for the measurement of teaching quality in students' perception. In survey research selection bias is one of the main problems and this is discussed extensively in methodological literature. Introductory textbooks of quantitative social research routinely include

Table 2 Comparison of the Course Ratings and Rankings

	SET			Ranking			Rank Change	
Course	t_1	t_2	imp.	t_1	t_2	imp.	$\|t_1 - t_2\|$	$\|t_2 - imp.\|$
1	1.52	1.48	1.53	1	1	1	0	0
2	1.65	1.49	1.55	5	2	2	3	0
3	1.93	1.56	1.77	7	3	7	4	4
4	1.95	1.57	1.72	8	4	5	4	1
5	2.14	1.62	1.98	20	5	15	15	10
6	2.05	1.65	1.74	15	6	6	9	0
7	1.53	1.69	1.70	2	7	4	5	3
8	1.55	1.69	1.66	3	8	3	5	5
9	2.06	1.70	1.88	14	9	12	5	3
10	2.08	1.76	1.85	17	10	10	7	0
11	1.60	1.78	1.81	4	11	8	7	3
12	2.12	1.82	1.84	19	12	9	7	3
13	1.71	1.89	1.87	6	13	11	7	2
14	2.10	1.89	1.91	18	14	13	4	1
15	2.04	1.98	1.96	13	15	14	2	1
16	2.01	2.00	1.99	10	16	16	6	0
17	2.05	2.07	2.18	16	17	18	1	1
18	2.03	2.13	2.04	12	18	17	6	1
19	2.55	2.20	2.26	24	19	21	5	2
20	2.35	2.25	2.23	22	20	20	2	0
21	1.97	2.31	2.19	9	21	19	12	2
22	2.18	2.34	2.27	21	22	22	1	0
23	2.01	2.44	2.36	11	23	24	12	1
24	2.39	2.52	2.31	23	24	23	1	1
25	2.73	2.53	2.50	25	25	25	0	0
26	2.86	2.58	2.59	26	26	26	0	0
Mean	2.04	1.96	1.99	–	–	–	5.0	2.7

a warning that treatment effects in evaluation studies might be over- or underestimated if assignment to the treatment or control group is not in the hand of the researcher and is instead driven by the respondents themselves. This applies equally for research in higher education as we demonstrate with an empirical application using SET. Relying on an additional evaluation of each course at the beginning of the term, we find that approximately one third of the original audience do not participate in the regular SET during the last quarter of the term. Further, our regression analyses show that course quality, students' motivation, course topic, climate between instructor and class, course- and workload, and timing of the course substantially influence absenteeism. Hence, data are certainly missing not at random; while on the level of individual courses strong variations occur when we adjust for selection bias, average changes of student ratings are small.

In contrast to that, we find substantial changes in rankings based on SET. We conclude from this finding, although SET are not totally free from bias, they are a quite reliable instrument to assess quality of teaching at the individual level. However, since extent and direction of selection bias varies with course format which induces substantial changes in course rankings, SET are ill-suited for the comparison of different courses. Consequently, the finding of biases induced by the self-selection of participants raises doubts whether evaluation results are appropriate to depict a precise order of courses or instructors as regards their teaching quality. The inexactness of course rankings strongly suggests to reconsider the current practice of evaluation-based tenure decisions, remuneration, and appropriation of funds. Increasing the stakes of SET might not only foster manipulations of the ratings by means of easy tests, lenient grading, and other presents to the students (e.g., finishing class earlier, reducing the amount of assignments, distributing candy for in-class tasks), but might even have the unintended consequence of demotivating instructors which feel treated unfairly. Thus, in contrast to the current practice, we recommend to mainly use SET results as a feedback instrument which can be put in context by the instructors.

To gain further insights on the self-selection of students additional research is required. First, one objective should be to develop statistical tools to correct for selection bias in SET. To achieve this purpose a deeper understanding of the determinants of class attendance is needed. To our knowledge (except for Wolbring 2012) this paper is the first which studies determinates of absenteeism from class and systematically relates them to the reliability and validity of SET. Second, the establishment of online-based teaching evaluation methods could present an opportunity to gain further insights about the reasons and consequences of student withdrawal, especially when e-mail lists of all initial participants are available. However, the availability of online-survey tools also raises new questions as especially regards new forms of non-response bias. Finally, information on students' study behavior and on the reasons for skipping class is also of interest for the development and enhancement of study programs and universities. High rates of absenteeism might often not reflect problems in an individual course, but rather deficits in the structure of a study program or the social and academic integration of students. Gathering information on the course level and systematically relating it to the structure of a study program might thus be helpful to further improve teaching quality and develop a coherent course of studies.

References

1. Abbott, A., & Tsay, A. (2000). Sequence Analysis and Optimal Matching Methods in Sociology. Sociological Methods & Research, 29(1), 3–33
2. Adams, M.J., & Umbach, P.D. (2012). Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. Researchh in Higher Education, 53(5), 576–591.
3. Arulampalam, W., Naylor, & Smith, J. (2012). Am I Missing Something? The Effects of Absence from Class on Student Performance. Economics of Education Review, 31(4), 363–375.
4. Babad, E. (2001). Student's Course Selection: Differential Considerations for First and Last Course. Research in Higher Education, 42(4), 469–492.
5. Babad, E., Darley, J., & Kaplowitz, H. (1999). Developmental Aspects in Students Course Selection. Journal of Educational Psychology, 91, 157–168.
6. Babad, E., Icekson, T., & Yelinek, Y. (2008). Antecedents and Correlates of Course Cancellation in a University Drop and Add Period. Research in Higher Education, 49(4), 293–316.
7. Babad, E., & Tayeb, A. (2003). Experimental Analysis of Students Course Selection. British Journal of Educational Psychology, 73(3), 373–393.
8. Becker, W.E., & Powers, J.R. (2001). Student Performance, Attrition, and Class Size Given Missing Student Data. Economics of Education Review, 20(4), 377–388.
9. Becker, W. E., & Walstad, W. B. (1990). Data Loss From Pretest to Posttest as a Sample Selection Problem. The Review of Economics and Statistics, 72(1), 184–188.
10. Berger, U., & Schleußner, C. (2003). Are Ratings of Lectures Confounded with Students Frequency of Participation? German Journal of Educational Psychology, 17(2), 125–131.
11. Bratti, M., & Staffolani, S. (2013). Student Time Allocation and Educational Production Functions. Annals of Economics and Statistics, 111/112, 103–140.
12. Coleman, J., & McKeachie, W. (1981). Effects of Instructor/Course Evaluations on Student Course Selection. Journal of Educational Psychology, 73, 224–226.
13. Devadoss, S., & Foltz, J. (1996). Evaluation of Factors Influencing Student Class Attendance and Performance. American Journal of Agricultural Economics, 78(3), 499–507.
14. Dolton, P., Marcenaro, O. D., & Navarro, L. (2003). The Effective Use of Student Time: a Stochastic Frontier Production Function Case Study. Economics of Education Review, 22(6), 547–560.
15. Douglas, S., & Sulock, J. (1995). Estimating Educational Production Functions with Correction for Drops. Journal of Economic Education, 26(2), 101–112
16. Elwert, F., & Winship, C. (2014). Endogenous Selection Bias. Annual Review of Sociology, 40.
17. Enders, C.K. (2010). Applied Missing Data Analysis. New York/London: Guilford Press.
18. Greimel-Fuhrmann, B., & Geyer, A. (2003). Students Evaluations of Teachers and Instructional Quality – Analysis of Relevant Factors Based on Empirical Evaluation Research. Assessment & Evaluation in Higher Education, 28(3), 229–238.
19. Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47(1), 153–161.
20. Isserstedt, W., Middendorff, E., Kandulla, M., Borchert, L., & Leszczensky, M. (2010). Die wirtschaftliche und soziale Lage der Studierenden in der Bundesrepublik Deutschland 2009. 19. Sozialerhebung des DSW durchgefhrt durch HIS Hochschul-Informations-System. Bonn/Berlin: BMBF.
21. Kearney, K. A., R. H. Hopkins, A. L. Mauss, & Weisheit, R.A. (1984). Self-generated Identification Codes for Anonymous Collection of Longitudinal Questionnaire Data. Public Opinion Quarterly, 48(1B), 370-378.
22. Kirby, A., & McElroy, B. (2003). The Effect of Attendance on Grade for First Year Economics Students in University College Cork. Economic and Social Review, 34(3), 311–326.
23. Leventhal, L., Abrami, P., & Perry, R. (1976). Do Teacher Rating Forms Reveal as much about Students as about Teachers? Journal of Educational Psychology, 441–445.
24. Leventhal, L., Abrami, P., Perry, R., & Breen, L. (1975). Section Selection in Multisection Courses: Implications for the Validation and Use of Teacher Rating Forms. Educational and Psychological Measurement, 35, 885–895.

25. Little, R.J.A., & Rubin, D.B. (2002). Statistical Analysis with Missing Data (2nd ed.). New York: Wiley.
26. Long, J.S., & Freese, J. (2006). Regression Models for Categorical Dependent Variables using Stata. College Station, Texas: Stata Press.
27. Marburger, D. R. (2001). Absenteeism and Undergraduate Exam Performance. Journal of Economic Education, 32(2), 99–108.
28. Marsh, H. (2007). Students Evaluations of University Teaching: A Multidimensional Perspective. In P.P. Raymond & J.C. Smart (Ed.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective (pp. 319–384). New York: Springer.
29. Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge: Cambridge University Press.
30. Reed, J. G. (1981). Dropping a College Course: Factors Influencing Students Withdrawal Decisions. Journal of Educational Psychology, 73(3), 376–385.
31. Romer, D. (1993). Do Students Go to Class? Should They? Journal of Economic Perspectives, 7(3), 167–174.
32. Royston, P. (2005). Multiple Imputation of Missing Values: Update. Stata Journal, 5(2), 188–201.
33. Schafer, J.L., & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. Psychological Methods, 7(2), 147–177.
34. Schmidt, R.M. (1983). Who Maximizes What? A Study in Student Time Allocation. American Economic Review: Papers and Proceedings, 73(2), 23–28.
35. Schnell, R., Bachteler, T., & Reiher, J. (2010). Improving the Use of Self-generated Identification Codes. Evaluation Review, 34(5), 391–418.
36. Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: the State of the Art. Review of Educational Research, 83(4), 598–642.
37. Stanca, L. (2003). The Effects of Attendance on Academic Performance: Panel Data Evidence for Introductory Microeconomics. Journal of Economic Education, 37(2), 251–266.
38. Taris, T. (2000). A Primer in Longitudinal Data Analysis. London: Sage.
39. Titus, M.A. (2007). Detecting Selection Bias, Using Propensity Score Matching, and Estimating Treatment Effects: An Application to the Private Returns to a Master's Degree. Research in Higher Education, 48(4), 487–521.
40. van Buuren, S. (2012). Flexible Imputation of Missing Data. Boca Raton: CRC Press.
41. Weiler, W.C., & Pierro, D.J. (1988). Selection Bias and the Analysis of Persistence of Part-time Undergraduate Students. Research in Higher Education, 29(3), 261–272.
42. Wilhelm, W.B. (2004). The Relative Influence of Published Teaching Evaluations and Other Instructor Attributes on Course Choice. Journal of Marketing Education, 26(1), 17–30.
43. Wolbring, T. (2012). Class Attendance and Students Evaluations of Teaching. Do No-Shows Bias Course Ratings and Rankings? Evaluation Review, 36(1), 72–96.
44. Wyatt, G. (1992). Skipping Class: An Analysis of Absenteeism among First-Year College Students. Teaching Sociology, 20(3), 201–207.
45. Yurek, L. A., Vasey, J., & Havens, D.S. (2008). The Use of Self-generated Identification Codes in Longitudinal Research. Evaluation Review, 32(5), 1–18.

Table A Descriptive Statistics for Students Present at t=1

Variable	Mean	SD	Min	Max	N
Missing at t=2	.533	.499	0	1	1335
Overall course rating at t=1	3.816	.706	1	5	1320
Prior interest in course topic	2.941	1.391	1	5	1267
Study-specific values	3.826	.880	1	5	1303
Log course preparation (in minutes)	1.947	1.850	0	6.4	1312
Exam relevance of attendance	3.416	1.382	1	5	1304
Bad exam performance due to absenteeism	1.761	1.217	1	5	1308
Course in quantitative methods	.401	.490	0	1	1335
Student-instructor-climate	4.027	.880	1	5	1317
Student-class-climate	3.984	.817	1	5	1318
First/last class of the day	.745	.436	0	1	1326
Break before course (in minutes)	40.110	66.700	0	560	1228
Courseload (in hours)	18.582	5.176	1	52	1288
Workload (in hours)	7.703	7.722	0	40	1303