 https://doi.org/10.3311/PPso.18817


**Periodica Polytechnica Social and Management Sciences**, 31(2), pp. 164–177, 2023

# Assessing the Effects of a Reformed System of Student Evaluation of Teaching

Gergely Dániel Lukáts[1], Zombor Berezvai[2*], Roland Molontay[1,3,4]

[1] Department of Stochastics, Faculty of Natural Sciences, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
[2] Institute of Marketing, Corvinus University of Budapest, Fővám tér 8., H-1093 Budapest, Hungary
[3] MTA-BME Stochastics Research Group, Műegyetem rkp. 3., H-1111 Budapest, Hungary
[4] Department of Management and Business Economics, Faculty of Economics and Social Sciences, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
[*] Corresponding author, e-mail: zombor.berezvai@uni-corvinus.hu

Received: 21 June 2021, Accepted: 23 March 2022, Published online: 16 June 2022

**Abstract**

The importance of student evaluation of teaching (SET) in higher education has increased substantially in the past decades. Despite this increase, there is no consensus on the most efficient way of executing these surveys or about the questions. This study analyses the impact of an SET system reform on the distribution of the responses and the response rate. The reform impacted the questions, the scale used for evaluation, the anchor labels of the scale elements, and the incentives for students to fill out the survey. Our results show that the number of extreme responses increased after the reform, which can be due to the elimination of anchor labels of the middle scale points. Insufficient effort bias also increased; however, the new motivation system doubled the response rate, which helped to collect more representative sets of evaluations. Taking into consideration the relatively small increase in insufficient responses, we believe this incentive can be a valid choice for SET surveys.

**Keywords**

higher education, student evaluation of teaching (SET), questionnaire design, rating scale

## 1 Introduction

Ever since the late 20th century, student evaluations of teaching (SET) have been used by a vast majority of universities across the world to assess the teaching effectiveness of their professors and gain feedback from students about the value and quality of their courses. The ratio of higher education institutes collecting student evaluations in the United States grew from 2% in 1973, to 68% in 1983, and by 1993 over 85% of universities relied on student evaluations (Seldin, 1989). Currently, nearly all higher education institutes collect SET scores around the world, and many use them for determining tenures or promotions.

There are two essential requirements for SET survey results to be adequate to its aim. First, they have to represent the true opinion of the students without any bias. Second, these opinions should reflect teaching effectiveness and quality and not other (irrelevant) factors and characteristics. This paper is only focusing on the first part of it.

Currently, there is no universally accepted way of conducting SET surveys, and different universities apply different methods and questions. Most universities use multidimensional SETs, such as the university examined, but there is no clear consensus on what survey design provides the highest construct-related validity. At some universities filling out SET surveys is mandatory, which may push some students to give insufficient effort when responding, while other universities that have non-mandatory SETs may have lower response rates, which could yield a non-representative pool of student participants (Stark and Freishtat, 2014). Additionally, in general, the choice of online or paper-based surveys can affect the response rate of students (Nowell, 2007; Spooren et al., 2013; Thielsch et al., 2018). Finally, Spooren and Christiaens (2017) pointed out that SET scores also depend on how students perceive and value the SET survey itself.

**Cite this article as:** Lukáts, G. D., Berezvai, Z., Molontay, R. (2023) "Assessing the Effects of a Reformed System of Student Evaluation of Teaching", Periodica Polytechnica Social and Management Sciences, 31(2), pp. 164–177. https://doi.org/10.3311/PPso.18817

The design of the SET system may have a substantial impact on the validity of the results. This paper aims to contribute to this stream of literature. The reform of the SET system at a leading Hungarian university provides the possibility to investigate the effects of the changes in a controlled fashion. In 2017, a comprehensive reform was carried out at the Budapest University of Technology and Economics (BME) involving the number and wording of the questions, the rating scale, and the incentives to fill out the SET surveys.

In this paper, we assess the effects of the reformed system of student evaluation of teaching at BME and compare the current system with the former. We identify the differences across the two systems in three main areas:

- Distribution of the responses and the ratio of extreme responses.
- Change in response rate and its difference between bachelor's and master's students.
- The change in the ratio of insufficient effort responses and its connection to response rates.

The remaining part of the paper is structured as follows. In Section 2, we review the history of SET at BME and describe its reform in detail. In Section 3, we provide a summary of the data together with some explanatory analysis to gain a basic understanding of the underlying data. In the following sections, we investigate the two aspects of the change: Section 4 is devoted to the impact of changes in scale, while Section 5 investigates the impact of new incentives to increase response rates. In Section 6, we discuss the results and provide some initial (normative) inferences on the design of the SET surveys. We also present some recommendations based on our findings that are generally applicable in higher education.

## 2 History and reform of SET at BME

A unified system of student evaluation of teaching was first introduced at the university level at BME in the fall semester of 1999. However, many faculties and departments conducted their own surveys before 1999. When student evaluations at BME were filled out on paper, they had a high response rate which made them representative of student opinions. The system was digitalized in 2005 to reduce the workload required to evaluate the results. However, the shift to a digitalized system had a major drawback: the response rates dropped significantly. A reason may be that students lost trust in the anonymity of their responses. Perhaps even more importantly, in the paper-based system, the instructor gave the surveys to the students during the final class, and they filled them out in class, while the online system required them to use their own free time at home. The low response rate (approximately 30–35% in general) was a major problem because it made the results less reliable. Moreover, the questionnaire was deemed too long, and the student's motivation was deemed low.

As a result, in 2017, the SET system was reformed with multiple changes (Table 1). The questionnaire was made shorter with new, reworded questions. Students were given an incentive to fill out the SET by having a head start when registering for courses in the next registration period, conditional on SET completion. The time for completing the SET surveys was shortened, and the website for publishing results was substantially updated and redesigned to show students that their feedback is being considered. This immediately raised the response rate from 32% to 60% for the fall semester of 2017.

However, one could raise concerns that these changes and incentives may have only artificially inflated the response rate with a significant number of careless new respondents who put insufficient effort when responding. This paper analyses these questions.

## 3 Data overview

The data used in this study are obtained from the official registers of BME. The data include the individual answers for all SET questions, for all courses and all faculties between 2013 and 2019. This means eight semesters of data and 5,086,057 observations for the previous survey design and an additional four semesters of data and 3,034,316 observations using the new survey design. Therefore, this study builds on over eight million observations of dozens of variables, each spanning six years.

Responses to multiple course evaluation surveys from individual students cannot be traced back and analyzed together due to the strict policy of anonymity of the SET system at BME. The timestamp attribute provides the exact time when the student filled out the survey. More precisely, this only gives the time when the survey was submitted and not the individual times of each question, which would have been more useful to detect insufficient effort. However, we could still use these data in later sections to determine which students filled out the survey immediately before course registration (supposedly for the sake of early course registration).

**Table 1** Changes in the SET system of BME in 2017

| Area | Previous system (2013–2017) | Reformed system (from 2017) | Reasons to change |
|---|---|---|---|
| Changes in questions | One open-ended (free text) question about the instructor and another one about the course. | Two open-ended (free text) questions about the course (and the one for the instructor remained). | To motivate students to share both positive and negative experiences. |
| | One question to rate the course in general. | Separate questions for each section (e.g., lecture, practice). | To get a more holistic understanding of the course. |
| | Separate instructor-specific questions for each section (even if they were taught by the same instructor). | Instructor-specific questions were combined for all sections if an instructor taught more sections (e.g., lecture, practice). | To make the survey shorter. |
| | Four, six, and seven questions for lectures, practice classes, and labs, respectively. | There are four questions for all course types. Several course-specific questions were removed (e.g., the question which asked about the relevance and usefulness of the materials taught). | To make the survey shorter. Moreover, the majority of students are normally not able to judge the relevance of the course materials. |
| Survey period | For courses with a mid-year mark, students could fill out the survey after their grade has been administered. For courses with an exam mark, students could fill out the survey after they have received a passing grade for the exam, or, if they failed, then they could fill out the survey after the exam period has terminated. The final deadline was the third day of the third week of the following semester. | The conditions for filling out the survey remained the same. However, the earliest start date was changed to the first day of the second week of the exam period, and the end of the survey period was changed to the sixth workday after the exam period had ended. | The survey period was shortened to publish the results earlier; therefore, instructors can utilize and incorporate the feedback in the next semester. |
| Publishing of results | The website used for publishing the results was not user-friendly. | All results (including historical results) are available on the renewed website in a transparent manner for all citizens of BME. | Students can see that their feedback is respected, and they can also collect information about instructors. |
| Grading scale | The questionnaire used a five-point Likert scale, with labelled categories: "Very Bad", "Bad", "Average", "Good" or "Excellent". | A six-point Likert scale is used, where only the endpoints of the scale are labelled with text ("Not at all"; "Fully"); other anchor labels are not used. | To avoid midpoints, but with no specific concern. |
| Incentive for students | Students were motivated with sweepstakes including high-value prizes (such as iPads or festival passes). Social media campaigns were also run to promote and raise awareness of the importance of the survey. | An early course registration opportunity was introduced for students who filled out the survey. | To design an incentive that is truly important to students. Popular elective courses can become full very quickly, and for mandatory courses with multiple possible sections, certain time slots can also become full, sometimes within an hour. |

The data structure of the two surveys differs heavily. The main difference is that in the previous survey, instructors received a separate evaluation for each lecture and practice course that they taught, and these were later combined for the instructor's overall score. In the new survey, instructors received one evaluation for the entire course, and the sections were evaluated separately from the instructors. When aggregating the evaluations from the previous system, we arrived at a Semester-Course-Section-Instructor granularity. This means that we count one observation for each instructor, in each section, in each course, in each semester. To keep the data consistent, we aggregated the new SET results with the same granularity, by counting each student's evaluation of the instructor once for each section that the instructor taught. In other words, while students could give the instructor different ratings for the lecture and practice courses in the previous questionnaire, we had to assume in the new SET survey that they always gave the same rating for both courses. In this way, the data for the two systems could be combined. Additionally, the numerical results in the previous version were scaled by 6/5 to make the scales consistent, with both being 1–6.

Both versions of the survey contain a question asking, "Did this instructor teach you?" to try to assess whether the correct instructor is listed in the administrative system. To make the data more reliable, we used this question to filter out all classes where the documented instructor did not actually teach the course, so we filtered out all instructors who did not have at least 80% 'Yes' response for a given section.

The survey data were also enriched by other course-, student, and instructor-related characteristics. An important attribute is the grade distribution, meaning how many students received a 1, 2, 3, 4 or 5 (worst-to-best) in the class.

It is worth noting that while SET scores have a Semester-Course-Section-Instructor granularity, data on grading do not have instructor-level granularity for sections with multiple instructors.

Additionally, other course-related characteristics include the degree level (undergraduate, graduate), course requirement (mid-year mark, exam grade, fail/pass), section type (lecture, practice), language (mostly Hungarian, English, and German), course credit, and section size.

In terms of student characteristics, for each section, we calculate the averages of cumulative grade point average (GPA), university admission point scores (for undergraduate students), and the ratio of student genders. This is possible for each Semester-Course-Section combination. With the average cumulative GPA calculated, we could define a relative grade variable, which is the difference between the average course grade in the section and the average cumulative GPA of the section.

As we have discussed, instructor characteristics can influence SET scores (Boring, 2017; Hornstein, 2017). In this study, we used the gender, age, and rank of the instructors. While there are 20 possible instructor ranks in the administrative system, we grouped them into five categories: Level 1, Level 2, Level 3, Level 4 and Other, which roughly correspond to PhD student or assistant lecturer, assistant professor, associate professor, full professor, and specific instructors that are outside of the university ranking system.

We categorized instructors' ages into factor levels with cut points at 25, 35, 50 and 65.

Since the analysis was done on course-level data (not based on individual responses), we aggregated the data to a Semester-Course-Section-Instructor granularity. We obtained 82,504 observations that we filtered to only include observations that satisfy some reliability criteria (at least five evaluations per section, at least 30% response rate on SET, maximum 15% of data is missing, the course credit is between 2 and 9, and at least 66% of the evaluations are valid). This resulted in a total of 26,659 observations, which is the final dataset that we use in the rest of this paper.

An overview of the attributes of the filtered dataset can be found in Table 2 and Table 3. We can observe that the response rate is much higher in the current SET system. The course size in the previous SET dataset is slightly higher because our filtering criteria require at least 5 evaluations. This is easier to fulfil in smaller courses after the SET reform. For the same reason, the percentage of practice courses increased from 49% to 60% in the current version. This can be explained by the fact that practice courses generally have smaller class sizes. This filtering is likely not causing any selection bias as other course level attributes (relative grade, course credit, course average grade, female student share, average cumulative GPA, average admission point, and instructor age) do not show any major change. In the case of cumulative GPA and admission points, the difference is not significant at 5% level. For the other variables, the difference is significant due to the very large sample size, but the magnitude is minor.

The lack of young instructors in the dataset suggests that a significant number of instructors with missing age are PhD students, who did not have their birth date filled out in the administrative system.

Additionally, the correlation matrices show a similar structure in the two examined periods (Fig. 1). The average grade, cumulative GPA, and admission point score variables all positively correlate with each other, which is expected since all these variables measure academic abilities. The relative grade and average grade variables are also

**Table 2** Descriptive statistics for the numerical variables (Semester-Course-Section-Instructor data)

| Variable | Previous SET system | | | | | Current SET system | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Obs. | Mean | SD | Min | Max | Obs. | Mean | SD | Min | Max |
| SET score (not rescaled) | 14,395 | 4.07 | 0.62 | 1.00 | 5.00 | 12,264 | 4.88 | 0.83 | 1.38 | 6.00 |
| Relative grade | 14,395 | 0.10 | 0.63 | −2.14 | 2.71 | 12,264 | 0.00 | 0.63 | −2.34 | 2.52 |
| Response rate | 14,395 | 0.48 | 0.13 | 0.30 | 1.00 | 12,264 | 0.61 | 0.17 | 0.30 | 1.00 |
| Course credit | 14,395 | 3.80 | 1.29 | 2.00 | 9.00 | 12,264 | 4.05 | 1.24 | 2.00 | 9.00 |
| Course size (# students) | 14,395 | 55.43 | 76.13 | 5.00 | 708.00 | 12,264 | 42.44 | 56.86 | 5.00 | 702.00 |
| Course avg. grade | 14,395 | 3.30 | 0.82 | 1.09 | 5.00 | 12,264 | 3.21 | 0.85 | 1.00 | 5.00 |
| Female student share | 14,395 | 0.31 | 0.24 | 0.00 | 1.00 | 12,264 | 0.29 | 0.23 | 0.00 | 1.00 |
| Avg. cumulative GPA | 14,395 | 3.20 | 0.53 | 1.49 | 4.84 | 12,264 | 3.20 | 0.55 | 1.69 | 4.96 |
| Avg. admission points | 11,635 | 419.76 | 25.05 | 288.20 | 488.00 | 10,487 | 419.41 | 26.20 | 350.00 | 496.00 |
| Instructor age (year) | 13,149 | 45.85 | 13.21 | 20.00 | 88.00 | 11,313 | 46.80 | 12.75 | 20.00 | 91.00 |

**Table 3** Relative and absolute frequency of the factor variables used (Semester-Course-Section-Instructor data)

| Variable | Factor level | Previous SET | | Current SET | |
|---|---|---|---|---|---|
| | | Freq. | Prop. | Freq. | Prop. |
| Course period | Morning | 2,659 | 18.47% | 2,142 | 17.47% |
| | Midday | 7,583 | 52.68% | 6,693 | 54.57% |
| | Evening | 2,425 | 16.85% | 1,885 | 15.37% |
| | Weekend | 1,417 | 9.84% | 1,431 | 11.67% |
| | (Missing) | 311 | 2.16% | 113 | 0.92% |
| Course type | Lecture | 7,339 | 50.98% | 4,981 | 40.61% |
| | Practice | 7,056 | 49.77% | 7,283 | 59.39% |
| Course requirement | Exam | 7,165 | 49.77% | 6,634 | 54.09% |
| | Mid-year mark | 7,230 | 50.23% | 5,630 | 45.91% |
| Course level | Undergraduate | 10,407 | 72.30% | 8,549 | 69.71% |
| | Graduate | 3,242 | 22.52% | 2,316 | 18.88% |
| | Other | 746 | 5.18% | 1,399 | 11.41% |
| Instructor age | (0,25] | 277 | 1.92% | 98 | 0.80% |
| | (25,35] | 3,252 | 22.59% | 2,340 | 19.08% |
| | (35,50] | 4,600 | 31.96% | 4,720 | 38.49% |
| | (50,65] | 4,013 | 27.88% | 3,107 | 25.33% |
| | (65,100] | 1,007 | 7.00% | 1,041 | 8.49% |
| | (Missing) | 1,246 | 8.66% | 958 | 7.81% |
| Instructor gender | Male | 10,693 | 74.28% | 9,434 | 76.92% |
| | Female | 2,456 | 17.06% | 1,872 | 15.26% |
| | (Missing) | 1,246 | 8.66% | 958 | 7.81% |
| Instructor rank | Level1 | 2,497 | 17.35% | 1,652 | 13.47% |
| | Level2 | 3,283 | 22.81% | 2,929 | 23.88% |
| | Level3 | 4,100 | 28.48% | 3,873 | 31.88% |
| | Level4 | 973 | 6.76% | 927 | 7.56% |
| | Other | 2,296 | 15.95% | 1,925 | 15.70% |
| | (Missing) | 1,246 | 8.66% | 958 | 7.81% |
| Semester | 2013/14/1 | 2,157 | 14.98% | – | – |
| | 2013/14/2 | 2,129 | 14.79% | – | – |
| | 2014/15/1 | 2,440 | 16.95% | – | – |
| | 2014/15/2 | 1,597 | 11.09% | – | – |
| | 2015/16/1 | 2,081 | 14.46% | – | – |
| | 2015/16/2 | 1,434 | 9.96% | – | – |
| | 2016/17/1 | 1,357 | 9.43% | – | – |
| | 2016/17/2 | 1,200 | 8.34% | – | – |
| | 2017/18/1 | – | – | 3,356 | 27.36% |
| | 2017/18/2 | – | – | 2,597 | 21.18% |
| | 2018/19/1 | – | – | 3,375 | 27.52% |
| | 2018/19/2 | – | – | 2,936 | 23.94% |

highly correlated by design. We can observe that variables that measure student ability have decent-sized correlations with response rates. This implies that students with higher grades are more likely to fill out the SET survey. It can

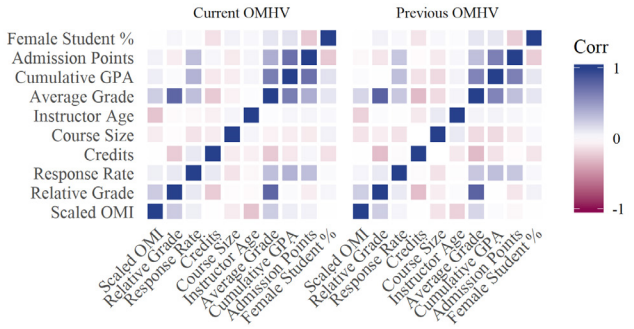


**Fig. 1** Correlation matrices of the variables considered

be reassuring that in the current SET system, this correlation has increased from 0.31 to 0.37, which could mean that many of the new students responsible for the increase in response rates are also academically advanced.

## 4 Impact of the changes in scale

The BME SET survey uses a Likert scale to gauge student satisfaction with the instructor and the course. Previously, a five-point Likert scale was used where each response category was labelled, while the current scale was a six-point Likert scale, which only uses labels at the endpoints.

Creating a forced-choice scale by omitting the midpoint rating (3) can contribute to higher reliability (e.g., Masters, 1974; Weems and Onwuegbuzie, 2001) as several respondents misuse the midpoint. Connected to this, Daher et al. (2015) measured the reliability of the Spiritual Well-Being Scale with different rating scale categories and found that six categories proved to be more reliable than three or four. In contrast, Weijters et al. (2010) found that including midpoints lowers extreme response bias. Similarly, Chen et al. (2015) found based on an eye-tracking experiment that a five-point Likert scale is optimal; it produces the lowest level of extreme response bias. Nadler et al. (2015) reported that a four-point Likert scale, a four-point Likert scale with an option for "no opinion", and a five-point Likert scale produced similar results in a political opinion survey. Reliability was higher for the five-point Likert scale, but the difference was not statistically significant either.

Chyung et al. (2017) summarized several studies analyzing the impact of midpoints on Likert-type scales. They suggested that respondents might misuse the midpoint if they are ambivalent about the topic, or if their answer depends on other factors. This can be true for SET surveys as the instructor often has positive and negative attributes at the same time and the grade can also influence the rating (Ewing, 2012; Berezvai et al., 2021). An important

tactic suggested by Chyung et al. (2017) is to include the "I don't know" option in forced-choice scales to offer the opportunity to not answer.

Anchor labels can also affect the results and the extreme response bias of surveys, as shown by Huck and Jacko (1974), and Weijters et al. (2010). However, some other studies (e.g., Newstead and Arnold, 1989; Chang, 1997) found no or very minor effects of the labels. According to the results of Weng (2004), labelling all anchors can be favorable from a reliability point of view, especially when the number of answer options is high. Wouters et al. (2014) found that labelling only the endpoints was associated with higher means and higher variances for the answers. Furthermore, the types of labels can also impact the results. Tsekouras (2017) showed that emotional labels increased extreme response bias compared to neutral labels.

Regarding student evaluations of teaching specifically, Newstead and Arnold (1989) found that labelling all anchors or only the endpoints causes no difference in instructor ratings. In contrast, the results of Zipser and Mincieli (2018) indicate that after modifying the question about the overall evaluation of the instructor and labelling all anchors of the scale, a decrease was caused in average ratings. Furthermore, Dolnicar and Grün (2009) found that the cultural background of students influenced their proneness to extreme response bias, with Asian respondents having milder response styles than American, Australian, and Hispanic students.

To summarize the available evidence, despite the widespread usage and popularity of Likert-type scales, there is a lack of consensus on the optimal format and structure for obtaining unbiased results. There are also different ways to measure the reliability of different scales. Weijters et al. (2010) focused on the bias from three key response styles: extreme response style (ERS), net acquiescence response style (NARS), and misresponse to reversed items (MR). ERS is the tendency to disproportionately use extreme response categories in a rating scale (such as the endpoints), NARS is the bias arising from tendency to agree with items, and MR stands for the bias defined when respondents answer the same question differently, depending on the direction of the scale.

In this paper, we attempt to quantify the level of ERS in BME SET survey responses and determine whether it has increased following the SET reform to evaluate the effect of scale changes on the reliability of the results. Most studies on rating scale bias and design used controlled experiments (e.g., Weems and Onwuegbuzie, 2001; Weng, 2004; Wouters et al., 2014). Our data differ significantly because although our sample size is extremely large, we only have two surveys, which we cannot change for the sake of our study. Thus, we cannot isolate the differences since multiple changes occurred simultaneously. Therefore, we cannot identify the individual effects of each change, but we can quantify the effect of the overall SET reform on ERS.

As a first step, we examined the distribution of the responses for the question asking about the instructor's overall performance, separately in the two surveys. We can observe from Fig. 2 that the right tail extreme value is considerably larger in the current SET. On the left tail, the lowest response also appears to have grown, although it is still negligible, even when combined with the second-lowest response. The graph implies that students completing

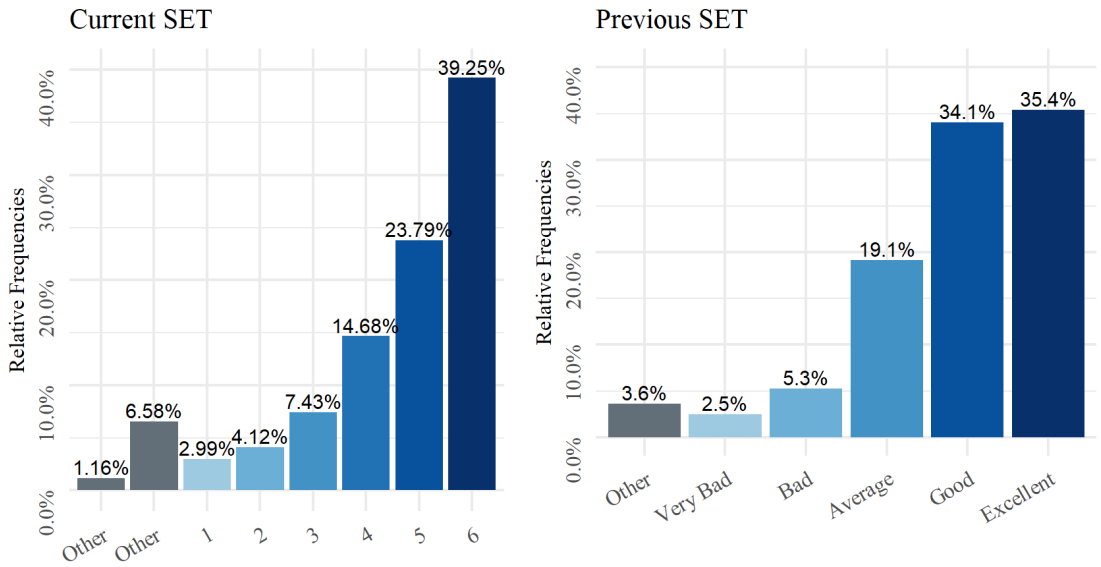


**Fig. 2** Distribution of the SET scores, (a) Current SET ($n$ = 12,264), (b) Previous SET ($n$ = 14,395)

the new survey have a higher tendency to give extreme responses. Looking towards the middle of the distribution, we can see in the previous survey there was a fairly large jump between "average" and "good", so 70% of students ended up giving evaluations using only the last two categories. In the new version, while the tails of the distribution are higher, the center of the distribution seems to have linearly expanded into three categories, with smaller jumps. Interestingly, the average (rescaled) SET score is almost identical in the two systems (4.88; the difference is not significant even at 10% level).

As seen in Fig. 2, we can observe that the extreme response ratio has increased. To identify the impact of the SET reform, we will use regression analysis and control for other observable factors. Following Weijters et al. (2010), we will measure ERS as the log odds ratio since it is not as skewed as the simple odds ratio, and it approximately follows a standard normal distribution. Namely, we define ERS as follows:

$$\mathrm{ERS} = \ln\left(\frac{\#\,\text{Extreme responses} + 1}{\#\,\text{Non-extreme responses} + 1}\right), \tag{1}$$

where extreme responses are defined as 1 or 6 in the current SET survey and as "Very bad" or "Excellent" in the previous version.

Using the weighted ordinary least squares (OLS) method, we estimate Eq. (2):

$$\mathrm{ERS}_{it} = \beta_0 + \beta_1 \cdot 1\{\text{Current SET}\} + \beta \cdot X_{it} + \varepsilon_{it}, \tag{2}$$

where $t$ denotes different semesters, while $i$ denotes the individual Course-Section-Instructor combinations within a given semester. $\beta_1$ is the coefficient of interest showing the effect of the reform on ERS. $X_{it}$ contains all the course, student and instructor-specific control variables specified in Tables 2 and 3. Finally, $\varepsilon_{it}$ refers to the idiosyncratic error term.

Standard regression methodology gives equal importance to all the observations (in our case to every Semester-Course-Section-Instructor combination). However, the number of students participating in a section and, therefore, evaluating the instructors can vary heavily across sections. Although we use section level data, every evaluation should have the same importance to avoid heteroskedasticity of the error terms corresponding to the section level averages. This can be achieved by weighting the observations with the number of respondents in each section and instructor-level clustered standard errors were also used to account for the individual characteristics of the instructors.

Instructors could have various characteristics that are constant in time and can affect the extent of ERS but cannot be captured by the observed instructor-specific variables. For example, if the instructors talk about controversial topics in class regularly, they could have more extreme responses. Similarly, physical attractiveness and humor could also influence the levels of ERS. Due to the short time period, weighted fixed effects regression could capture these traits, which requires the estimation of Eq. (3):

$$\begin{aligned}\mathrm{ERS}_{ijt} &= \beta_0 + \beta_1 \cdot 1\{\text{Current SET}\} \\ &+ \beta \cdot X_{ijt} + \alpha_j + \varepsilon_{ijt},\end{aligned} \tag{3}$$

where $j$ denotes the individual instructors who taught course $i$ in semester $t$, $a_j$ picks up on the effects of all characteristics of the instructor that did not change over the 12 semesters observed. Similar to the weighted OLS, we used the same weights in this regression and clustered the standard errors by instructors. The dataset was filtered further than the dataset used for the OLS estimation because we only included instructors who taught at least two semesters during the previous SET version and at least one semester after the SET system reform. Additionally, instructors had to teach at least four semesters altogether to be included in the sample.

The results of the regressions described above can be found in Table 4. In the previous SET survey, 2 out of 5 categories were considered extreme responses, while in the current version, only 2 out of 6 categories were considered extreme responses, so with everything else unchanged, we would expect the coefficient of the current version indicator to be negative. However, at a university level, it is positive and highly significant in both regressions, with a coefficient of 0.211 from the OLS regression and 0.151 from the FE regression. This means that the ratio of extreme responses in the SET survey was higher between 2017 and 2019, likely due to the reform. Thus, we conclude that introducing a forced-choice scale and labelling only the extreme values on the scale may have increased the ratio of extreme responses. This result is contrary to the finding of Newstead and Arnold (1989) and partially to Zipser and Mincieli (2018). Differences in the SET questionnaire, the location, or the time between the studies might have contributed to the different results. However, as SET surveys differ across universities, this calls attention to the importance of local studies helping university staff create the best possible SET survey.

**Table 4** Results of the weighted OLS and weighted fixed effect regressions

| Independent variable | Weighted OLS | Weighted fixed effect |
|---|---|---|
| Intercept | −0.535*** (0.135) | −0.227** (0.081) |
| Version: Current | 0.211*** (0.030) | 0.151*** (0.021) |
| Relative grade | 0.423*** (0.039) | 0.314*** (0.024) |
| Credit | 0.064* (0.028) | −0.000 (0.012) |
| log (Course size) | −0.083*** (0.022) | −0.187*** (0.016) |
| Female student share | −0.134 (0.122) | −0.466*** (0.068) |
| Female instructor × Female student share | 0.329 (0.290) | – |
| Language: English | 0.053 (0.239) | −0.054 (0.139) |
| Language: Other | −0.050 (0.093) | 0.013 (0.073) |
| Period: (Missing) | 0.844*** (0.126) | 0.313** (0.116) |
| Period: Evening | 0.086 (0.046) | 0.072* (0.030) |
| Period: Weekend | 0.028 (0.053) | 0.061* (0.029) |
| Period: Morning | 0.005 (0.046) | 0.024 (0.023) |
| Type: Practice | 0.139*** (0.036) | −0.031 (0.021) |
| Requirement: Exam | 0.156** (0.053) | 0.046 (0.026) |
| Degree: Other | 0.212** (0.067) | 0.031 (0.044) |
| Degree: Graduate | 0.058 (0.046) | 0.066 (0.035) |
| Age ∈ (25,35] | −0.001 (0.124) | 0.230* (0.104) |
| Age ∈ (35,50] | −0.102 (0.132) | 0.262* (0.129) |
| Age ∈ (50,65] | −0.467*** (0.133) | 0.303* (0.144) |
| Age > 65 | −0.436** (0.146) | 0.228 (0.158) |
| Missing instructor data | 0.325* (0.137) | −0.104 (0.156) |
| Gender: Female | −0.189 (0.140) | - |
| Rank: Level2 | −0.069 (0.084) | 0.138* (0.070) |
| Rank: Level3 | 0.051 (0.094) | −0.005 (0.096) |
| Rank: Level4 | 0.304* (0.154) | −0.041 (0.125) |
| Rank: Other | 0.037 (0.091) | 0.053 (0.078) |
| *R*-square | 0.162 | 0.682 |
| Sample size (*n*) | 23,659 | 22,880 |

Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

## 5 Impact of new incentives to increase the response rate and their effects on insufficient effort responses

Response rates are a particularly important factor in student evaluation of teaching, as sufficiently high response rates can ensure the validity of the scores. Zumrawi et al. (2014) simulated the minimum response rates required to achieve this objective. As the proportion of positive answers is approximately 70–80% (Fig. 2), and smaller courses dominate the sample (Table 2), the response rate in the old system (approximately 30%) is deemed to be too low. The increase in the response rate was an objective of the reform (Table 1) that was successfully achieved. Ernst (2014) explored the motivations of students to fill out SET surveys and the reform addressed several of them.

- The website used for publishing SET results was redesigned and made more user-friendly so students could easily see that their feedback was being considered.
- The questionnaire was shortened fairly, which meant that students needed to sacrifice less time to respond.
- The survey period was shortened which makes it possible to publish the results earlier so students can observe a more immediate impact of their responses.
- However, the main driver behind response rate growth may be the newly introduced early course registration for students who filled out the survey. Similar incentives are generally useful in substantially increasing response rates (Zipser and Mincieli, 2018; Lipsey and Shepperd, 2021).

If the early course registration incentive is the main reason for the response rate growth, then we suspect that there will be a significant increase in the number of responses immediately before the course registration dates, when procrastinating students remember that they need to fill out the SET questionnaires to be eligible for early course registration. Fig. 3 illustrates the time evolution of responses both before (first two rows) and after (last row) the incentive was introduced. We can observe that the response rate increased drastically after the SET reform. Moreover, we can also see that the peaks in the number of responses used to be rather hectic before the reform and were mostly driven by SET reminder emails sent out to the students (marked by light blue circles). After the reform, we can observe that responses increased by over 500% in each semester in the week before the course registration dates and fell back dramatically on the remaining few days after course registration started. (The first peak, colored dark blue, occurred in early January 2018 when the early course registration policy was first announced to the students.) Based on the time series, we can conclude that the early course registration incentive had a very significant impact on the response rates of the current SET survey. In fact, we can see that over 50% of the responses were submitted in the week just before the early course registration started. Naturally, we assume that many of these students would fill out the survey regardless of the incentive, but the number of responses in the peaks approximately matches the amount of response rate growth after the SET reform.


**Fig. 3** Distribution of SET response by time

Furthermore, response rates by different subgroups can also be analyzed to evaluate the effect of the new incentive. We examined the difference in response rates by degree level (undergraduate and graduate). Due to the higher number of undergraduate students, there are more courses with multiple possible sections compared to graduate courses. While only 44% of graduate courses have at least two sections, this ratio is 61% for bachelor's courses. This means that early course registration is on average more valuable for undergraduate students, as they might prefer one particular section over another, and early course registration can help them to ensure their seat in the desired section.

Fig. 4 shows that response rates for graduate and undergraduate programs were approximately the same for each semester in the previous SET system. However, starting from the reform (the fall semester of 2017), the response rates separated by approximately 20%-points. This is a clear indication that the motivation system was the main driver of response rate growth.

Data indicate that after the reform, the number of participants doubled, perhaps most of them responding due to the course registration incentive, and not just because they would like to provide useful feedback. This could cause these students to respond without sufficient effort. Seemingly random and careless responses provided by unmotivated students can threaten the quality of the SET survey results. This incentive could, therefore, raise a major

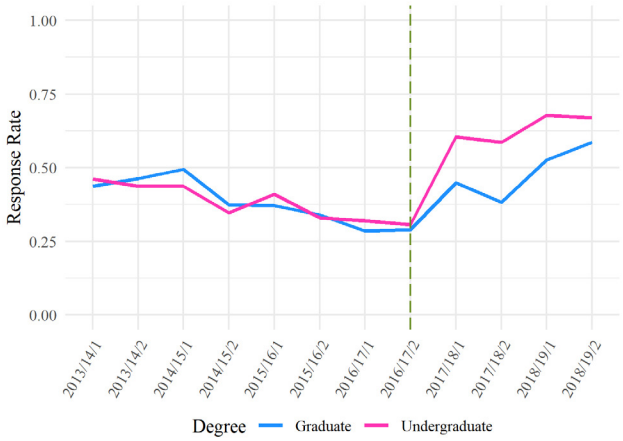


**Fig. 4** Response rate by degree level

concern around the reliability of the SET system. In this section, we investigate whether the incentive increased the number of students giving insufficient effort when responding and whether the SET survey is still reliable.

Huang et al. (2012) identified the most frequent approaches for detecting insufficient effort responses (IER) as the infrequency approach (identifying items where most attentive respondents should provide the same response), response time approach (looking at the time each participant spent responding to the survey) and response pattern approach (analyzing the pattern of response options). Based on these approaches and the available data, we introduce two tests for detecting inattentive respondents in BME SET.

- Test 1: Since the question "Did this instructor teach you?" should have a straightforward answer, if the majority response was at least 80% then we label students who chose the minority answer as potentially inattentive. Since this question did not change in the reform, this test can be calculated for both periods.
- Test 2: There were two questions connected to the evaluation of students' performance on the course. The first question asked whether the evaluations were in line with the knowledge taught by the instructor, on a 1 to 6 Likert scale. A different question asked about the forms and quantities of evaluations. This was a multiple-choice question where one of the options was that evaluations were in line with the knowledge taught. This test assesses the average rating of the first question among those who selected this particular option in the latter question. This test can be calculated for the after-reform period only.

First, we assess the two tests graphically in Figs. 5 and 6, which is followed by a more detailed comparison (Tables 5 and 6). These tests were run on individual responses, i.e., using approximately 8 million observations.

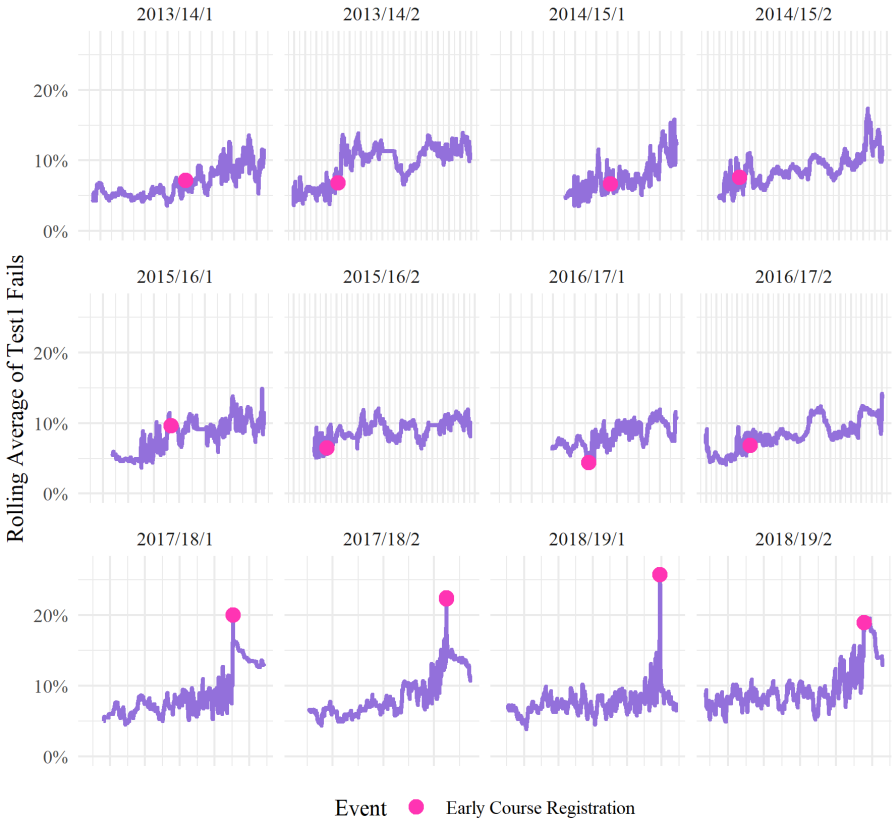


**Fig. 5** Percent of responses failed on Test 1 (rolling average of the past 1,000 responses in the given point in time)

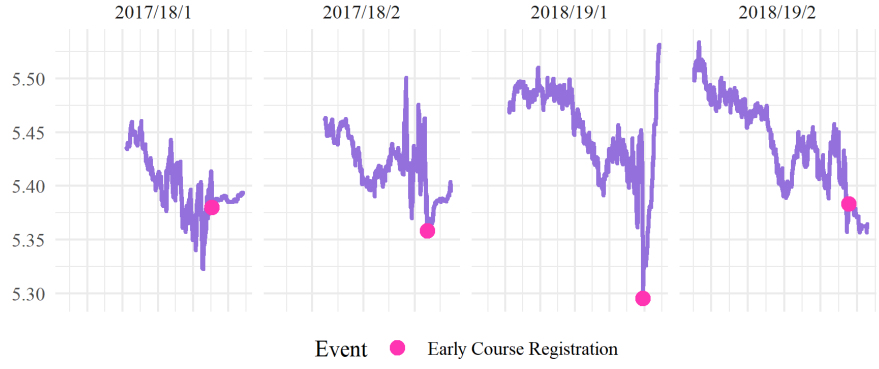


**Fig. 6** Average rating of the question "The evaluation was in line with the knowledge taught" (1 – not at all; 6 – completely) among those selecting the option "the form of evaluation was in line with the knowledge taught" in a later multiple-choice question (rolling average of the past 1,000 responses in the given point in time)

**Table 5** Detecting insufficient efforts in responses based on Test 1

| Semester | Response time | Insufficient effort | |
|---|---|---|---|
| | | Freq. | Prop. |
| 2013/14/1 | General | 4,581 | 8.44% |
| 2013/14/2 | General | 3,994 | 8.17% |
| 2014/15/1 | General | 4,489 | 7.83% |
| 2014/15/2 | General | 3,138 | 8.29% |
| 2015/16/1 | General | 4,113 | 8.00% |
| 2015/16/2 | General | 2,927 | 8.07% |
| 2016/17/1 | General | 2,907 | 7.96% |
| 2016/17/2 | General | 2,409 | 7.85% |
| 2017/18/1 | Early Registration | 2,240 | 9.81% |
| | General | 3,234 | 7.55% |
| 2017/18/2 | Early Registration | 3,796 | 11.81% |
| | General | 1,505 | 7.78% |
| 2018/19/1 | Early Registration | 3,372 | 11.64% |
| | General | 3,284 | 7.49% |
| 2018/19/2 | Early Registration | 1,581 | 13.25% |
| | General | 4,217 | 8.71% |

**Table 6** Detecting insufficient efforts in responses based on Test 2

| Semester | Response time | Average | Number of responses |
|---|---|---|---|
| 2017/18/1 | Early Registration | 5.35 | 8,775 |
| | General | 5.38 | 18,690 |
| 2017/18/2 | Early Registration | 5.40 | 11,987 |
| | General | 5.42 | 8,536 |
| 2018/19/1 | Early Registration | 5.37 | 10,685 |
| | General | 5.44 | 18,495 |
| 2018/19/2 | Early Registration | 5.38 | 4,294 |
| | General | 5.43 | 19,295 |

Fig. 5 indicates that the proportion of failed responses peaked when early course registration started in the new system. The same pattern was not observable at all in the old system; hence, the incentive led to an increase in the number of insufficient effort responses.

To determine the size and significance of the changes, we compared the SET results submitted in the three days up to the start of early course registration ("Early registration") with the results submitted at another time ("General"). We calculated the percentage of students who failed on Test 1 separately for the two categories and compared the results in Table 5.

While the proportion of insufficient effort responses was approximately 8% in the old system, this did not change significantly if we compared the general period of the new system ($p = 0.0614$; difference: -0.16%). However, the early registration period shows a significantly higher proportion of insufficient effort responses ($p = 0.000$; difference: 3.38% compared to the old system; $p = 0.000$; difference 3.53% compared to the general period of the new system).

Results are reinforced by Test 2. Fig. 6 clearly shows that once the early course registration period was approaching, the average rating decreased indicating an increase in inattentive responses. The difference is also statistically significant ($p = 0.000$; Table 6), albeit not very large in absolute terms (lower by 0.04 on a 1–6 scale).

In conclusion, we can state that the percent of insufficient effort responses did not deviate when comparing the old system and the new system, apart from the 3 days before the early course registration. Additionally, from both tests, we can conclude that the percentage of students identified as possibly inattentive was significantly higher among those who responded directly before the course registration date. Hence, the redesign of the questionnaire did not impact the proportion of insufficient effort responses, and only the incentive system made a difference. Nevertheless, the increase in absolute terms is not so high. Even in this subsample, only a small fraction of students failed the insufficient effort tests, and the increase was much less than the increase in total response rates, which is the advantage of the new incentive.

## 6 Discussion of the results

In this study, we investigated the impact of changes in the SET survey system at a Central European university (BME). The analysis focused on two main aspects of the reform: the change in the grading scale and the change in the incentive to increase the response rate.

Regarding the change in the grading scale, we observed that the frequency of the two extreme responses increased substantially, and that the frequency distribution had a different shape in the new system. The increase in extreme responses might not be favorable; however, since we do not know the true distribution, we cannot state that the change was disadvantageous.

Generally, providing anchor labels for all response categories in an SET survey might be favorable as this can better guide respondents. Regarding the five- or six-point Likert scale, neither the literature nor this analysis provided clear and irrefutable answers.

In our opinion, changing to a six-point rating scale was a valid choice, because the rating distribution was more spread out, while previously it was concentrated around the top two ratings. However, the choice to only keep anchor labels at the two ends of the scale may have

potentially increased extreme response bias. Thus, we recommend designing SET questionnaires with anchor labelling for all categories.

The response rate growth of the survey is likely due, by an overwhelming majority, to the new student incentive of early course registration. Our results show that insufficient effort responses were somewhat more frequent among those who responded just before the start of the early course registration. However, this difference is – albeit statistically significant – not very large in absolute terms. The increase in response rate was approximately 30%-points if compared the semesters before and after the reform. With this increase, the SET survey reached the minimum desired response rate for validity (Zumrawi et al., 2014).

According to the first test, which is the best one for capturing insufficient effort in our opinion, the ratio of insufficient effort responses was below 10% overall and the increment was even smaller, below 4%-points. This shows that, overall, there is a sufficient gain in the reliability of the data due to the new incentive. This is also supported by Zipser and Mincieli (2018) who found that incentives did not significantly impact SET scores. On the other hand, the substantial increase in response rates can help to collect more reliable and representative responses and opinions.

The insufficient effort responses can be tracked and discarded from the SET scores in several ways to avoid potential bias. One way is the one applied by BME, to ask whether the given instructor taught the section. A second potential way is to save and store response time data so that insufficient effort responses can be identified more accurately and easily. With these strategies, the potential negative impacts of an incentive system could be mitigated, in our opinion, and the benefits of higher response rates could be further harvested.

On the other hand, it is important to note that different universities might use very different SET surveys, which concerns the generalizability of the results. Furthermore, since we analyzed a reform impacting several aspects of the SET survey at the same time, it is not possible to separate the effects from each other and draw a causal relationship. The latter is generally very difficult in social and educational research (Morrison, 2012; Morrison and van der Werf, 2016). This is an important limitation of our study.

Finally, even high response rates and adequate questions do not necessarily lead to unbiased SET results (Boring 2017; Hornstein 2017). Despite their extensive usage, many doubt the validity of SET in determining instructor quality or teaching effectiveness (Uttlet al., 2017; Wang and Williamson, 2022), since various factors irrelevant to teaching quality have been associated with SET scores such as class size (de Carvalho Andrade and de Paula Rocha, 2012), lenient grading (Berezvai et al., 2021; Boring et al., 2016; de Carvalho Andrade and de Paula Rocha, 2012; Ewing, 2012), gender (Boring, 2017), age (de Carvalho Andrade and de Paula Rocha 2012), likability (Feistauer and Richter, 2018), and physical attractiveness (Wolbring and Riordan, 2016). Some universities have introduced rating interpretation guides to make the viewer of the SET scores aware of these biases (Palmer and Smith, 2013). Furthermore, using a computer simulation framework, Esarey and Valdes (2020) demonstrated that unbiased, reliable, and valid student evaluations can still be unfair, and the authors urge the use of multiple dissimilar measures to evaluate the teaching performance of faculty members. For example, to complement SET surveys, Tóth et al. (2013) highlighted the importance of organizing brainstorming sessions on the quality of education involving students, to understand the problems highlighted by SET surveys. Furthermore, putting more emphasis on comprehensive university peer review of teaching programs could also complement the SET survey results (Tóth et al., 2017).

**Acknowledgement**

We would like to thank Mihály Szabó and Bálint Csabay for their help in data collection and interpretation.

The work of R. Molontay is supported by the ÚNKP-21-4-II New National Excellence Program of the Ministry for Innovation and Technology, Hungary.

**References**

Berezvai, Z., Lukáts, G. D., Molontay, R. (2021) "Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching", Assessment & Evaluation in Higher Education, 46(5), pp. 793–808. https://doi.org/10.1080/02602938.2020.1821866

Boring, A. (2017) "Gender biases in student evaluations of teaching", Journal of Public Economics, 145, pp. 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., Stark, P. B. (2016) "Student Evaluations of Teaching (mostly) do not Measure Teaching Effectiveness", ScienceOpen Research, pp. 1–11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Chang, L. (1997) "Dependability of Anchoring Labels of Likert-Type Scales", Educational and Psychological Measurement, 57(5), pp. 800–807. https://doi.org/10.1177/0013164497057005005

Chen, X., Yu, H., Yu, F. (2015) "What is the optimal number of response alternatives for rating scales? From an information processing perspective", Journal of Marketing Analytics, 3(2), pp. 69–78. https://doi.org/10.1057/JMA.2015.4

Chyung, S. Y., Roberts, K., Swanson, I., Hankinson, A. (2017) "Evidence-based survey design: The use of a midpoint on the Likert scale", Performance Improvement, 56(10), pp. 15–23. https://doi.org/10.1002/PFI.21727

Daher, A. M., Ahmad, S. H., Winn, T., Selamat, M. I. (2015) "Impact of rating scale categories on reliability and fit statistics of the Malay Spiritual Well-Being Scale using Rasch Analysis", Malaysian Journal of Medical Sciences, 22(3), pp. 48–55.

de Carvalho Andrade, E., de Paula Rocha, B. (2012) "Factors Affecting the Student Evaluation of Teaching Scores: Evidence from Panel Data Estimation", Estudios De Economia, 42(1), pp. 129–150. https://doi.org/10.1590/S0101-41612012000100005

Dolnicar, S., Grün, B. (2009) "Response style contamination of student evaluation data", Journal of Marketing Education, 31(2), pp. 160–172. https://doi.org/10.1177/0273475309335267

Ernst, D. (2014) "Expectancy theory outcomes and student evaluations of teaching", Educational Research and Evaluation, 20(7–8), pp. 536–556. https://doi.org/10.1080/13803611.2014.997138

Esarey, J., Valdes N. (2020) "Unbiased, reliable, and valid student evaluations can still be unfair", Assessment & Evaluation in Higher Education, 45(8), pp. 1106–1120. https://doi.org/10.1080/02602938.2020.1724875

Ewing, A. M. (2012) "Estimating the impact of relative expected grade on student evaluations of teachers", Economics of Education Review, 31(1), pp. 141–154. https://doi.org/10.1016/j.econedurev.2011.10.002

Feistauer, D., Richter, T. (2018) "Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest", Studies in Educational Evaluation, 59, pp. 168–178. https://doi.org/10.1016/j.stueduc.2018.07.009

Hornstein, H. A. (2017) "Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance", Cogent Education, 4(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., DeShon, R. P. (2012) "Detecting and Deterring Insufficient Effort Responding to Surveys", Journal of Business and Psychology, 27(1), pp. 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huck, S. W., Jacko, E. J. (1974) "Effect of Varying the Response Format of the Alpert-Haber Achievement Anxiety Test", Journal of Counseling Psychology, 21(2), pp. 159–163. https://doi.org/10.1037/h0036288

Lipsey, N., Shepperd, J. (2021) "Examining strategies to increase student evaluation of teaching completion rates", Assessment & Evaluation in Higher Education, 46(3), pp. 424–437. https://doi.org/10.1080/02602938.2020.1782343

Masters, J. R. (1974) "The Relationship between Number of Response Categories and Reliability of Likert-Type Questionnaires", Journal of Educational Measurement, 11(1), pp. 49–53. https://doi.org/10.1111/j.1745-3984.1974.tb00970.x

Morrison, K. (2012) "Searching for causality in the wrong places", International Journal of Social Research Methodology, 15(1), pp. 15–30. https://doi.org/10.1080/13645579.2011.594293

Morrison, K., van der Werf, G. (2016) "Searching for causality in educational research", Educational Research and Evaluation, 22(1–2), pp. 1–5. https://doi.org/10.1080/13803611.2016.1195081

Nadler, J. T., Weston, R., Voyles, E. C. (2015) "Stuck in the Middle: The Use and Interpretation of Mid-Points in Items on Questionnaires", The Journal of General Psychology, 142(2), pp. 71–89. https://doi.org/10.1080/00221309.2014.994590

Newstead, S. E., Arnold, J. (1989) "The Effect of Response Format on Ratings of Teaching", Educational and Psychological Measurement, 49(1), pp. 33–43. https://doi.org/10.1177/0013164489491004

Nowell, C. (2007) "The Impact of Relative Grade Expectations on Student Evaluation of Teaching", International Review of Economics Education, 6(2), pp. 42–56. https://doi.org/10.1016/S1477-3880(15)30104-3

Palmer, S., Smith, C. (2013) "Updating RIGs: including the systematic influence of online study on student evaluation of teaching", Educational Research and Evaluation, 19(1), pp. 77–90. https://doi.org/10.1080/13803611.2012.746712

Seldin, P. (1998) "How Colleges Evaluate Teaching: 1988 vs. 1998: Practices and Trends in the Evaluation of Faculty Performance", AAHE BULLETIN, 41(7), pp. 3–7.

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), pp. 598–642. https://doi.org/10.3102/0034654313496870

Spooren, P., Christiaens, W. (2017) "I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students'perceptions of a teaching evaluation process and their relationships with SET scores", Studies in Educational Evaluation, 54, pp. 43–49. https://doi.org/10.1016/j.stueduc.2016.12.003

Stark, P. B., Freishtat, R. (2014) "An evaluation of course evaluations", ScienceOpen Research, pp. 1–7. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Thielsch, M. T., Brinkmöller, B., Forthmann, B. (2018) "Reasons for responding in student evaluation of teaching", Studies in Educational Evaluation, 56, pp. 189–196. https://doi.org/10.1016/j.stueduc.2017.11.008

Tóth, Z. E., Andor, G., Árva, G., Árva, G. (2017) "Peer review of teaching at the Budapest University of Technology and Economics Faculty of Economic and Social Sciences", International Journal of Quality and Service Sciences, 9(3/4), pp. 402–424. https://doi.org/10.1108/IJQSS-02-2017-0014

Tóth, Z. E., Jónás, T., Bérces, R., Bedzsula, B. (2013) "Course evaluation by importance-performance analysis and improving actions at the Budapest University of Technology and Economics", International Journal of Quality and Service Sciences, 5(1), pp. 66–85. https://doi.org/10.1108/17566691311316257

Tsekouras, D. (2017) "The effect of rating scale design on extreme response tendency in consumer product ratings", International Journal of Electronic Commerce, 21(2), pp. 270–296. https://doi.org/10.1080/10864415.2016.1234290

Uttl, B., White, C. A., Gonzalez, D. W. (2017) "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related", Studies in Educational Evaluation, 54, pp. 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wang, G., Williamson A. (2022) "Course evaluation scores: valid measures for teaching effectiveness or rewards for lenient grading?", Teaching in Higher Education, 27(3), pp. 297–318. https://doi.org/10.1080/13562517.2020.1722992

Weems, G. H., Onwuegbuzie, A. J. (2001) "The Impact of Midpoint Responses and Reverse Coding on Survey Data", Measurement and Evaluation in Counseling and Development, 34(3), pp. 166–176. https://doi.org/10.1080/07481756.2002.12069033

Weijters, B., Cabooter, E., Schillewaert, N. (2010) "The effect of rating scale format on response styles: The number of response categories and response category labels", International Journal of Research in Marketing, 27(3), pp. 236–247. https://doi.org/10.1016/j.ijresmar.2010.02.004

Weng, L.-J. (2004) "Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability", Educational and Psychological Measurement, 64(6), pp. 956–972. https://doi.org/10.1177/0013164404268674

Wolbring, T., Riordan, P. (2016) "How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching", Social Science Research, 57, pp. 253–272. https://doi.org/10.1016/j.ssresearch.2015.12.009

Wouters, K., Maesschalck, J., Peeters, C. F. W., Roosen, M. (2014) "Methodological Issues in the Design of Online Surveys for Measuring Unethical Work Behavior: Recommendations on the Basis of a Split-Ballot Experiment", Journal of Business Ethics, 120(2), pp. 275–289. https://doi.org/10.1007/S10551-013-1659-5

Zipser, N., Mincieli, L. (2018) "Administrative and structural changes in student evaluations of teaching and their effects on overall instructor scores", Assessment & Evaluation in Higher Education, 43(6), pp. 995–1008. https://doi.org/10.1080/02602938.2018.1425368

Zumrawi, A. A., Bates, S. P., Schroeder, M. (2014) "What response rates are needed to make reliable inferences from student evaluations of teaching?", Educational Research and Evaluation, 20(7–8), pp. 557–563. https://doi.org/10.1080/13803611.2014.997915