

Evaluation Review

<http://erx.sagepub.com/>

Class Attendance and Students' Evaluations of Teaching: Do No-Shows Bias Course Ratings and Rankings?

Tobias Wolbring

Eval Rev published online 12 April 2012

DOI: 10.1177/0193841X12441355

The online version of this article can be found at:

<http://erx.sagepub.com/content/early/2012/04/12/0193841X12441355>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Evaluation Review* can be found at:

Email Alerts: <http://erx.sagepub.com/cgi/alerts>

Subscriptions: <http://erx.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

>> [OnlineFirst Version of Record](#) - Apr 12, 2012

[What is This?](#)

Class Attendance and Students' Evaluations of Teaching: Do No-Shows Bias Course Ratings and Rankings?

Evaluation Review

00(0) 1-25

© The Author(s) 2012

Reprints and permission:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0193841X12441355

<http://erx.sagepub.com>



Tobias Wolbring¹

Abstract

Background: Many university departments use students' evaluations of teaching (SET) to compare and rank courses. However, absenteeism from class is often nonrandom and, therefore, SET for different courses might not be comparable. *Objective:* The present study aims to answer two questions. Are SET positively biased due to absenteeism? Do procedures, which adjust for absenteeism, change course rankings? *Research Design:* The author discusses the problem from a missing data perspective and present empirical results from regression models to determine which factors are simultaneously associated with students' class attendance and course ratings. In order to determine the extent of these biases, the author then corrects average ratings for students' absenteeism and inspect changes

¹ Institute of Sociology, Ludwig-Maximilians-University Munich, Munich, Germany

Corresponding Author:

Tobias Wolbring, Institute of Sociology, Ludwig-Maximilians-University Munich, Konradstr, 6, Munich 80801, Germany

Email: tobias.wolbring@soziologie.uni-muenchen.de

in course rankings resulting from this adjustment. *Subjects:* The author analyzes SET data on the individual level. One or more course ratings are available for each student. *Measures:* Individual course ratings and absenteeism served as the key outcomes. *Results:* Absenteeism decreases with rising teaching quality. Furthermore, both factors are systematically related to student and course attributes. Weighting students' ratings by actual absenteeism leads to mostly small changes in ranks, which follow a power law. Only a few, average courses are disproportionally influenced by the adjustment. Weighting by predicted absenteeism leads to very small changes in ranks. Again, average courses are more strongly affected than courses of very high or low in quality. *Conclusions:* No-shows bias course ratings and rankings. SET are more appropriate to identify high- and low-quality courses than to determine the exact ranks of average courses.

Keywords

absenteeism, course rankings, missing data, sample selection bias, students' evaluations of teaching

Introduction

Students' evaluations of teaching (SET) are a widely used instrument to assure and foster teaching quality at universities around the globe. More specifically, many departments do not only report SET results to the faculty but also use average course ratings to compare and rank courses. Meanwhile, even decisions about merit pay, tenure, payment-related budgeting, and teaching awards (not only in the United States but also in Asia, Australia, and Europe) are based on these measures of teaching quality (Hénard 2010; Wilkesmann and Schmid 2011). Thus, SET are a central component of a New Public Management in higher education (de Boer, Endres, and Schimank 2007; Lane and Kivisto 2008; Sporn 2011), whereby the wide use of SET is implicitly based on at least three empirical assumptions:

Assumption 1: SET are valid measures (of students' perception)¹ of teaching quality.

Assumption 2: SET are fair measures (of students' perception) of teaching quality.

Assumption 3: The quality of different courses (as perceived by the students) can be compared based on SET.

In this article, we focus on the effects of students' attendance on SET and test all three assumptions:² Assumption 1 would be challenged, if SET were systematically biased by the absence of less interested and dissatisfied students—a so-called sample selection bias (Becker and Walstad 1990; Heckman 1979) or problem of nonignorable missing data (Little and Rubin 1987; Rubin 1976). The fundamental consequence of such a nonrandom data loss is that the measure does not solely measure what it is intended to measure, namely teaching quality.

Furthermore, if factors, which teachers can hardly control (e.g., students' employment, their leisure activities, number of courses taken, weekday and time of the course, weather), influence absenteeism and (mediated by this) SET, then the fairness of the procedure (Assumption 2) is also questionable. In this case, the mean rating of a course could be worse (than an SET of a course similar in quality), simply because the evaluation takes place during a sunny Friday morning, when especially interested or satisfied students are absent.³

Finally, if absenteeism rates strongly vary between courses and are correlated with teaching quality, measurement biases could not only lead to imprecise course *ratings* but also to incorrect course *rankings*. In that case, Assumption 3 would be wrong and, therefore, one should not compare SET of different courses—at least as long as SET are not adjusted for differential absenteeism and dropout.⁴

We further develop our argument in four steps: First, we discuss the problem from a missing data perspective and motivate our empirical analyses. Having explained our research strategy and the SET data used, we present empirical results from regression models to determine which factors are simultaneously associated with students' class attendance and SET. In order to determine the extent of these biases, we then correct average ratings for actual or predicted student absenteeism and inspect changes in course rankings resulting from this adjustment. We will close with some general conclusions from these results for SET and give specific recommendations for the use of course rankings based on SET.

Students' Absenteeism As a Case of Missing Data

The fact that some students are absent when the SET takes place can be regarded as a problem of missing data. One does not know how those students would have rated were they present. In their seminal book, Little and Rubin (1987; see also Rubin 1976) distinguish three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR),

and not missing at random (NMAR). NMAR refers to a situation, where the missing data mechanism is systematically related to variables of interest Y , even after controlling for other variables X . In contrast to that, data are MAR, if the probability of missing values is related to other variables X , but not directly to the outcome Y . In other words, Y is conditionally independent from the process of data loss. Finally, MCAR refers to the case of unconditional independence. In this case, missing values in the data set can be regarded as the result of a simple random process. Obviously, this is the most desirable type of missingness. However, in practice, the MCAR assumption almost never holds, whereas cases of MAR and NMAR are far more common. The same is true for students' ratings of teaching quality.

In the case of SET, the MAR assumption would hold, if after controlling for all factors, which simultaneously affect course attendance and SET ratings, both variables were unrelated. Clearly, the plausibility of the MAR assumption and the success of any adjustment critically depend on the covariate selection. Therefore, ideally one would take into account all relevant information about both those students, who completed the SET, and those who did not. However, we have no information on absent students in our data. Therefore, we have to rely on the bold assumption that we have identified all (or at least the most important) variables, which simultaneously influence attendance and SET ratings.

The tenability of this MAR assumption is at least worthy of discussion, since, even after partialling out the effects of all selected covariates, one might still suspect a nonignorable nonresponse leading to biased estimates. First, one can never be sure to have identified all relevant variables. Second, course attendance, performance, and SET ratings seem to be directly related (Babad, Ickson, and Yelinek 2008; Berger and Schleußner 2003). And third, students dropping class or withdrawing from studies are probably not comparable to those, who miss a significant part of the course, but finish it. Especially, it appears quite plausible that dropouts would evaluate the quality of a course worse than infrequent visitors do. We are not aware of any study which explores this specific hypothesis. However, research shows that dissatisfied and low performing students have a higher propensity to drop class (Bosshardt 2004; Dobkin, Gill, and Marion 2007; Reed 1981) and withdraw from studies (Thomas, Adams, and Birchenough 1996; Tinto 1993; Yorke 1999).

As becomes clear from this discussion, one cannot formally test whether the data are MAR. However, one can perform calculations and analyses that provide some insight. For example, one can check whether the data are MCAR by determining whether missingness is related to variables that

likely predict outcomes as well. Prior research suggests that these variables might include course and instructor characteristics such as course size, mandatory attendance rules, difficulty of course contents, weekday and time of the course, and physical attractiveness of the instructor (e.g., Arulampalam, Naylor, and Smith 2008; Becker and Powers 2001; Bosshardt 2004; Devadoss and Foltz 1996; Hamermesh and Parker 2005; Wolbring 2010). Moreover, students' individual dispositions, prior interest in the course content, general study motivation, grades, courseload, and workload appear to be important predictors of attendance and course ratings (Becker and Powers 2001; Berger and Schleußner 2003; Devadoss and Foltz 1996; Esser 1997; Kirby and McElroy 2003; Reed 1981; Romer 1993; Wolbring 2010). However, all studies cited only report determinants of either class attendance or SET ratings, but not of both of them.

Therefore, in order to explore whether the attributes mentioned above are simultaneously related to attendance and course ratings, we estimated regression models with these variables as regressors and with class attendance and SET ratings as dependent variables. If the data are MAR, then the results should be corrected for bias. Since our results suggest such undesirable influences, we adjust ratings by weighting the data by actual and predicted class attendance.⁵ However, before presenting our empirical findings, we describe data and statistical methods used.

Data and Methods

In this section, we give a description of the SET data used, motivate the choice of regression models employed to determine influential factors on class attendance and SET rating, and explain the weighting approach to adjust for bias.

Data and Operationalizations

The following analyses are based on SET at the Faculty of Social Sciences of the University of Munich. In total, approximately 18,000 observations for over 680 courses with at least 10 participants are available for the years 2008–2010 (covering four terms). Besides specific items about a variety of dimensions of teaching quality and instructor behavior, the Munich SET questionnaire contains a global rating question about overall course quality (“All in all, how do you rate the overall quality of this course?” *very good* [1.0] – *insufficient* [5.0]). Moreover, for each course, the

number of classes missed is also assessed in the SET form (“Up to now, how many classes of this course did you miss?”).

Furthermore, the data set contains the following variables (for an overview of their distribution see Table A1 in Appendix A):

- *PERFORMANCE RECORD*: “Do you need a performance record?” *yes* (1); *no* (0).
- *COURSE SIZE*: Number of students who completed a SET form in the course.
- *DEPARTMENT*: Communication sciences, political sciences, or sociology.
- *COURSE DAY and TIME*: Weekday and starting time of the course.
- *SUMMER TERM*: Equals 1 if the course is held during a summer term (on average about 12 classes). Otherwise winter term (on average about 14 classes).
- *COURSE PACE*: “I could not follow the pace of the course.” *totally agree* (1)–*totally disagree* (5).
- *COURSE DIFFICULTY*: “The course was too difficult” *totally agree* (1)–*totally disagree* (5).
- *PRIOR INTEREST*: “I chose the course, because I was interested in its content” *totally agree* (1)–*totally disagree* (5).
- *INSTRUCTOR KNOWN*: “I chose the course, because I already knew the instructor” *totally agree* (1)–*totally disagree* (5).
- *INSTRUCTOR’S PHYSICAL ATTRACTIVENESS*: Only available for a subset of lecturers. 11 male and 9 female students from the University of Bern and 31 male and 34 female students from the University of Bern, who didn’t know the instructors, rated instructors’ portrait photos (taken from the internet) on an 11-point-attractiveness-scale (0 [*very unattractive*]–10 [*very attractive*]). For analyses we use each instructor’s average rating as a measure for her physical attractiveness.⁶
- *PREPARATION FOR THE COURSE*: “On average, how many minutes per week did you prepare for the course?”
- *COURSELOAD*: “How many hours per week do you take for credit in this semester?”
- *WORKLOAD*: “How many hours per week do you work for payment in this semester?”
- *SEMESTER*: Student’s subject-related semester of study.

Based on prior findings on determinants of SET ratings and attendance, we included these variables as regressors in our models.

Regression Models

Perceived course quality and absenteeism are the dependent variables in our regression models. In a first step, we estimated simple linear regressions.⁷ To take into account the fact that students usually complete more than one SET questionnaire, we clustered standard errors around students.⁸ We can do this, since—compared to regular SET data for other universities and besides the rich information on student and course attributes—our data set has another major advantage: the Munich SET questionnaire contains a self-generating panel identifier (e.g., Kearney et al. 1984; Yurek, Vasey, and Havens 2008).⁹

Moreover, the availability of a panel identifier allows us to control for unobserved interindividual heterogeneity, which often biases results from cross-sectional analyses. In the case of SET, individual differences in general abilities, socioeconomic background, and intrinsic motivation could influence class attendance and ratings. For that reason and in addition to simple linear regression models, we estimated models with individual fixed effects for each subject (Allison 2009; Angrist and Pischke 2009; Wooldridge 2002) to test for the robustness of our results.¹⁰

Having identified factors, which simultaneously influence both class attendance and SET ratings, by this means, we then correct for bias due to influential covariates. We do this by weighting observations by actual and predicted absenteeism in the following way.

Weighting Approach

Assume that the number of previous classes of a course a student, who completed the SET questionnaire for this course, has missed reveals information about other students, who are absent during the SET of this course. If this assumption holds, based on the number of classes c_j of a course j previous to the SET and the number of classes m_{ij} a student i missed, we can calculate individual probabilities p_{ij} to be present during the SET of this course:

$$p_{ij} = \frac{c_j - m_{ij}}{c_j}.$$

Thereby, a natural starting point is to use the actual number of classes missed (as self-disclosed in the questionnaire) as a proxy for m_{ij} . However, one might be worried that this measure has a significant random component (e.g., illness, problems with public transportation) and/or is biased (e.g., due

to social desirability and fear of negative consequences). Therefore, besides a weighting approach using the self-assessed frequency of absenteeism, we also weight by the systematic part of absenteeism, namely, the predicted values we derived from the simple linear regression model in Table 1. Furthermore, we always set $c_j = 11$ for the winter term and $c_j = 9$ for the summer term, because the summer term is usually 2 weeks shorter than the winter term.

We then weight individual SET quality ratings by p_{ij} . More specifically, we use the inverse probability of being present $1/p_{ij}$ to weight up students, who miss class more often. This form of weighting reflects the fact that students with a lower p_{ij} are underrepresented in our SET data. Thus, in order to receive a less biased measure of teaching quality more weight has to be given to absent students.¹¹

As common in many departments, the next step is to rank courses (with at least 10 observations) separately for each term using average course ratings (unadjusted and weighted by $1/p_{ij}$). Finally, we compute absolute mean deviations between the ranked orderings, which are based on these two measures.

Results

Having explained our empirical strategy, we now first present findings on determinants of students' absenteeism and course ratings. By determining whether class attendance is related to variables that likely predict course ratings as well one can check whether the data are MCAR. Since our findings suggest that our SET data are probably MAR, but definitely not MCAR, we then adjust course rankings for absenteeism and explore changes in ranks resulting from this bias correction.

Determinants of Students' Absenteeism

First, we estimated simple linear regression models for students' absenteeism. As can be seen in Table 1, most effects are in accordance with theoretical expectations: Most importantly, course quality is significantly correlated with absenteeism from class. If a student rates a course with the German grade 1.0 (+; *very good*) instead of the German grade 5.0 (–; *insufficient*), he misses .4 less classes, on average. Given that the average student misses 1.1 classes, this effect is surely not irrelevant in practice. If we do not control for other covariates, which also influence the SET ratings, we get a highly significant correlation between quality and attendance of .7. Using

Table 1. Regression Results (Dependent Variable: Frequency of Absenteeism)

Variable	Simple linear regression		Fixed-effects regression 1		Fixed-effects regression 2	
	Coef.	SE	Coef.	SE	Coef.	SE
Performance record (1 = required)	-0.204***	(0.041)	-0.342***	(0.056)	-0.379***	(0.069)
Course size	0.002***	(0.0001)	0.002***	(0.0002)	0.002***	(0.0002)
Course day (ref = Monday)						
Tuesday	-0.036	(0.027)	-0.122***	(0.029)	-0.165***	(0.049)
Wednesday	0.059*	(0.029)	0.038	(0.032)	-0.02	(0.047)
Thursday	-0.001	(0.030)	-0.033	(0.033)	-0.037	(0.051)
Friday	0.127*	(0.057)	0.138*	(0.063)	0.260*	(0.102)
Course time (ref = 8/9 a.m.)						
10 a.m.	-0.157***	(0.037)	-0.181***	(0.037)	-0.174***	(0.050)
12/1 p.m.	-0.135***	(0.040)	-0.188***	(0.041)	-0.114 ⁺	(0.059)
2/3 p.m.	-0.246***	(0.040)	-0.303***	(0.043)	-0.310***	(0.062)
4 p.m.	-0.124**	(0.042)	-0.168***	(0.043)	-0.203**	(0.066)
6/7/8 p.m.	-0.198***	(0.051)	-0.259***	(0.054)	-0.407***	(0.079)
Course pace	-0.019	(0.012)	-0.020	(0.014)	-0.012	(0.020)
Course difficulty	-0.010	(0.014)	-0.033*	(0.016)	-0.008	(0.023)
Prior interest	0.057***	(0.010)	0.057***	(0.011)	0.085***	(0.016)
Instructor known	0.013	(0.008)	0.006	(0.011)	0.033*	(0.015)
Preparation for the course	-0.002***	(0.0002)	-0.002***	(0.0002)	-0.003***	(0.0004)
Courseload	-0.006*	(0.002)	-0.007	(0.004)	-0.008	(0.005)
Workload	0.015***	(0.001)	0.006 ⁺	(0.003)	0.005	(0.004)
Semesters	0.050***	(0.005)	0.155***	(0.016)	0.181***	(0.019)
Course quality	0.104***	(0.019)	0.123***	(0.021)	0.132***	(0.028)
Instructors' attractiveness					-0.040*	(0.017)
Constant	0.869***	(0.106)	0.943***	(0.155)	1.008***	(0.215)
R ²	.096		.107		.116	
Adjusted R ²	.094		.105		.114	
Observations	14,104		14,028		9,967	
Individuals	—		6,724		5,331	

Note. Regression models with dependent variable "frequency of absenteeism from classes". Nonstandardized coefficients (for negative binomial model: incidence-rate ratios), standard errors in parentheses.

⁺ $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$. Controlled for department (communication sciences, political sciences, and sociology), summer term. Negative binomial and simple linear regression model with robust standard errors clustered around students. Fixed-effects models with robust standard errors.

similar global rating questions (such as “I learned more than in other classes” or “I could explain the central concepts of the course to others”) instead we even find slightly stronger effect.

Moreover, needing a performance record significantly reduces absenteeism. The effect of social control is similar in size: if course size rises, it becomes more difficult for instructors to keep track of individual absenteeism. In result, an increase of course size by 100 students is associated with an increase in average absenteeism by 0.2. Further analyses (not reported) show that course size is only a predictor of absenteeism if students need a performance record—a finding which further strengthens the interpretation of the negative class size effect as the result of decreasing social control.

Furthermore, weekday and starting time of the course significantly affect class attendance. Classes on Monday, Wednesday, and Friday, as well as early in the morning (8/9 a.m.) are attended less regularly. However, in contrast to our theoretical expectations and previous findings, course pace and course difficulty have no significant effect.¹² However, if we test for their effects on absenteeism without including further controls in the model, both are significantly related to course attendance. According to these results, absenteeism is higher, if the course pace is too fast and courses are too difficult.

As well in accordance with theoretical expectations and previous findings, students, who were more interested in the content of the course, who knew the instructor when choosing the course, who better prepared for the course, and who spent less hours working for payment, attended class more frequently. Furthermore, dropout increases with higher semester and lower number of courses attended. However, this last result could only reflect the fact that students self-select into courses and that more interested and motivated students elect more courses.

For that reason, we next estimated a model with individual fixed effects to control for unobserved heterogeneity and test for the robustness of our results.¹³ All our main results remain remarkably stable; most correlations between independent variables (e.g., course quality, performance record, and course time) and students' absenteeism even become closer in the fixed-effects context.

In a last step, we analyzed a subsample of our data, for which measures of instructors' physical attractiveness are available. Interestingly, students obviously value physical attractiveness, because they more regularly attend classes of more physically attractive instructors. Thus, to complement recent studies, which show that more attractive instructors are significantly rated better in SET (Hamermesh and Parker 2005; Wolbring 2010), students

seem to have a preference for more physically attractive instructors and derive utility from attending their classes and interacting with them.

To sum up our results, course quality is clearly associated with absenteeism. Moreover, we found that various individual and structural factors, which are not under the control of the instructor, affect class attendance. In a next step, we ask whether absenteeism is simultaneously determined with course ratings. If this is the case, the Munich SET data are not MCAR and should be corrected for bias.

Determinants of Course Ratings

Simple linear regression models and fixed-effects models in Table 2 actually show that some of the structural and individual factors, which influence attendance, are also significantly related to students' ratings of course quality: Needing a performance record for the course not only increases the frequency of attendance but is also associated with slightly worse ratings. However, the effect is only significant in the fixed-effects model and rather small. In contrast to that, course size and student's semester of study go hand in hand with more absenteeism and better ratings. Finally, some factors have both positive effects on class attendance and course ratings. This is true for students' prior interest in course contents, the amount of time they spend to prepare for class, and physical attractiveness of the instructor. Finally, looking at standardized coefficients (not reported), one can see that course size, semester of study, and weekly preparation for the course are most closely correlated with SET.

Thus, our analyses show that the SET data are not MCAR. Because of this, one has to question the validity and fairness of SET as a measure of teaching quality (Assumptions 1 and 2). As well, it suggests itself to use information about students' class attendance to at least correct SET ratings for known influences. If data were MAR, this correction would even remove all bias induced by missing students. However, before drawing such a practical conclusion for SET, we have to prove first that the bias induced by absenteeism on SET is reasonably strong, that is, that a correction would really make a difference for the ranking of courses. Furthermore, it is not clear how those different biases add up, since some variables have a positive (negative) influence on attendance and a negative (positive) influence on the ratings, whereas others have a positive effect on both outcomes.

Table 2. Regression Results (Dependent Variable: Course Quality)

Variable	Simple linear regression		Fixed-effects regressions 1		Fixed-effects regressions 2	
	Coef.	SE	Coef.	SE	Coef.	SE
Performance record (1 = required)	0.026	(0.016)	0.078**	(0.027)	0.080*	(0.036)
Course size	-0.001***	(0.0001)	-0.001***	(0.0001)	-0.001***	(0.0001)
Course day (ref = Monday)						
Tuesday	-0.074***	(0.015)	-0.103***	(0.019)	-0.086***	(0.026)
Wednesday	0.015	(0.015)	-0.001	(0.018)	-0.04	(0.025)
Thursday	-0.027 ⁺	(0.015)	-0.045*	(0.019)	-0.083**	(0.026)
Friday	-0.060*	(0.030)	-0.097*	(0.039)	-0.107*	(0.053)
Course time (ref = 8/9 a.m.)						
10 a.m.	0.019	(0.018)	-0.029	(0.022)	-0.064*	(0.026)
12/1 p.m.	0.039*	(0.020)	0.008	(0.025)	0.019	(0.031)
2/3 p.m.	0.099***	(0.020)	0.070**	(0.025)	0.001	(0.032)
4 p.m.	0.106***	(0.022)	0.102***	(0.027)	0.097**	(0.034)
6/7/8 p.m.	0.202***	(0.030)	0.244***	(0.038)	0.228***	(0.041)
Course pace	-0.066***	(0.006)	-0.070***	(0.009)	-0.076***	(0.010)
Course difficulty	-0.086***	(0.008)	-0.091***	(0.011)	-0.082***	(0.012)
Prior interest	0.128***	(0.005)	0.145***	(0.007)	0.148***	(0.008)
Instructor known	0.137***	(0.004)	0.165***	(0.007)	0.170***	(0.007)
Preparation for the course	-0.001***	(0.0001)	-0.001***	(0.0001)	-0.001***	(0.0002)
Courseload	0.002	(0.001)	0.002	(0.002)	0.002	(0.003)
Workload	-0.0002	(0.001)	-0.0004	(0.002)	0	(0.002)
Semester	-0.0003	(0.002)	-0.051***	(0.008)	-0.054***	(0.010)
Instructors' attractiveness					-0.032***	(0.009)
Constant	1.893***	(0.051)	2.016***	(0.083)	2.200***	(0.107)
R ²	.262		.283		.304	
Adjusted R ²	.261		.282		.302	
Observations	14,159		14,159		10,013	
Individuals	—		6,771		5,351	

Note: Regression models with dependent variable "course quality" ranging from 1.0 (very good) to 5.0 (insufficient). Nonstandardized coefficients, standard errors in parentheses.

⁺p < .1. *p < .05. **p < .01. ***p < .001. Controlled for department (communication sciences, political sciences, sociology), summer term.

Adjusting Course Rankings for Absenteeism

As described in the data and methods section, we weight observations by actual and predicted absenteeism, rank all courses with at least 10 participants on the basis of unadjusted and adjusted mean course ratings, and compare resulting rankings.

Mean and median differences in ranks are presented in Table 3. The absolute average deviation of the adjusted ranking (based on actual and not predicted absenteeism) from a ranking generated with the commonly used procedure is rather small and ranges from 4.0 to 7.1 ranks. This is equal to an average change in ranks between 3.7% and 6.5% relative to the maximum of possible rank changes. The position of one quarter of all courses only changes by a maximum of 1–3 ranks, the position of 75% of all courses changes by a maximum of 5–8 ranks. Furthermore, the median is always below the mean, which (in addition to the high standard deviation) indicates that the distribution in ranks is positively skewed. In other words, there are a few courses, which are assigned to significantly different ranks with the two procedures: 25% (5%) of all courses are faced with differences of at least 5–8 (13–22) ranks. For some courses, even more extreme variation can be observed (up to 22–59 differences in ranks).

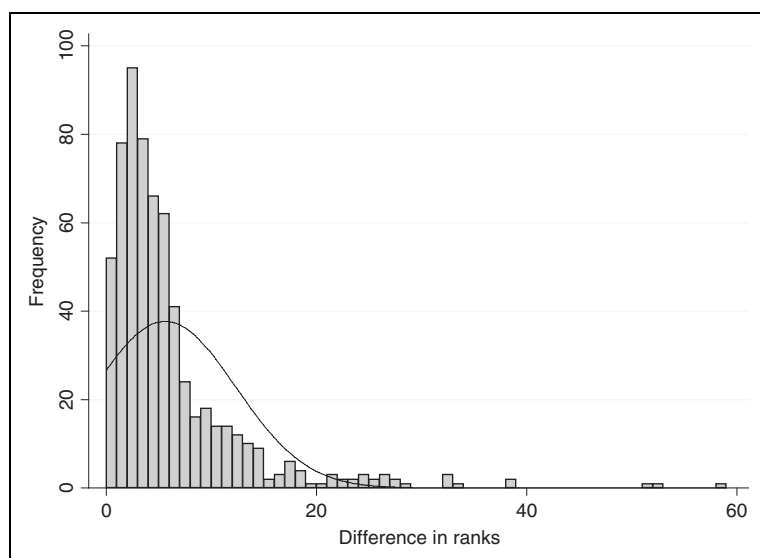
Figure 1 further illustrates this aspect. The distribution of differences in ranks obviously deviates from a normal distribution, since it has a longer tail and is positively skewed. Additional analyses show that the distribution follows a power law function (e.g., Mitzenmacher 2003; Newman 2005). An equation of the form $\log Y = \hat{C} - \hat{\beta} \log X$ with $\hat{C} = 1.94$ and $\hat{\beta} = 1.80$ adequately describes the relationship between differences in ranks X and the relative frequency of differences in ranks Y ($R^2 = .91$).¹⁴ This implies an extremely unequal distribution of rank differences and, therefore, that the rank of a small number of courses is disproportionately influenced by the used ranking procedure.

One might ask what course type is disproportionately affected by a change in the ranking mechanism. Our answer is courses with average teaching quality. We find that the positions of courses with a mean grade between 1.5 and 2.0 (2.0 and 2.5) on average change by 7.2 (4.6) ranks, whereas the positions of courses at the upper (lower) tail of the distribution only vary by 2.3 (1.8) ranks. Because ratings at the center of the distribution are closer to each other than more extreme ratings, weighting has stronger effects on the rank of average courses. Thus, the usual ranking procedure is more reliable for the identification of very good and very poor courses than for average courses.

Table 3. Absolute Differences in Ranks Using Unadjusted and Adjusted Mean Ratings

Term	Absenteeism	N	M	SD	min	5%	25%	Median	75%	95%	max
Winter 08/09	Actual	128	4.03	3.99	0	0	1	3	5	13	22
	Predicted	123	.44	.76	0	0	0	0	1	2	4
Summer 09	Actual	146	7.14	6.69	0	1	3	5	8	22	51
	Predicted	139	.56	.79	0	0	0	0	1	2	4
Winter 09/10	Actual	179	4.98	6.00	0	0	2	3	6	20	38
	Predicted	166	.78	.90	0	0	0	1	1	2	4
Summer 10	Actual	181	6.19	8.43	0	0	1	3	8	18	59
	Predicted	172	.68	.94	0	0	0	0	1	2	7
Total (per term)	Actual	634	5.63	6.71	0	0	2	4	7	18	59
	Predicted	600	.64	.87	0	0	0	0	1	2	7

Note. Differences in ranks between league tables based on unadjusted mean ratings and with mean ratings adjusted for actual and predicted absenteeism. Predicted absenteeism based on simple linear regression in Table 1. Columns min to max refer to relative frequency of changes in ranks. For example, looking at column "75%" and the row "total" one can see that for 75% of all courses the position changed by a maximum of 7 ranks if we weigh by actual absenteeism.

**Figure 1.** Distribution of differences in ranks.

In general, we find similar patterns if we weight by predicted frequencies of attendance: average courses are affected more strongly than courses with a mean rating at the tails of the distribution. However, with a maximum difference of 7 ranks and an average difference of 0.4–0.8 ranks, the size of the changes becomes extremely small (again see Table 3).

The main reason for this is the much smaller variation of predicted values of absenteeism compared to the actual frequency of absenteeism. Technically speaking, the second weighting approach only used the—according to our models—systematic component of absenteeism. Since the explanatory value of those models is rather low ($R^2 \approx .1$), much unexplained variance remains in the data. This residual variation is presumably at least partly systematic. So, we propose that the different weighting approaches are based on the assumption of two rather extreme constellations, namely that absenteeism is either completely or only to a very low degree systematic. We think the truth lies in between both extremes, since attendance behavior has a significant random component but is presumably also more systematic than our explanatory model suggests. Therefore, we argue that both weighting schemes are informative as measures of uncertainty in making comparisons based on incomplete data.

Conclusions

In this article, we investigated whether SET are biased by students being absent and whether procedures which adjust for varying frequency of attendance between courses produce different rankings than the widely used procedure which is based on students' raw mean ratings.

To begin with, we discussed the problem from a missing data perspective. Based on SET data for the Faculty of Social Sciences at the University of Munich, we then tested whether class attendance and quality ratings are determined by the same factors. This is the case for the need of a performance record, course size, course time, prior interest, preparation for the course, semester of study, and physical attractiveness of the instructor. Furthermore, attendance is also directly related to the quality of teaching.

Thus, it suggests itself that data are not MCAR and SET are biased. Therefore, Assumption 1 (SET are valid measures of teaching quality) does not hold. As well, Assumption 2 (SET are fair measures of teaching quality) is questionable, since some of these factors are not fully under the instructor's control. So, for example, more interested students rate courses better and also attend them more frequently. Similarly, courses with more participants are rated worse and are attended less regularly.

However, as long as course rankings do not change significantly if we adjust for absenteeism, validity, and fairness issues concerning students' ratings are not problematic from a pragmatic point of view. Therefore, we then switched from the micro to the macro level and asked whether course rankings based on a global measure of teaching quality change if we adjust for actual and predicted absenteeism (instead of using the raw mean). In other words, we tested Assumption 3, which posits that the quality of different courses can be compared based on SET. Weighting by actual frequency of absenteeism, we found that, on average, changes in ranks are low but follow a power law. The rank of a small number of mostly average courses is disproportionally influenced by the choice of the ranking procedure. This finding leads us to the conclusion that SET are appropriate to identify high- and low-quality courses, but not to determine the exact ranks of average courses. Using only the systematic part of absenteeism, which is associated with other covariates, we find similar patterns. However, resulting changes in ranks are so small, that weighting does not seem to be necessary at all. We posit that the truth lies in between those two rather extreme estimates.

However, we acknowledge that our data have weaknesses and, therefore, are not optimal for our endeavor. First, we had to rely on self-assessed information on students' absenteeism. This gives a lower bound of true absenteeism, since (although anonymity is fully guaranteed) some students fear negative consequences if admitting higher frequency of absenteeism. Furthermore, it is not always the case that past attendance predicts current attendance in the expected way. For example, if students need a performance record and course attendance is compulsory, past absenteeism can even positively influence the probability of current attendance. However, based on the assumption that class attendance is mainly the result of a cost-benefit analysis, we would expect that past attendance reveals information on the utility, which students derive from the course and other alternatives. Therefore, we think that in this case, past behavior is predictive for current attendance.

Second, in our empirical models, we could not control for all factors, which might be relevant for students' attendance and ratings. For example, we could not map peer group effects, although they are relevant for both class attendance decisions and students' well-being in class. Given a residual variance of about 90%, it seems to be natural to think about further omitted variables.

Finally, and most importantly, our weighting approach is worth discussing. On one hand, one could argue that students who miss more classes

should be given less and not more weight, since they are less equipped to evaluate the quality of teaching. The decision depends on whether the effect of perceived quality on attendance or of attendance on perceived quality is stronger. Although we have no proof for this, we assumed the former being stronger. On the other hand, even if one accepts giving absent students more weight the approach per se is probably problematic, because we had to draw inferences on students absent during the SET. More specifically, since no information on those students was available, we assumed that ratings by students present during the SET can be informative about the rating behavior of absent students. In other words, we assumed that students are MAR. However, one could argue that nonresponse is nonignorable, since students, who drop class or withdraw from studies, are also among those persons, who did not participate at the SET. Obviously, the decision of dropping out of university may not at all be driven by concerns about teaching quality of a specific course. However, if this is the case, these no-shows are unproblematic for ratings and ranking. In contrast to that, cancelling a course should be influenced more strongly by quality concerns. This nonignorable nonresponse is more problematic for ratings and rankings. However, we expect that changes in ranks would become more extreme if we controlled for this type of missingness. Thus, it suggests itself that our results give a lower bound of actual bias if part of the data is NMAR.

A more direct approach would have been to evaluate early and late in the course. As Kohlan (1973) showed SET “stabilize very early in the course”. Similarly, Costin (1968) reports medium to strong correlations between midsemester and end of course ratings. Thus, on one hand, given a design with two measurements in time one can compare the ratings of students, who completed both questionnaires, with those who only completed the first questionnaire. Based on such data, more reliable results on the exact size of the bias induced by absenteeism can be produced, since these estimates would rest on less bold assumptions than our analysis does. On the other hand, one can also utilize this additional information on absent students for the first measurement in time and can try to answer the counterfactual question how the evaluation of those students who are absent during the second SET would have looked (if they were present). Thus, it goes without saying that we plan further research with such a design with two measurements in time.

Appendix A

Table A1. Descriptive Statistics

Variable	M	SD	Min	Max	N
Performance record (1 = required)	0.91	0.29	0	1	18,756
Course size	72.22	95.20	1	426	19,186
Department					
Communication sciences	0.31	0.46	0	1	19,111
Political sciences	0.32	0.47	0	1	19,111
Sociology	0.36	0.48	0	1	19,111
Course day					
Monday	0.25	0.43	0	1	17,863
Tuesday	0.26	0.44	0	1	17,863
Wednesday	0.25	0.43	0	1	17,863
Thursday	0.21	0.41	0	1	17,863
Friday	0.03	0.17	0	1	17,863
Course time					
8/9 a.m.	0.11	0.31	0	1	17,891
10 a.m.	0.35	0.48	0	1	17,891
12/1 p.m.	0.18	0.38	0	1	17,891
2/3 p.m.	0.16	0.37	0	1	17,891
4 p.m.	0.14	0.35	0	1	17,891
6/7/8 p.m.	0.07	0.25	0	1	17,891
Summer term	0.48	0.50	0	1	19,137
Course speed	3.91	1.05	1	5	19,021
Course difficulty	4.13	0.93	1	5	19,009
Prior interest	2.32	1.24	1	5	17,783
Instructor known	3.21	1.41	1	5	17,303
Instructors' attractiveness	4.52	1.05	2	7.45	13,374
Preparation for the course	62.43	66.78	0	998	17,977
Courseload	16.58	5.79	0	55	17,886
Workload	8.46	8.45	0	60	18,005
Semester	3.77	2.60	1	18	17,919
Course quality	1.98	0.69	1	5	18,095
Absenteeism	1.14	1.22	0	10	18,215

Acknowledgments

The author is grateful to Marc Keuschnigg for sharing his vast knowledge about power laws with us and also for providing STATA do-files to determine the exact functional form in our data. The author also appreciates the help he received from Christiane Bozoyan, Roger Berger, and the participants of their courses on “deviant behavior” at the University of Bern and on “methods of social research” in Leipzig,

who rated the physical attractiveness of the instructors for us. Furthermore, the author would like to thank Patrick Riordan and Martina Kroher for their language editing support and for useful comments, Edgar Treischl for excellent research assistance, and Professor Helmut Küchenhoff for a helpful discussion on the usefulness of different weighting approaches. Last but not least, the author also received helpful suggestions from two anonymous reviewers and important support from Professor Michael Foster, who encouraged the author to frame students' absenteeism as a case of missing data and to organize the article in a straightforward way.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Considering SET as measures of teaching quality in students' perception appears to be more appropriate, since this definition highlights that objective teaching quality and its subjective perception need not be identical. Furthermore, this formulation emphasizes the interplay between students, instructor, and environmental conditions in the learning process (Biggs 1985; Ramsden 1979, 1991; Rindermann and Schofield 2001).
2. Thereby, the range of these assumptions is, of course, much wider than the test we suggest. This might be one reason why by now a vast amount of research on SET and, especially, on their validity exists (e.g., Kulik 2001; Marsh 1987; Merritt 2008). We cannot deal with all the details of this research branch. Nonetheless, if one is able to show, that a single factor (such as students' class attendance) biases SET, this is already sufficient to call the common practice of ranking into question.
3. This would also imply problems with the reliability of the measurement, because in this case SET then differ markedly at two different points in time.
4. As one of the reviewers pointed out, there are further arguments against respective comparisons (see also Footnote 2). For example, it is a matter of fact that each of the student ratings refers to an individual context of teaching and learning and is performed by different groups of students.
5. We are aware of the fact that other ranking procedures than a simple weighting have been suggested in the literature. Goldstein and Spiegelhalter (1996) argue for random effects models to take "account of model-based uncertainty in

making comparisons” (for a discussion of this and alternative statistical methodologies, see Deely and Smith 1998; Draper and Gittoes 2004). In order to partial out biases, McPherson and Jewell (2007) propose to adjust SET for factors that are beyond the control of the instructor (e.g., instructor race) or can be gamed by them (e.g., grades).

6. Rating portrait photos by a group of raters is a common way to measure physical attractiveness and is called the “truth of consensus method” (e.g., Henss 1992; Patzer 1985, 2007). Further analyses show that raters highly agree in their ratings (Cronbach’s $\alpha = .95$). We conclude from this fact that our measurements are reliable. However, one has to acknowledge that the attractiveness of a face is only one aspect of the broader theoretical construct “physical attractiveness.” This also includes further dimensions such as height, weight, body composition, and gesture. Because only photos of the face are available, we concentrate on this dimension of attractiveness, which seems to be the most important dimension of overall physical attractiveness (e.g., Hamermesh 2011).
7. Since the number of classes missed is a typical case of count data with overdispersion (Cameron and Trivedi 1998; Long 1997), we ran negative binomial and zero-inflated negative binomial models as sensitivity analyses. The results are very similar and predicted values from both count data models are highly correlated with those from the simple linear regression model ($r = .977$ and $r = .942$). Because of this and since the use of a nonlinear link function does raise additional estimation issues (involving distributional assumptions) that are generally underappreciated, we only report findings from ordinary regressions. Results from analyses using nonlinear regression models are available on request from the author.
8. Following the suggestion of one of the reviewers, we also estimated a mixed model with random effects on course and individual level to take into account the fact that students are also nested into classes. Our findings remain the same if we control for both forms of clustering, although standard errors are higher than in simple linear regression models. However, standard errors are even more conservative in the fixed-effects models we report.
9. By asking students in the SET for a series of time constant, easily retrievable characteristics (gender, year of birth, degree of studies, the first two digits of their mother’s birthday, the first two initials of their parents’ first name, and the number of older brothers and sisters), we can link SET to individual students and are able to observe their rating and attendance behavior for different courses over time.
10. By adding a subject-specific intercept α_i for each individual i , the fixed-effects approach allows to take into account the unobserved heterogeneity ascribable to the presence of repeated observations on the same student (repeated over time t

and repeated over more courses j). Formally, fixed-effects models can be written as $y_{ijt} = \beta X + \alpha_i + \varepsilon_{ijt}$, where β is a vector of parameters, X is a matrix of covariates and ε_{ijt} are independent and normally distributed errors. Since fixed-effects models relate changes in X for individual i to changes in y for the very same individual, but not differences in X to differences in y between individuals, interindividual differences are averaged out and do not bias fixed-effects estimates. Thus, the main motivation for estimating fixed-effects models is usually not to determine the between-unit variability in the outcome exactly, but to work with the within-unit variability only to produce more reliable estimates.

11. On the other hand, one could argue that students who miss more classes are less equipped to evaluate the quality of teaching. Moreover, a lecturer could be not very good in the perception of those students since they missed previous classes and as a result lack basic knowledge to continue to attend the course. From this perspective students who miss more classes should be given less weight, not more. However, the decision to weigh absent students up or down depends on whether the effect of perceived quality on attendance or of attendance on perceived quality is stronger. We do not deny that the latter could be present, but we think the former dominates the latter and, thus, decided to give absent students more weight.
12. However, it is possible that students drop out of too difficult courses. Unfortunately, we cannot control for this process with our data.
13. It is worth mentioning that the average number of observations per individual is rather small. In the fixed-effects context, this could lead to inconsistent estimates (Cameron and Trivedi 2005). For that reason, we reestimated all fixed-effects models for a subsample of person, for which 10 or more observations were available. Our main results are qualitatively very similar to those in Tables 1 and 2.
14. To determine the exact functional form in our data, we estimated simple linear regression models. We used a procedure suggested by Clauset, Shalizi, and Newman (2009) to determine the cutoff point, which lies around a difference in ranks of 5. We also followed their example when testing whether a maximum likelihood approach yields better estimates (for a similar, but more comprehensive application of power law analyses in the field of books, see Keuschnigg 2012).

References

- Allison, P. D. 2009. *Fixed Effects Regression Models*. Thousand Oaks, CA: SAGE.
- Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics. An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

- Arulampalam, W., R. A. Naylor, and J. Smith. 2008. "Am I Missing Something? The Effects of Absence from Class on Student Performance." IZA Discussion Paper No. 3749. Accessed June 27, 2011, <https://www.econstor.eu/dspace/bitstream/10419/35597/1/581797353.pdf>
- Babad, E., T. Icekson, and Y. Yelinek. 2008. "Antecedents and Correlates of Course Cancellation in a University 'Drop and Add' Period." *Research in Higher Education* 49:293–319.
- Becker, W. E., and J. R. Powers. 2001. "Student Performance, Attrition, and Class Size Given Missing Student Data." *Economics of Education Review* 20:377–88.
- Becker, W. E., and W. B. Walstad. 1990. "Data loss from Pretest to Posttest as a Sample Selection Problem." *Review of Economics and Statistics* 72:184–88.
- Berger, U., and C. Schleußner. 2003. "Are Ratings of Lectures Confounded with Students' Frequency of Participation?" *German Journal of Educational Psychology* 17:125–31.
- Biggs, J. B. 1985. "The Role of Metalearning in Study Processes." *British Journal of Educational Psychology* 55:185–212.
- Bosshardt, W. 2004. "Student Drops and Failure in Principles Courses." *Journal of Economic Education* 35:111–28.
- Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.
- Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics. Methods and Applications*. Cambridge: Cambridge University Press.
- Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. "Power-law Distributions in Empirical Data." *SIAM Review* 51:661–703.
- Costin, F. 1968. "A Graduate Course in the Teaching of Psychology: Description and Evaluation." *Journal of Teacher Education* 19:425–32.
- de Boer, H., J. Endres, and U. Schimank. 2007. "On the Way towards New Public Management? The Governance of University Systems in England, the Netherlands, Austria, and Germany." In *New Forms of Governance in Research Organizations*, edited by D. Jansen, 137–54. Dordrecht, The Netherlands: Springer.
- Deely, J. J., and A. F. M. Smith. 1998. "Quantitative Refinements for Comparisons of Institutional Performance." *Journal of the Royal Statistical Society* 161:5–12.
- Devadoss, S., and J. Foltz. 1996. "Evaluation of Factors Influencing Student Class Attendance and Performance." *American Journal of Agricultural Economics* 78: 499–507.
- Dobkin, C., R. Gil, and J. Marion. 2007. "Causes and Consequences of Skipping Class in College." Working Paper, UC Santa Cruz. Accessed January 4, 2012, http://people.ucsc.edu/~cdobkin/Papers/Class_Attendance.pdf.

- Draper, D., and M. Gittoes. 2004. "Statistical Analysis of Performance Indicators in UK Higher Education." *Journal of the Royal Statistical Society* 167:449–74.
- Esser, H. 1997. "Zweifel an der Evaluation der Lehre." *Wirtschaftswissenschaftliches Studium: Zeitschrift für Ausbildung und Hochschulkontakt* 26:45–49.
- Goldstein, H., and D. J. Spiegelhalter. 1996. "League Tables and Their Limitations: Statistical Issues in Comparisons of institutional Performance." *Journal of the Royal Statistical Society* 159:385–443.
- Hamermesh, D. S. 2011. *Beauty Pays. Why Attractive People Are More Successful*. Princeton, NJ: Princeton University Press.
- Hamermesh, D. S., and A. M. Parker. 2005. "Beauty in the Classroom. Instructors' Pulchritude and Putative Pedagogical Productivity." *Economics of Education Review* 24:369–76.
- Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–61.
- Hénard, F. 2010. *Learning Our Lesson. Review of Quality Teaching in Higher Education*. Paris, France: Organization for Economic Co-operation and Development.
- Henss, R. 1992. *Spieglein, Spieglein an der Wand... Geschlecht, Alter und physische Attraktivität*. Weinheim: Psychologie Verlags Union.
- Kearney, K. A., R. H. Hopkins, A. L. Mauss, and R. A. Weisheit. 1984. "Self-Generated Identification Codes for Anonymous Collection of Longitudinal Questionnaire Data." *Public Opinion Quarterly* 48:370–8.
- Keuschnigg, M. 2012. *Das Bestseller-Phänomen. Die Entstehung von Nachfragekonzentration im Buchmarkt*. Wiesbaden: VS Verlag.
- Kirby, A., and B. McElroy. 2003. "The Effect of Attendance on Grade for First Year Economics Students in University College Cork." *Economic and Social Review* 34:311–26.
- Kohlan, R. G. 1973. "A Comparison of Faculty Evaluations Early and Late in the Course." *Journal of Higher Education* 44:587–95.
- Kulik, J. A. 2001. "Student Ratings: Validity, Utility, and Controversy." *New Directions for Institutional Research* 109:9–25.
- Lane, J. E., and J. A. Kivisto. 2008. "Interests, Information, and Incentives in Higher Education: Principal-Agent Theory and Its Potential Applications to the Study of Higher Education Governance." In *Higher Education: Handbook of Theory and Research*. Vol. XXIII, edited by C. J. Smart, 141–79. New York: Springer.
- Little, R. J., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.
- Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: SAGE.

- Marsh, H. 1987. "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research." *International Journal of Educational Research* 11:253–388.
- McPherson, M. A., and R. T. Jewell. 2007. "Leveling the Playing Field: Should Student Evaluation Scores be Adjusted?" *Social Science Quarterly* 88:868–81.
- Merritt, D. J. 2008. "Bias, the Brain, and Student Evaluations of Teaching." *St. John's Law Review* 82:235–87.
- Mitzenmacher, M. 2003. "A Brief History of Generative Models for Power Law and Lognormal Distributions." *Internet Mathematics* 1:226–51.
- Newman, M. E. J. 2005. "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physic* 46:323–51.
- Patzer, G. L. 1985. *The Physical Attractiveness Phenomena*. New York: Plenum Press.
- Patzer, G. L. 2007. *Why Physically Attractive People are More Successful. The Scientific Explanation, Social Consequences, and Ethical Problems*. New York: Edwin Mellen Press.
- Ramsden, P. 1979. "Student Learning and Perceptions of the Academic Environment." *Higher Education* 8:411–27.
- Ramsden, P. 1991. "A Performance Indicator of Teaching Quality in Higher Education: The Course Experience Questionnaire." *Studies in Higher Education* 16:129–50.
- Reed, J. G. 1981. "Dropping a College Course: Factors Influencing Students' Withdrawal Decisions." *Journal of Educational Psychology* 73:376–85.
- Rindermann, H., and N. Schofield. 2001. "Generalizability of Multidimensional Student Ratings of University Instruction across Courses and Teachers." *Research in Higher Education* 42:377–99.
- Romer, D. 1993. "Do Students Go to Class? Should They?" *Journal of Economic Perspectives* 7:167–74.
- Rubin, D. 1976. "Inference with Missing Data." *Biometrika* 63:581–92.
- Sporn, B. 2011. "Governance and Administration: Organizational and Structural Trends." In *International Handbook of Higher Education*. Vol. 1, edited by J. F. Forest and P. G. Altbach, 141–57. Dordrecht, The Netherlands: Springer.
- Thomas, M., S. Adams, and A. Birchenough. 1996. "Student Withdrawal from Higher Education." *Educational Management Administration & Leadership* 24:207–21.
- Tinto, V. 1993. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. Chicago: University of Chicago Press.
- Wilkesmann, U., and C. J. Schmid. 2011. "The Impacts of New Governance on Teaching at German Universities. Findings from a National Survey in Germany." *Higher Education*, doi:10.1007/s10734-011-9423-1
- Wolbring, T. 2010. "Physical Attractiveness, Gender and the Evaluation of Teaching. A Replication Study of Hamermesh's and Parker's (2005) and

- Klein's and Rosar's (2006) Findings Analyzing Individual Data." *Zeitschrift für Evaluation* 9:29–48.
- Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.
- Yorke, M. 1999. *Leaving Early: Undergraduate Non-completion in Higher Education*. New York: Falmer Press.
- Yurek, L. A., J. Vasey, and D. S. Havens. 2008. "The Use of Self-Generated Identification Codes in Longitudinal Research." *Evaluation Review* 32:1–18.

Bio

Tobias Wolbring is a doctoral student in Sociology at the Ludwig-Maximilians-University Munich, Germany. His research interests include students' evaluations of teaching, causal inference (especially experiments and panel data), and economic sociology.