Galkiewicz, Agata; Marcus, Jan; Siedler, Thomas

Working Paper

E-Learning at Universities: Does Starting with Difficult Questions Affect Student Performance?

IZA Discussion Papers, No. 17479

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Galkiewicz, Agata; Marcus, Jan; Siedler, Thomas (2024) : E-Learning at Universities: Does Starting with Difficult Questions Affect Student Performance?, IZA Discussion Papers, No. 17479, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/308338

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17479

E-Learning at Universities: Does Starting with Difficult Questions Affect Student Performance?

Agata Galkiewicz
Jan Marcus
Thomas Siedler

NOVEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17479

E-Learning at Universities: Does Starting with Difficult Questions Affect Student Performance?

Agata Galkiewicz
University of Potsdam and IAB Nürnberg

Jan Marcus
Freie Universität Berlin, Berlin School of Economics and IZA

Thomas Siedler
University of Potsdam, Berlin School of Economics and IZA

NOVEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

E-Learning at Universities: Does Starting with Difficult Questions Affect Student Performance?

To reduce cheating in written tests and exams, assessors often randomly vary the order of questions across students. However, little is known about the potential unintended side effects of question order. This paper examines whether randomizing students to start with an easier or harder question makes a difference to overall assessment performance in incentivized testing situations under time pressure. Using data from more than 8,000 online tests and exams administered in econometrics and statistics courses at two of Germany's largest universities, we find no evidence that the difficulty of the first question(s) has an effect on overall assessment performance. Our findings are good news for people designing (online) assessments, because randomizing the order of questions can be used as an effective tool to mitigate cheating, but does not affect students' overall performance.

JEL Classification: A22, I23

Keywords: education, university students, question order, randomization, e-learning, teaching of economics

Corresponding author:
Jan Marcus
Free University of Berlin
Kaiserswerther Str. 16-18
14195 Berlin
Germany
E-mail: jan.marcus@fu-berlin.de.

I. Introduction

In recent years, many universities worldwide shifted from traditional pen-and-paper exams to online testing as a means of assessment. This transition is likely to persist to a certain degree, as students' and professors' interest in online teaching and assessments remains high (Hill and LoPalo 2024). However, the shift to online assessments, where students are not directly monitored by exam supervisors, has raised concerns about potential cheating opportunities (Martinelli et al. 2018; Bilen and Matros 2021). Many lecturers and institutions have therefore adopted the practice of randomizing the ordering of questions to discourage cheating in assessments (Gruss and Clemons 2023). The randomization of question order is relevant not only in online testing, but also in tests and exams (pen-and-paper or computer) taken in the classroom, especially when students sit close together in lecture halls and can copy from their neighbors with relative ease.

A potential problem may arise here if the order of assessment questions directly impacts students' performance in the assessment. One particular concern is the negative impact of starting with a difficult question. It is feared that such an arrangement may increase stress and anxiety, undermine confidence, or lead to time mismanagement, as students might spend too much time on the challenging question, leaving insufficient time for the remaining questions. Some psychological studies, for example, find evidence of a retrospective bias: students were more likely to believe that they had answered more questions correctly if the answers were sorted from the easiest to the hardest compared to a randomized question order (Weinstein and Roediger 2010; Bard and Weinstein 2017).

In light of these concerns, this paper aims to improve our understanding of potential unintended side-consequences of question ordering in online, incentivized assessments under time pressure. In our randomized field experiments, we examine whether it makes a difference for the overall assessment performance if students have to start with a more difficult question. We make use of a unique dataset that includes detailed information about individual performance in 45 incentivized online single- and multiple-choice assessments. The main dataset covers 8,396 assessments (tests and exams) taken by undergraduate students at two large German universities in economics and statistics over the course of seven semesters from November 2020 to July 2024. Most importantly, the order of questions was randomized in each assessment, thereby allowing us to examine the effects of difficulty-based question order in 45 controlled, incentivized field experiments.[1]

We find little evidence that it makes a difference for overall assessment performance whether the first question is difficult or not. This finding holds across different definitions

[1]Our assessments are "incentivized" because passing the exams is required for successful completion of the bachelor's degree. Additionally, students who participated successfully in the tests earned bonus points that contributed to their final exam grade.

of question difficulty and a broad range of specifications. There is also no evidence that a difficult first question affects the distribution of our outcome variable or that it has a negative effect for certain subgroups. Moreover, we find no evidence that having two difficult questions at the beginning of the assessment has a significant impact on overall performance. The zero effect remains consistent during and after the COVID-19 pandemic and applies to both universities, supporting the generalizability of the results.

This paper seeks to contribute to several strands of literature. First, we aim to contribute to studies examining how teaching practice and (online) learning technologies in higher education influence students' academic performance.[2] Second, we seek to contribute to the literature on economic education research in general (Hoxby 2014; Allgood et al. 2015; Deming et al. 2015; Johnson and Meder 2024), and on how faculty members can use e-learning methods to complement traditional lectures and classes in particular. Third, we aim to add to studies on the potential side effects of measures to prevent cheating (Harmon and Lambrinos 2008; Swoboda and Feiler 2016; Martinelli et al. 2018; Lucifora and Tonello 2020; Bilen and Matros 2021; Cagala et al. 2024). Lastly, the most closely related body of literature we add to covers studies on the effects of difficulty-based question order on assessment performance, which has been a subject of interest in different academic disciplines. Overall, the findings regarding the effects of question difficulty order on students' performance are rather mixed, with a tendency toward more and more studies not finding a significant impact on assessment performance (Perlini et al. 1998; Weinstein and Roediger 2010; Sad 2020; Gruss and Clemons 2023). In contrast, earlier work by Hodson (1984) and Hambleton and Traub (1974) find that placing easy questions at the beginning results in improved test performance.[3] Recent work by Anaya et al. (2022) suggests that, in settings in which performance is not explicitly incentivized, ordering questions from the easiest to the most difficult results in the best scores and reduces the likelihood of premature dropouts among test takers, concluding that "order difficulty needs to be seriously considered" (Anaya et al. 2022: 11).

Our paper makes several specific contributions to the literature. First, many related studies are based on relatively small samples.[4] Second, while recent research estimates causal effects of difficulty-based question ordering by exploiting the randomization of

[2]See, for example, Brown and Liedholm (2002); Hernández-Julián and Peters (2012); Figlio et al. (2013); Banerjee and Duflo (2014); Alpert et al. (2016); Swoboda and Feiler (2016); Bettinger et al. (2017); Feld et al. (2019); Kofoed et al. (forthcoming); Eau et al. (2022); Elzinga and Harper (2023); De Paola et al. (2023); Hill and LoPalo (2024).

[3]There is also some evidence for test fatigue, meaning that test performance gets worse as the assessment progresses (see, e.g., Zamarro et al. 2019). However, this is less likely in our setting due to the relatively short duration of our assessments. Nevertheless, we also show that having a difficult question at the end does not affect overall performance.

[4]With the exception of recent work by Gruss and Clemons (2023) and Anaya et al. (2022).

question order across students, our knowledge about external validity is quite limited. We address these issues by presenting evidence from large-scale field experiments, conducted at two major universities in Germany. The assessments at the two universities are in different, but related subjects (statistics and econometrics). Third, we present evidence for incentivized assessments under time pressure. Fourth, we do not examine test performance in a laboratory setting, but our field experiments look at real-world performances.

The remainder of the paper is organized as follows. Section II describes the data and the empirical strategy. Section III presents the main findings, robustness checks and an analysis of effect heterogeneity. Section IV concludes.

II. Data and empirical strategy

A. Data

The dataset used for our main empirical analyses pertains to the results of online assessments (tests and exams) taken by students enrolled in bachelor's programs at two of the largest universities in Germany, the *University of Hamburg* (UHH) and the *Free University of Berlin* (FU). The UHH data includes information from the bachelor courses Applied Econometrics I and II. Both courses are mandatory for students studying for a bachelor's in Economics. This means that passing the final exams in these courses is a necessary requirement for them to receive their bachelor's degree. Our FU data covers two statistics courses: Introduction to Statistics (Statistics I) is a compulsory course for undergraduate students of economics and business administration. Therefore, students must pass the course to graduate with a bachelor's degree. Inferential Statistics (Statistics II) is also compulsory for economics majors and is taken by many business students.

Like most universities in the German higher education system, UHH and FU distinguish between summer and winter semesters.[5] Our UHH data covers a span of three semesters: winter semester 2020/21, summer semester 2021, and winter semester 2021/22, while the FU data cover four semesters: Statistics I in the summer semesters 2023 and 2024 as well as Statistics II in the winter semesters 2022/23 and 2023/24. Figure 1 provides a timeline of the courses and assessments. Since we have data for the years 2020–2024, we can identify causal effects both during and after the COVID-19 pandemic.

The dataset comprises the results of online assessments conducted through the e-learning platform mcEmpirics, which is mainly designed for (bachelor's) students and lecturers in Economics. The assessments included both true/false and multiple-choice questions. Students had to participate in these assessments at a fixed and scheduled time

[5]Teaching in the summer semester typically takes place between mid-April and mid-July. Teaching in the winter semester begins in mid-October and ends in mid-February.

but there was no fixed location, so they could take the tests and exams with their computers or tablets wherever they wanted.[6] At both universities, six tests were conducted each semester, except for the summer term of 2021 at UHH, which had only five tests. Additionally, in the winter term of 2020/2021 and the summer term of 2021 at UHH, two final exams were also offered per semester, with students required to take only one.[7] Participation in the tests was voluntary, but students could improve their final exam grade through successful participation in the tests. Each test consisted of 12–20 questions, while the final exams contained around 22–29 questions each.[8] In a given semester, there was essentially no overlap between assessments with respect to the questions asked.

In order to limit the amount of cheating, the assessments included two specific features. First, the order of the questions was randomized. In each assessment, each participating student received the same set of questions—but the order in which the questions appeared was completely randomized by a computer algorithm. Second, once a student moved on to the next question, they could not return to previous questions. Combined with the time constraint for completing the assessment, these two measures were designed to prevent students from exchanging their answers to the questions in the assessment. We consider these measures to be important in reducing the risk of cheating, because—in the present setting—the marginal return from cheating is likely to be high. For example, Harmon et al. (2010) show that randomizing the order of questions is one of the most highly rated tools for reducing cheating, while Harmon and Lambrinos (2008) argue that with randomly selected questions and time constraints, online exams without proctoring might not be any more prone to cheating than proctored exams. Further, Bilen and Matros (2021) point out that many universities require students to switch on their cameras for self-recording during online tests and exams to deter cheating. However, such regulations often violate students' privacy rights and online proctoring services are therefore not allowed at many universities, including UHH and FU.

Two other features of our setting are worth noting: First, students have an incentive to perform well on the assessments. Passing the exams is a necessary condition for the successful completion of the bachelor's degree and students received bonus points for good test results which counted toward their final exam grade. Second, the assessments are administered within a limited time window, which ensures that all students are equally

[6]The assessments took place at predetermined and previously announced times. For example, the first test in our sample took place on November 11, 2020, from 11.00 to 11.45 a.m..

[7]In each semester, students could chose between a first and a second exam date. If they failed the exam they sat on the first date, they could re-take it on the second date. If they failed the exam on the second date, the next retake exam would take place almost a year later. Nevertheless, students sometimes prefer the second exam date to avoid having too many exams in the same time period.

[8]It is important to note that, in line with most assessments in educational settings, wrong answers are not penalized with negative points in the performance evaluations (Weinstein and Roediger 2010).

affected by general external conditions (e.g., heat) during the assessment. Students were given 60 minutes to complete the exams, while for the tests they had 30–45 minutes, depending on the test and the number of questions.

All in all, our dataset contains information on 8,396 assessments taken, which originate from 45 tests (exams). The FU sample is larger with 6,583 individual assessments and 24 tests, compared to the UHH sample with 1,813 assessments taken that originate from 17 tests and four exams. Overall, the UHH sample includes 276 students and 335 different questions, while the FU sample contains information on 1,064 students and 448 different questions. Table A.1 in the Appendix shows how the number of students is distributed across universities and the different assessments.

B. Variables

Outcome. Our main outcome variable "total score", $Y_{a,i}$, measures the overall performance of individual i in assessment a and tells us i's proportion of correct answers in that assessment:

$$Y_{a,i} = \frac{\sum_{q=1}^{Q_a} \text{Correct}_{a,i,q}}{Q_a}, \tag{1}$$

where Q_a is the total number of questions in an assessment and $Correct_{a,i,q}$ is a dummy variable that takes the value one if individual i gave the correct answer to question q in assessment a, and zero otherwise. Overall, about 63.1% of the questions in the 8,396 assessments in the pooled sample were answered correctly, or $\bar{Y} = 0.631$.

Treatment. An advantage of our setting is that we do not need external assessors to determine the difficulty of a specific question. Instead, we measure the difficulty of a question in a data-driven way by relying on the proportion of incorrect answers to that question. More specifically, we compute $D_{a,i}$, the difficulty of the first question in assessment a for individual i, by looking at the answers to that question given by all other participants in the same assessment (thus excluding individual i in a leave-one-out fashion).

$$D_{a,i}^{cont} = \frac{\sum_{s \neq i} \text{Incorrect}_{a,q,s}}{n_{a,q} - 1}, \tag{2}$$

where $Incorrect_{a,q,s}$ is a binary variable that takes the value one if individual $s \neq i$ did not give the correct answer to question q in assessment a, and zero otherwise.[9] $n_{a,q}$ denotes the number of all individuals who gave an answer to question q in assessment a. Theoretically, $D_{a,i}^{cont}$ can range from 0 (no wrong answers) to 100% (only wrong answers).

[9] $Incorrect_{a,i,q}$ is the complementary event to $Correct_{a,i,q}$ from Equation 1, i.e. $Incorrect_{a,i,q} = 1 - Correct_{a,i,q}$.

In our setting, $D_{a,i}^{cont}$ is between 6.38% and 77.32% in 95% of cases.

As well as presenting results for the continuous treatment variable $D_{a,i}^{cont}$, we also do so for several binary treatment indicators. To construct the first binary treatment indicator ($D_{a,i}^{50\%}$), we split the continuous treatment variable $D_{a,i}^{cont}$ at its assessment-specific median to indicate whether the first question an individual faces is in the more difficult half of questions in assessment a, or not. The second binary treatment variable $D_{a,i}^{30\%}$ indicates whether $D_{a,i}^{cont}$ is in the top 30% of difficult questions in assessment a. Similarly, $D_{a,i}^{20\%}$ and $D_{a,i}^{10\%}$ indicate whether or not a question is in the top 20% or top 10% of difficult questions in an assessment.

The binary indicators compare groups that differ to a greater or lesser extent with respect to the difficulty of the first question, thus better capturing the possibility that small differences in question difficulty (as measured by a continuous treatment variable) may not have an impact. In addition, binary indicators have the advantage that they make it easier to compare the distribution of covariates between two groups, which allows us to check whether randomization was successful with respect to these covariates.

Further variables. For the UHH sample, we can use additional variables to ascertain whether the treatment and control groups are balanced, to control for them in our analyses, and to study effect heterogeneity. Our UHH data include information about whether a student used mcEmpirics to prepare for the assessment prior to the first test in a given semester. More specifically, students can use mcEmpirics individually for their studies by logging in at the platform homepage and by playing quizzes and answering questions. When using the platform on their own, students get immediate feedback after each answer (i.e., whether their answer was correct, a short explanation, and suggestions for further reading). We use the information on whether students studied on their own by means of the e-learning website mcEmpirics to create a proxy binary measure of motivation. This measure takes the value one if a student took at least one mcEmpirics quiz before the first assessment, and zero otherwise.[10] Further, we link the assessment data to administrative course records based on university email addresses. Since students were required to register for mcEmpirics using their university email address, we can link the administrative data perfectly to each mcEmpirics participant. The administrative data contains information about what the student is studying, the semester they are in, and their name. We use the administrative data mainly to construct control variables and to analyze whether randomization worked. Because of data protection issues, we do not merge the names with the assessment data, but we do construct a gender dummy based

[10] Quizzes contain ten single-choice and/or multiple-choice questions and allow students to familiarize themselves with mcEmpirics and the content of the course by providing an impression of question types and difficulty levels.

on the first names.

C. Empirical strategy

Our empirical strategy takes advantage of the fact that individuals are randomly assigned to first questions of different levels of difficulty. Thus, we estimate linear regressions of the following form

$$Y_{a,i} = \beta D_{a,i} + \gamma_a + \varepsilon_{a,i}, \tag{3}$$

where $Y_{a,i}$ (proportion of correct answers for individual i in assessment a) and $D_{a,i}$ (difficulty of the first question) are defined as above. γ_a denotes a set of 45 dummy variables for the 45 different assessments. We include the γ_a in all our specifications as the construction of the key variables is on the assessment level. This implicitly controls for day, subject, and university fixed effects. $\varepsilon_{a,i}$ denotes the error term. Since the error terms for the same individual may be correlated across assessments within the same semester, we present heteroscedasticity-robust standard errors clustered at the individual level.

From an identification point of view, one single assessment would be sufficient. However, we pool information across the 45 assessments to increase sample size and thus statistical power. We also present results separately for the UHH and FU samples, as well as for each of the 45 individual assessments.

While the regression in equation 3 focuses on average effects, we also examine differences across the entire distribution of assessment performance. To test for equality of distributions, we conduct two-sample Kolmogorov-Smirnov tests (see Massey 1951), reporting approximate asymptotic p-values.

Our empirical strategy strongly depends on the successful randomization of assessment takers. We provide two sets of evidence that the randomization was successful. First, Table A.2 in the Appendix shows for each assessment that none of the questions were more likely than the others to appear as the first question. More specifically, the table presents p-values for a chi-squared test with the null hypothesis that all questions in the given assessment are equally likely to appear as the first question.[11]

Second, for the UHH sample we have some background information about the assessment takers and we can examine whether they differ for individuals who start with a difficult first question. Table A.3 shows the means for these variables separately for the treatment and control groups as well as the difference between these means, based on the different definitions of the treatment variable (i.e., $D_{a,i}^{50\%}$, $D_{a,i}^{30\%}$,$D_{a,i}^{20\%}$, and $D_{a,i}^{10\%}$).

[11] The chi-squared test statistic is significant at the 5% level for five out of the 45 assessments. According to a binomial distribution with $p = 0.05$ and $N = 45$, we would expect, on average, 2.25 assessments to be significant at the 5% level. The probability of observing 5 or more significant tests is 7.3%.

Although our data set does not contain many background variables, it is reassuring to see that none of the mean differences are statistically significant from zero at the 5% level, indicating that the variables are balanced between the treatment and control groups.

III. Results

A. Main results

Table 1 presents our main results. Column 1 displays the estimated effect of a difficult first question on overall performance for the UHH sample, column 2 for the FU sample and column 3 for the pooled sample. Panel A (Median Split; $D_{a,i}^{50\%}$) defines a difficult first question as a question that is in the more difficult half of the questions in the respective assessment, and zero otherwise. The point estimate in panel A, column 1, suggests that in our UHH sample a difficult first question reduces the proportion of correct answers in the assessment by 0.4 percentage points. However, this effect is not statistically significant at conventional levels, indicating that starting with a difficult question has no effect on overall assessment performance. This result holds when using alternative definitions of a difficult first question (see panels B, C, and D in Table 1) and when examining the continuous difficulty variable (panel E). The point estimates are small, close to zero, and statistically insignificant. The pattern remains consistent across panels in the much larger FU sample and also in the pooled sample. None of the estimated effects is statistically significant and all point estimates are very small (i.e., ≤ 0.01). In the following, we focus on the pooled results; the Online Appendix presents all results separately for the UHH and FU samples. Notably, there is little difference between the results for the two samples.

While Table 1 focuses on average effects, we now turn to studying distributional treatment effects. To do this, we plot the full distribution of our outcome separately for the treatment and control groups. Figure 2 shows that the distribution of overall assessment performance for the treatment and control groups is very similar for all binary treatment indicators. Further, Kolmogorov-Smirnov tests for equality of distributions show that the distributions are not statistically significantly different at the 5% significance level. This supports the conclusion that starting with a difficult first question does not affect overall performance in the same assessment.

It may be that the experience of (not) having a difficult question at the beginning of an assessment affects participation and performance in subsequent assessments, even if it does not affect performance in the assessment at hand. To explore this further, we analyze whether a difficult first question affects the probability of taking the subsequent test, the overall performance in the subsequent test, and – for the UHH sample – the performance in the *final exam*. Table A.4 in the Appendix consistently suggests that a

difficult first question does not affect any of these additional outcomes, regardless of the measure of difficulty.

B. Robustness checks

Next, we examine the robustness of our findings. Our main specification uses standard errors that are clustered at the individual level to account for potential correlations between observations of the same individual across different assessments in the same semester. An alternative approach is to cluster the standard errors at the assessment level, as the treatment variable is defined at that level. This is what we do in column 1 of Table 2 for our pooled sample. In column 2, we cluster standard errors both at the individual and at the assessment level. However, these alternative approaches have little effect on the estimated standard errors in our setting, and all point estimates remain insignificant.

Column 3 in Table 2 takes advantage of the fact that we have multiple observations of the same individual, which allows us to include individual fixed effects. However, this is not our preferred specification, because individuals who never (or always) have a difficult first question do not contribute to the variation in our treatment variable. This is particularly relevant for individuals who only take one or two assessments, as well as for the results in panels B, C, and D, where only about 30%, 20%, and 10%, respectively, are defined as being treated in a given assessment. When individual fixed effects are included, a difficult first question has no significant effect on overall assessment performance.

While our main focus is on the difficulty of the *first* question, in columns 4 and 5 of Table 2, we also consider the *second* question. Specifically, we distinguish between three groups: one where both the first and the second question are easy, one where both the first and the second question are difficult, and one where one of the first two questions is easy and the other is difficult. The results in columns 4 and 5 are from the same regression, with students who received one easy and one difficult question to begin with serving as the reference group.[12] There is no evidence that starting with two difficult (or two easy) questions impacts the total score. An alternative way to capture the difficulty level of the first questions is by taking the average of correct answers for the first two (or three) questions and then calculating the treatment effect based on the median, 3rd decile split, etc. of this average (see Gruss and Clemons (2023) for a similar specification). For example, the treatment effect based on the median therefore captures whether the first two (or three) questions were, on average, in the upper half of the difficulty distribution, or not. The results of this additional sensitivity analysis are displayed in columns 6 and 7. The empirical results confirm that question-order difficulty has zero impact on students'

[12]Since the tests consist of 12–20 questions each, only one question can be in the first decile of difficulty. Therefore, we do not show results for panel D (1st decile split).

total score. Similarly, column 8 shows that *ending* with a difficult question does not affect the overall performance either.

While our binary treatment indicators divide the sample into two groups (easy versus difficult first question), Figure 3 shows the results for an alternative specification, in which the treatment variable can take three values: easy, medium, and difficult.[13] Again, there is no evidence that having a difficult (or easy) first question affects the total score.

Our binary treatment indicators use relative thresholds, meaning that they are based on a comparison of the difficulty level of a question relative to the other questions in the same assessment. In contrast, the next set of robustness checks builds on absolute thresholds instead. We now define a question as difficult if the proportion of fellow students who correctly answered the question is below a specific threshold (e.g., 50%) and as easy if this proportion is equal to or above that specific threshold. Again, we work with different thresholds. Figure 4 shows that when applying these absolute thresholds, our conclusion still does not change: a difficult first question has no negative effect on overall performance and does not result in a penalty.

C. Effect heterogeneity

We now examine whether the estimated average null effect masks important heterogeneity in the effect. First, we analyze whether the effect differs by course. Table 3 clearly shows that having a difficult first question has no effect on overall assessment performance in any of the four courses–Econometrics I, Econometrics II, Statistics I, or Statistics II. Second, we take a more granular approach by estimating the effect separately for each of the 45 assessments. Figure 5 displays the estimated effect sizes against their respective standard errors, similar to funnel plots used in meta-analysis. Several noteworthy patterns emerge from the figure: The estimated effect sizes are consistently small and symmetrically distributed around zero. Additionally, few point estimates are statistically significant at the 5% level, as indicated by the gray line. Given that we have 45 assessments, we would expect approximately 2.25 ($= 45 \cdot 0.05$) point estimates to be statistically significant at the 5% level under the null hypothesis of no effect. Figure 5 aligns well with this expectation. Overall, these findings reinforce the conclusion that starting with a difficult first question does not significantly affect overall performance.

Third, we look at whether the effect differs by individual-level background characteristics. This information is only available for the UHH sample and, hence, we exclusively use the UHH sample to analyze whether the effect differs by gender, between students

[13] More specifically, we estimate a regression similar to equation 3, where we replace the single treatment indicator with indicators for questions in the 1st and 3rd deciles of question difficulty in the respective assessment. Thus, medium difficulty is the reference category.

with higher and lower motivation, by type of assessment (tests versus exams), or across different cohorts. Table A.5 shows that for no group is the effect of a difficult first question statistically significant at the 5% level, regardless of the definition of a difficult first question.[14] In addition, Figure 6 shows the difference in treatment effects between groups for the different grouping variables.[15] It also shows that in each panel, none of the differences in the effects are statistically significant at the 5% level. Overall, there is no evidence that a difficult first question has a negative effect on the total score for any subgroup, which means that all group average treatment effects (GATEs) are zero.

IV. Conclusion

This paper addresses an often overlooked aspect of testing: how changing the order of questions in assessments might affect students' educational performance. While shuffling questions is a common practice to mitigate cheating at universities around the world, little attention has been paid to its potential unintended consequences. This study tackles a basic question: Does it make a difference whether students start with easy or hard questions?

To answer this, we use a unique dataset containing comprehensive individual performance data from several incentivized online binary-choice and multiple-choice assessments in statistics and econometrics courses based on the e-learning website mcEmpirics. Contrary to our expectations, our analysis finds no evidence that students who were randomly given a more difficult question at the start performed worse than those who started with an easier question. This finding remains consistent across different subgroups and universities as well as different definitions of question difficulty. Moreover, we find little evidence that the difficulty level of the first and second question matters for student performance, and we find also no evidence that the effect differs for assessments during and after the COVID-19 pandemic. Overall, our research suggests that starting with a difficult question has no discernible effect on subsequent test-taking behavior.

Compared to other settings, two elements of our field experiment made it more likely that a difficult first question would have a negative effect on overall performance. First, students could not go back to previous questions, so they did not have the option of skipping the first question if they found it difficult and returning to it later. Second, students were under time pressure, as they had to complete the assessment within a

[14] For the gender-specific analysis, we exclude 56 observations whose first names have an ambiguous gender assignment.

[15] The figure shows the point estimate (and its 95% confidence interval) for the interaction between a difficult first question and the grouping variable, based on a regression that includes interactions of the respective grouping variable with all other right-hand-side variables from equation 3.

relatively short period, which meant that they did not have much time to think about difficult questions. For these reasons, it seems likely that our findings would hold true in settings where students can go back to previous questions or take as much time as they want.

Our findings are good news for people who design (online) incentivized assessments under time pressure, because randomizing the order of questions can be used as an effective tool to mitigate cheating, but does not have unintended side effects in terms of affecting students' overall performance. The empirical results are also encouraging for people who have to take these assessments, because their overall performance is not affected by the luck or misfortune of having an easy or difficult question at the beginning of the test or exam (and thus not an element of chance). This is important, as online assessments with randomized question-ordering are a common practice in (higher) education when stakes are high. We therefore conclude that randomizing the order of questions is an inexpensive but effective tool that is unlikely to have unintended side effects on student's educational performance.

Conflict of Interest Statement

Thomas Siedler is the founder of the e-learning platform mcEmpirics. Although this affiliation facilitated access to the data analyzed in this paper, it had no impact on the outcomes or conclusions drawn from our research. Thomas Siedler and Jan Marcus taught the courses at the *University of Hamburg*, while Jan Marcus was responsible for the courses at *Free University Berlin*.

Data and Code Availability Statement:

Should our manuscript be accepted for publication, we will provide a replication package, including raw data and statistical software code.

References

Allgood, Sam, William B. Walstad, and John J. Siegfried, "Research on Teaching Economics to Undergraduates," *Journal of Economic Literature*, 2015, *53* (2), 285–325.

Alpert, William T., Kenneth A. Couch, and Oskar R. Harmon, "A Randomized Assessment of Online Learning," *American Economic Review*, May 2016, *106* (5), 378–382.

Anaya, Lina, Nagore Iriberri, Pedro Rey-Biel, and Gema Zamarro, "Understanding Performance in Test Taking: The Role of Question Difficulty Order," *Economics of Education Review*, October 2022, *90*, 102293.

Banerjee, Abhijit V. and Esther Duflo, "(Dis)organization and Success in an Economics MOOC," *American Economic Review*, May 2014, *104* (5), 514–518.

Bard, Gabriele and Yana Weinstein, "The Effect of Question Order on Evaluations of Test Performance: Can the Bias Dissolve?," *Quarterly Journal of Experimental Psychology*, October 2017, *70* (10), 2130–2140.

Bettinger, Eric P., Lindsay Fox, Susanna Loeb, and Eric S. Taylor, "Virtual Classrooms: How Online College Courses Affect Student Success," *American Economic Review*, September 2017, *107* (9), 2855–2875.

Bilen, Eren and Alexander Matros, "Online Cheating Amid COVID-19," *Journal of Economic Behavior & Organization*, February 2021, *182*, 196–211.

Brown, Byron W. and Carl E. Liedholm, "Can Web Courses Replace the Classroom in Principles of Microeconomics?," *American Economic Review*, May 2002, *92* (2), 444–448.

Cagala, Tobias, Ulrich Glogowsky, and Johannes Rincke, "Detecting and Preventing Cheating in Exams: Evidence from a Field Experiment," *Journal of Human Resources*, 2024, *59* (1), 210–241.

De Paola, Maria, Francesca Gioia, and Vincenzo Scoppa, "Online Teaching, Procrastination and Student Achievement," *Economics of Education Review*, June 2023, *94*, 102378.

Deming, David J., Claudia Goldin, Lawrence F. Katz, and Noam Yuchtman, "Can Online Learning Bend the Higher Education Cost Curve?," *American Economic Review*, May 2015, *105* (5), 496–501.

Eau, Grace, Derek Hoodin, and Tareena Musaddiq, "Testing the Effects of Adaptive Learning Courseware on Student Performance: An Experimental Approach," *Southern Economic Journal*, 2022, *88* (3), 1086–1118.

Elzinga, Kenneth G. and Daniel Q. Harper, "In-Person versus Online Instruction: Evidence from Principles of Economics," *Southern Economic Journal*, 2023, *90* (1), 3–30.

Feld, Jan, Nicolás Salamanca, and Ulf Zölitz, "Students are Almost as Effective as Professors in University Teaching," *Economics of Education Review*, December 2019, *73*, 101912.

Figlio, David, Mark Rush, and Lu Yin, "Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning," *Journal of Labor Economics*, October 2013, *31* (4), 763–784.

Gruss, Richard and Josh Clemons, "Does Question Order Matter on Online Math Assessments? A Big Data Analysis of Undergraduate Mathematics Final Exams," *Journal of Computer Assisted Learning*, 2023, *39* (5), 1539–1552.

Hambleton, Ronald K. and Ross E. Traub, "The Effects of Item Order on Test Performance and Stress," *The Journal of Experimental Education*, September 1974, *43* (1), 40–46.

Harmon, Oskar R. and James Lambrinos, "Are Online Exams an Invitation to Cheat?," *The Journal of Economic Education*, April 2008, *39* (2), 116–125.

Harmon, Oskar R, James Lambrinos, and Judy Buffolino, "Assessment Design and Cheating Risk in Online Instruction," *Online Journal of Distance Learning Administration*, 2010, *13* (3), 23–33.

Hernández-Julián, Rey and Christina Peters, "Targeting Teaching: Does the Medium Matter? Online versus Paper Coursework," *Southern Economic Journal*, 2012, *78* (4), 1333–1345.

Hill, Andrew J. and Melissa LoPalo, "The Effects of Online vs In-Class Testing in Moderate-Stakes College Environments," *Economics of Education Review*, February 2024, *98*, 102505.

Hodson, Derek, "The Effect of Changes in Item Sequence on Student Performance in a Multiple-Choice Chemistry Test," *Journal of Research in Science Teaching*, 1984, *21* (5), 489–495.

Hoxby, Caroline M., "The Economics of Online Postsecondary Education: MOOCs, Nonselective Education, and Highly Selective Education," *American Economic Review*, May 2014, *104* (5), 528–533.

Johnson, Marianne and Martin E. Meder, "Twenty-Three Years of Teaching Economics with Technology," *International Review of Economics Education*, March 2024, *45*, 100279.

Kofoed, Michael S., Lucas Gebhart, Dallas Gilmore, and Ryan Moschitto, "Zooming to Class?: Experimental Evidence on College Students' Online Learning during COVID-19," *American Economic Review: Insights*, forthcoming.

Lucifora, Claudio and Marco Tonello, "Monitoring and Sanctioning Cheating at School: What Works? Evidence from a National Evaluation Program," *Journal of Human Capital*, 2020, pp. 584–616.

Martinelli, César, Susan W. Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo, "Cheating and Incentives: Learning from a Policy Experiment," *American Economic Journal: Economic Policy*, February 2018, *10* (1), 298–325.

Massey, Frank J., "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, 1951, *46* (253), 68–78.

Perlini, Arthur H., David L. Lind, and Bruno D. Zumbo, "Context Effects on Examinations: The Effects of Time, Item Order and Item Difficulty," *Canadian Psychology / Psychologie canadienne*, 1998, *39* (4), 299–307.

Sad, Sueleyman Nihat, "Does Difficulty-Based Item Order Matter in Multiple-Choice Exams? (Empirical Evidence from University Students)," *Studies in Educational Evaluation*, March 2020, *64*, 100812.

Swoboda, Aaron and Lauren Feiler, "Measuring the Effect of Blended Learning: Evidence from a Selective Liberal Arts College," *American Economic Review*, May 2016, *106* (5), 368–372.

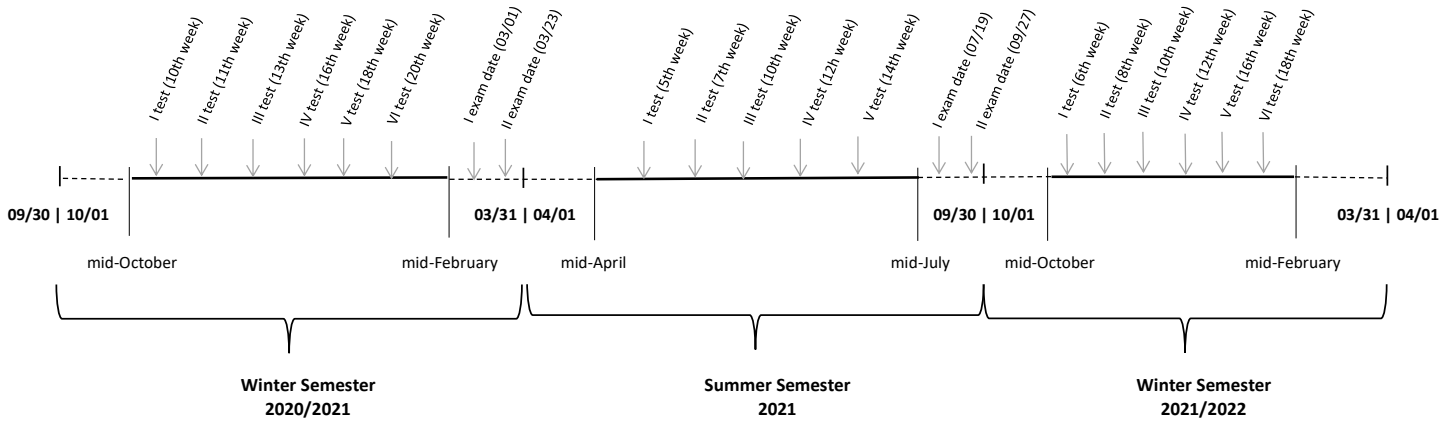
Weinstein, Yana and Henry L. Roediger, "Retrospective Bias in Test Performance: Providing Easy Items at the Beginning of a Test Makes Students Believe They Did Better on It," *Memory & Cognition*, April 2010, *38* (3), 366–376.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez, "When Students Don't Care: Reexamining International Differences in Achievement and Student Effort," *Journal of Human Capital*, 2019, *13* (4), 519–552.

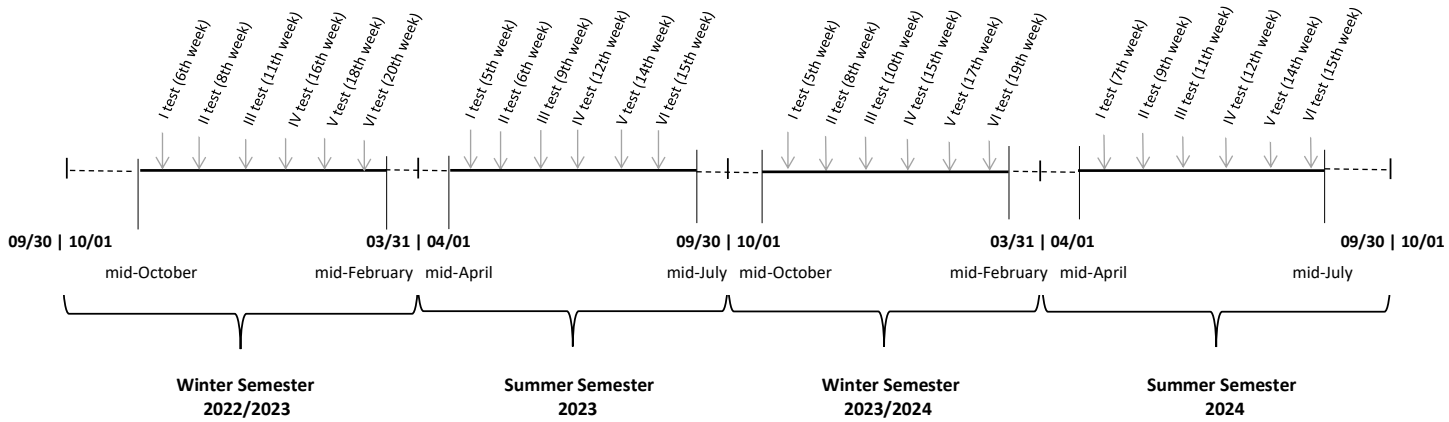
Figures and Tables

Figure 1: Timeline

(a) University of Hamburg



(b) Free University Berlin

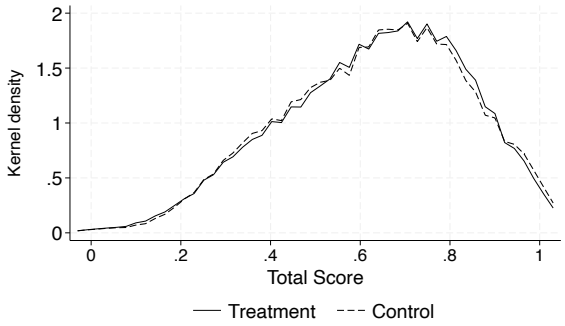


Notes: The figure shows the timeline of our study. In the German educational system, the winter semester starts on October 1 and ends on March 31, while the summer semester starts on April 1 and ends on September 30. The lecture period typically runs from mid-October to mid-February in the winter semester and from mid-April to mid-July in the summer semester. At the University of Hamburg, the data is collected over three consecutive semesters. In the winter semester 2020/21, six online tests and two exam dates were offered, in the summer semester 2021, five online tests and two exams, and in the winter semester 2021/22, six online tests and two exam dates were offered (as exams were not conducted online in this semester, they are not included in our study and are not shown in the figure). At the Free University Berlin, the data is collected over four consecutive semesters. In each semester, six online tests were conducted.

Figure 2: Distribution of Overall Assessment Performance by Treatment Status

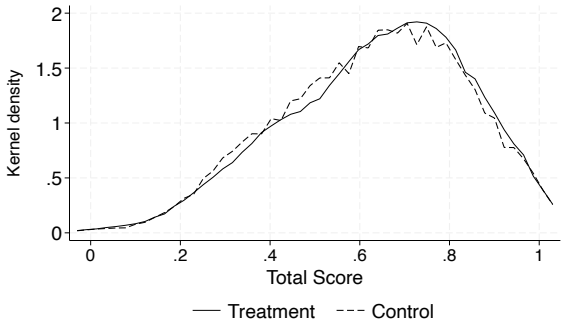
Panel A. Median Split ($D^{50\%}$)

(Kolmogorov-Smirnov test p-value: 0.914)



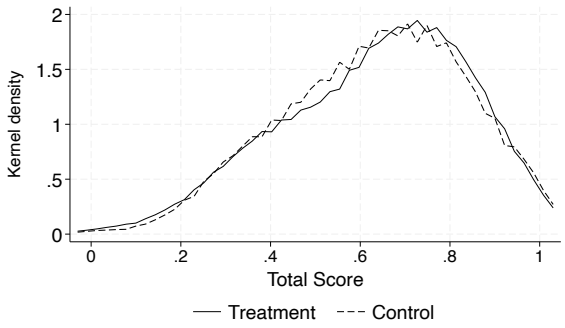
Panel B. 3rd Decile Split ($D^{30\%}$)

(Kolmogorov-Smirnov test p-value: 0.089)



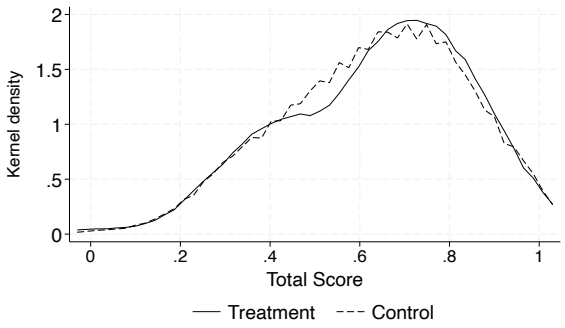
Panel C. 2nd Decile Split ($D^{20\%}$)

(Kolmogorov-Smirnov test p-value: 0.293)



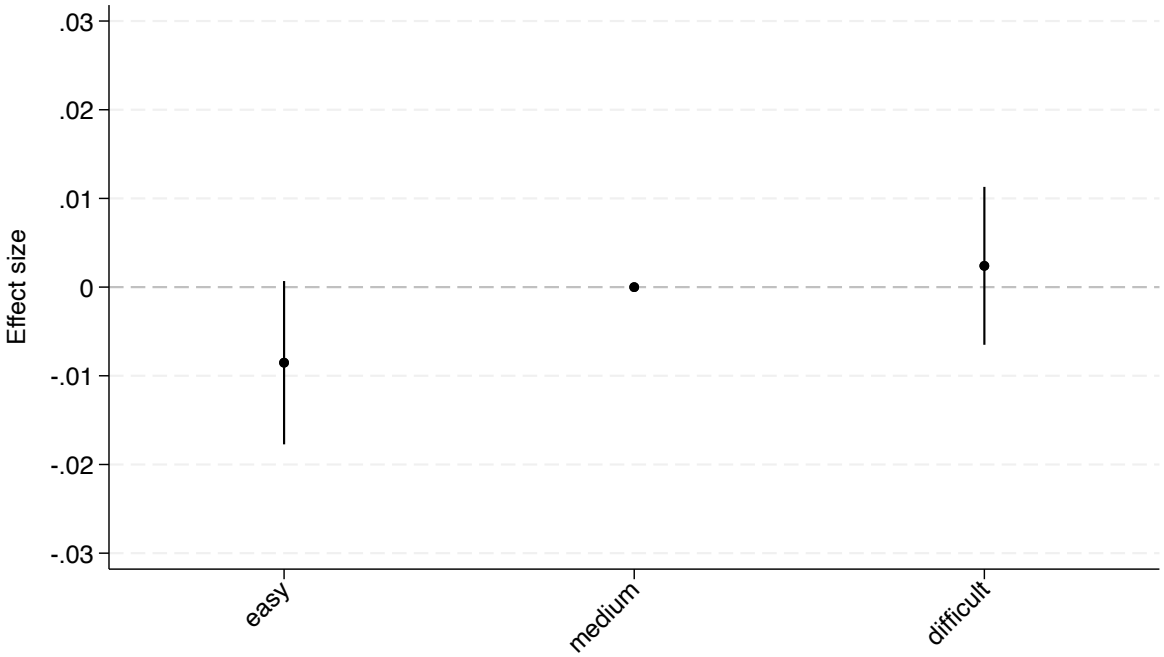
Panel D. 1st Decile Split ($D^{10\%}$)

(Kolmogorov-Smirnov test p-value: 0.338)



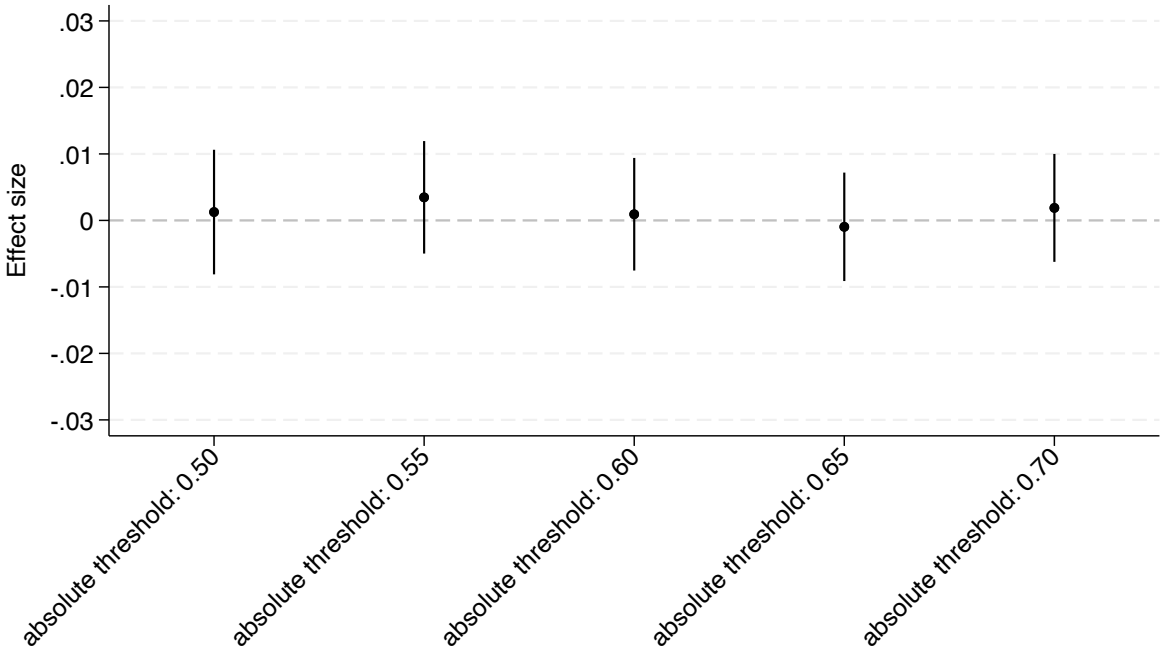
Notes: The figure displays kernel density of the total score (measured by the share of correct answers in the assessment) for treatment (solid line) and control (dashed line) groups. In panel A, the treatment indicator "difficult question" is based on the median value of the difficulty-based question distribution, in panel B on the 3rd decile value, in panel C on the 2nd decile value, and in panel D on the 1st decile value. The reported p-values are the approximate asymptotic p-values from the two-sample Kolmogorov-Smirnov tests.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure 3: Robustness Checks with Three Question Types



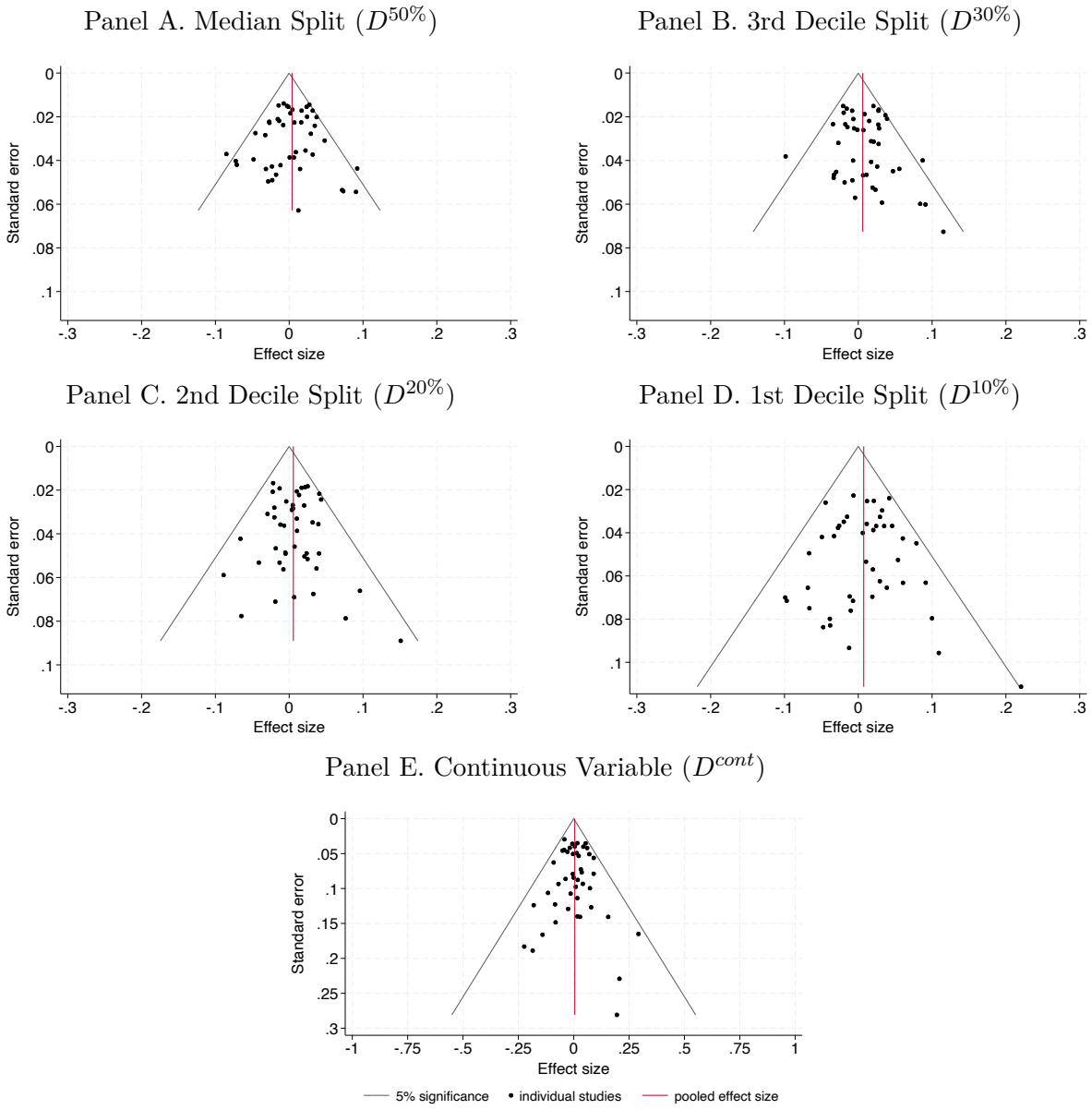
Notes: The figure presents point estimates (and 95% confidence intervals) from regressions of assessment performance (measured as the proportion of correct answers in the assessment) on an alternative treatment definition. This alternative treatment distinguishes between three question types instead of two (first question easy and first question difficult). First questions are defined as easy if the proportion of correct answers given by fellow students is higher than the 66th percentile of the distribution of these proportions for a given assessment, and are defined as difficult if this proportion is lower than the 33rd percentile (see Section B for more details). The remaining questions are defined as being of medium difficulty and form the baseline group. The number of observations is equal to 8,396.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure 4: Robustness Checks with Absolute Thresholds



Notes: The figure displays point estimates (and 95% confidence intervals) from regressions of assessment performance (measured by the share of correct answers in the assessment) on an alternative definition of the treatment variable. This alternative treatment is defined based on an absolute threshold. The first questions are defined as easy if the proportion of correct answers, given by fellow students in a given assessment, is higher than or equal to an absolute threshold, and are defined as difficult if this proportion is smaller than the threshold. The number of observations is 8,396.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

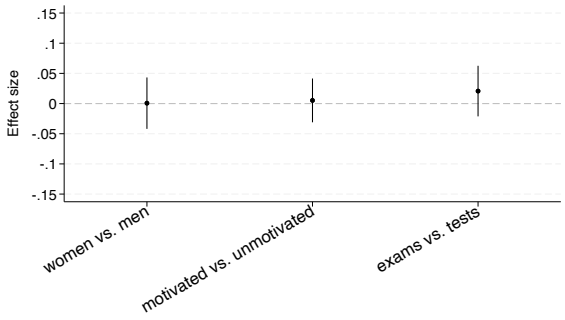
Figure 5: The Effect of a Difficult First Question on Test Performance Separately for Each Assessment



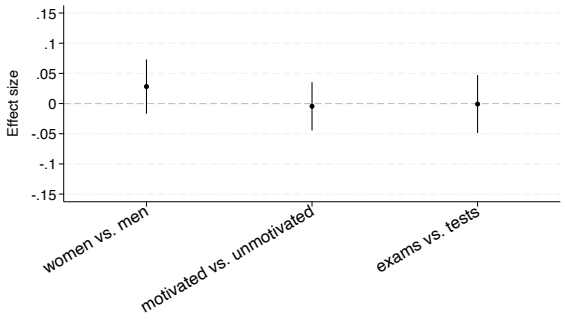
Notes: The figure displays the effect of a difficult first question on the total score separately for each of the 45 assessments. The black dots represents the estimated coefficients from these 45 regressions. The grey lines mark the 95% confidence level. The red line indicates the pooled effect size. In panel A, the treatment indicator "difficult question" is based on the median value of the difficulty-based question distribution, in panel B on the 3rd decile, in panel C on the 2nd decile, and in panel D on the 1st decile value. In panel E, the treatment variable is a continuous measure.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure 6: Heterogeneity: The Effect of a Difficult First Question
on Overall Test Performance
University of Hamburg

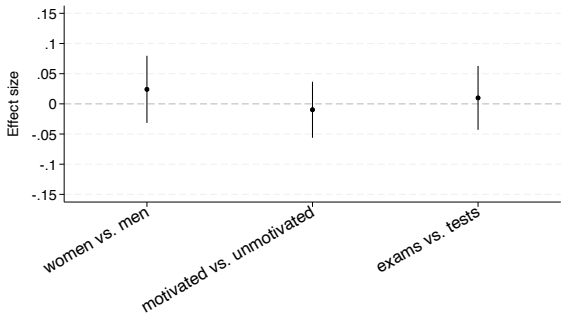
Panel A. Median Split ($D^{50\%}$)



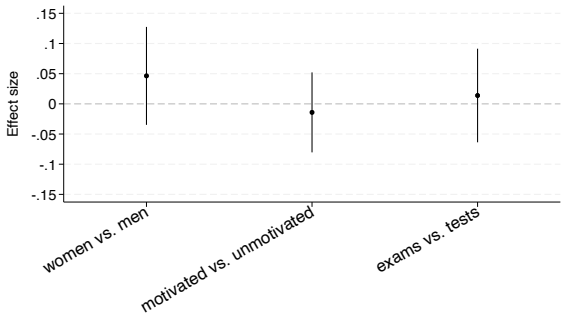
Panel B. 3rd Decile Split ($D^{30\%}$)



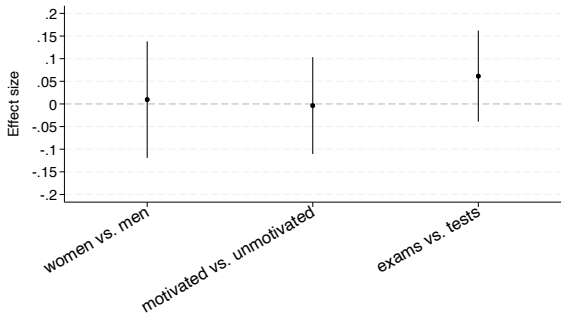
Panel C. 2nd Decile Split ($D^{20\%}$)



Panel D. 1st Decile Split ($D^{10\%}$)



Panel E. Continuous Variable (D^{cont})



Notes: The figure displays heterogeneity in the effect of a difficult first question on the total score. The point estimates are the interaction terms between different grouping variables and the treatment variable. In panel A, the treatment indicator "difficult question" is based on the median value of the difficulty-based question distribution, in panel B on the 3rd decile, in panel C on the 2nd decile, and in panel D on the 1st decile value. In panel E, the treatment variable is a continuous measure.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table 1: The Effect of a Difficult First Question on Assessment Performance

	University of Hamburg	Free University Berlin	Pooled Sample
	(1)	(2)	(3)
Panel A: Median Split ($D^{50\%}$)			
Difficult question	-0.004	0.006	0.004
	(0.009)	(0.004)	(0.004)
R-squared	0.162	0.340	0.294
Panel B: 3rd Decile Split ($D^{30\%}$)			
Difficult question	0.015	0.005	0.007
	(0.010)	(0.005)	(0.004)
R-squared	0.163	0.340	0.295
Panel C: 2nd Decile Split ($D^{20\%}$)			
Difficult question	0.004	0.006	0.006
	(0.012)	(0.005)	(0.005)
R-squared	0.162	0.340	0.294
Panel D: 1st Decile Split ($D^{10\%}$)			
Difficult question	0.008	0.010	0.010
	(0.017)	(0.007)	(0.007)
R-squared	0.162	0.340	0.295
Panel E: Continuous Variable (D^{cont})			
Difficult question	0.004	0.006	0.005
	(0.027)	(0.010)	(0.010)
R-squared	0.162	0.340	0.294
Outcome mean	0.644	0.628	0.631
Number of observations	1,813	6,583	8,396

Notes: The table presents the estimated effect of having a difficult question first on overall assessment performance (measured by the share of correct answers in the assessment). Column 1 only includes observations from the University of Hamburg, column 2 only observations from Free University Berlin. Column 3 shows the estimates for the pooled sample. All specifications include a binary indicator for each assessment. Different definitions of the treatment variable "difficult question" are presented in panels A–E. Robust standard errors (displayed in parentheses) allow for clustering at the individual level (** $p<0.01$, * $p<0.05$).
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table 2: Robustness Checks: The Effect of a Difficult Question(s) on Overall Assessment Performance

				First Two Questions		Average Difficulty		Last Question
	SE Clustering at Test Level	Two-way Clustering	Ind. FE	Two Easy Questions	Two Difficult Questions	First Two Questions	First Three Questions	Difficult Questions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Median Split ($D^{50\%}$)								
Treatment effect	0.004	0.004	0.000	-0.002	-0.001	0.004	0.004	0.004
	(0.004)	(0.004)	(0.003)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Share of observations				25%	22%	50%	50%	49%
Panel B: 3rd Decile Split ($D^{30\%}$)								
Treatment effect	0.007	0.007	0.004	-0.009*	-0.008	-0.001	0.006	0.006
	(0.004)	(0.004)	(0.003)	(0.004)	(0.007)	(0.004)	(0.005)	(0.004)
Share of observations				50%	7%	70%	70%	29%
Panel C: 2nd Decile Split ($D^{20\%}$)								
Treatment effect	0.006	0.006	0.004	-0.006	-0.013	-0.000	0.007	0.008
	(0.004)	(0.004)	(0.004)	(0.004)	(0.011)	(0.005)	(0.005)	(0.005)
Share of observations				66%	3%	80%	80%	19%
Panel D: 1st Decile Split ($D^{10\%}$)								
Treatment effect	0.010	0.010	0.009			0.006	0.007	0.013*
	(0.007)	(0.007)	(0.006)			(0.007)	(0.007)	(0.007)
Share of observations						90%	90%	9%
Panel E: Continuous Variable (D^{cont})								
Treatment effect	0.005	0.005	0.001			0.002	0.027	0.009
	(0.008)	(0.008)	(0.008)			(0.014)	(0.017)	(0.010)

Notes: In columns 1–3, results of the regression of the overall assessment performance (measured by the share of correct answers in the assessment) on the treatment variable "difficult question" are displayed. In column 1, standard errors are clustered at the assessment level instead of the individual level. In column 2, standard errors are clustered both at the assessment and individual level (two-way clustering). In column 3, individual fixed effects are added. In columns 4 and 5, the difficulty-based question order of the first two questions are considered; column 4 shows the effect of having two easy questions first and column 5 of having two difficult questions first. The baseline group in this specification consists of students who got one easy and one difficult question among the first two (irrespective of their order). Columns 6–7 are based on the average difficulty level of the first two and three questions, respectively. In column 8, the effect of the last question being difficult is estimated. Different definitions of the treatment variable "difficult question" are presented in panels A–E. All estimations are based on 8,396 observations and the sample mean of the outcome variable is 0.631. Robust standard errors (in parentheses) allow for clustering at the individual level. ** p<0.01, * p<0.05.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table 3: The Effect of a Difficult First Question on Assessment Performance by Course

	University of Hamburg		Free University Berlin	
	Econ. I	Econ. II	Statistics I	Statistics II
	(1)	(2)	(3)	(4)
Panel A: Median Split ($D^{50\%}$)				
Difficult question	0.006	-0.022	0.008	0.001
	(0.012)	(0.016)	(0.005)	(0.007)
Panel B: 3rd Decile Split ($D^{30\%}$)				
Difficult question	0.015	0.017	0.005	0.007
	(0.012)	(0.017)	(0.005)	(0.008)
Panel C: 2nd Decile Split ($D^{20\%}$)				
Difficult question	0.008	-0.004	0.004	0.011
	(0.015)	(0.021)	(0.007)	(0.009)
Panel D: 1st Decile Split ($D^{10\%}$)				
Difficult question	0.017	-0.013	0.010	0.010
	(0.020)	(0.032)	(0.009)	(0.012)
Panel E: Continuous Variable (D^{cont})				
Difficult question	0.011	-0.024	0.000	0.018
	(0.030)	(0.057)	(0.012)	(0.018)
Outcome mean	0.614	0.699	0.611	0.663
Number of observations	1,178	635	4,419	2,164

Notes: The table presents the estimated effect of having a difficult question first on overall assessment performance (measured by the share of correct answers in the assessment). Columns 1–2 includes observations from the University of Hamburg. In column 1 only the course Econometrics I is considered and in column 2 only the course Econometrics II. Columns 3–4 include observations from Free University Berlin. In column 3 only the course Statistics I is considered and in column 4 only the course Statistics II. All specifications include a binary indicator for each assessment. Different definitions of the treatment variable "difficult question" are presented in panels A–E. Robust standard errors (displayed in parentheses) allow for clustering at the individual level (** $p<0.01$, * $p<0.05$).
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

A. Appendix

Table A.1: Number of Students by University, Semester, and Assessment

	University of Hamburg			Free University Berlin			
	Econometrics I Winter Sem. 20/21	Econometrics II Summer Sem. 21	Econometrics I Winter Sem. 21/22	Statistics II Winter Sem. 22/23	Statistics I Summer Sem. 23	Statistics II Winter Sem. 23/24	Statistics I Summer Sem. 24
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Tests During Instruction Period							
Test I	98	117	98	135	347	179	403
Test II	95	104	111	166	377	230	438
Test III	80	93	99	166	325	234	421
Test IV	79	94	90	150	321	196	399
Test V	72	72	84	160	311	206	409
Test VI	58		64	157	272	185	396
Panel B: Final Exams							
Date I	73	95					
Date II	77	60					
$N_{assessments}$:	632	635	546	934	1,953	1,230	2,466
$N_{students}$:	149	153	125	189	421	261	480

Notes: The table presents the number of students participating in the assessments, the total number of assessments, and the number of students per semester.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table A.2: Chi-squared Test That All Questions Are Equally Likely to Be the First

	University of Hamburg			Free University Berlin			
	Econometrics I Winter Sem. 20/21	Econometrics II Summer Sem. 21	Econometrics I Winter Sem. 21/22	Statistics II Winter Sem. 22/23	Statistics I Summer Sem. 23	Statistics II Winter Sem. 23/24	Statistics I Summer Sem. 24
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Tests During Instruction Period							
Test I	0.426	0.573	0.595	0.026	0.273	0.770	0.046
Test II	0.883	0.469	0.155	0.211	0.089	0.133	0.717
Test III	0.190	0.032	0.230	0.454	0.224	0.260	0.969
Test IV	0.185	0.318	0.736	0.256	0.550	0.554	0.038
Test V	0.944	0.810	0.215	0.334	0.334	0.270	0.531
Test VI	0.249		0.897	0.074	0.013	0.649	0.520
Panel B: Final Exams							
Date I	0.385	0.997					
Date II	0.832	0.466					

Notes: The table presents p-values for chi-squared tests with the null hypothesis that all questions in the given assessment are equally likely to appear as the first question.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table A.3: Balance Test

University of Hamburg

	Treatment (1)	Control (2)	Difference (1)-(2)
Panel A: Median Split ($D^{50\%}$)			
Total score	0.639	0.648	-0.010
Female	0.341	0.337	0.004
Semester	4.808	5.005	-0.197
Economics program	0.952	0.950	0.003
Motivated student	0.611	0.575	0.036
Number of observations	840 (46%)	973 (54%)	
Panel B: 3rd Decile Split ($D^{30\%}$)			
Total score	0.656	0.639	0.017
Female	0.339	0.338	0.001
Semester	4.885	4.925	-0.041
Economics program	0.959	0.948	0.011
Motivated student	0.617	0.581	0.036
Number of observations	512 (28%)	1,301 (72%)	
Panel C: 2nd Decile Split ($D^{20\%}$)			
Total score	0.652	0.642	0.010
Female	0.346	0.337	0.009
Semester	4.828	4.934	-0.106
Economics program	0.958	0.949	0.009
Motivated student	0.611	0.587	0.025
Number of observations	337 (19%)	1,476 (81%)	
Panel D: 1st Decile Split ($D^{10\%}$)			
Total score	0.643	0.644	-0.001
Female	0.312	0.341	-0.030
Semester	4.929	4.912	0.016
Economics program	0.946	0.951	-0.005
Motivated student	0.619	0.588	0.031
Number of observations	168 (9%)	1,645 (91%)	

Notes: The outcome variable "total score" measures the share of correct answers in the assessment (test or exam). "Female" is a binary variable equal to one if a student is is a female, and zero otherwise. In case of 56 observations where information on gender is missing, the sample mean is imputed. "Semester" is a count variable indicating the semester in which a student is enrolled. "Economics program" is a binary variable equal to one if a student's major is economics, and zero otherwise. "Motivation" is equal to one if a student played quizzes before the first formal assessment, and zero otherwise. If at least one quiz had been played, a student is defined as highly-motivated, if no quizzes had been played, a student is defined as having low motivation. Different definitions of the treatment variable "difficult first question" are presented in panels A–D, separately. The total number of observations is 1,813 assessment results. ** $p<0.01$, * $p<0.05$.

Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table A.4: The Effect of a Difficult First Question on Other Educational Outcomes

	Next Test Participation	Next Test Result	Final Exam Result					
			Pooled	1st Test	2nd Test	3rd Test	4th Test	5th Test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Median Split ($D^{50\%}$)								
Difficult question	0.003	0.008	-0.014	0.005	-0.024	-0.003	-0.054*	-0.018
	(0.008)	(0.004)	(0.011)	(0.027)	(0.026)	(0.028)	(0.027)	(0.031)
Panel B: 3rd Decile Split ($D^{30\%}$)								
Difficult question	-0.005	0.006	0.003	0.008	0.019	-0.007	-0.008	-0.029
	(0.009)	(0.005)	(0.012)	(0.030)	(0.029)	(0.031)	(0.029)	(0.033)
Panel C: 2nd Decile Split ($D^{20\%}$)								
Difficult question	-0.007	0.009	-0.012	-0.009	-0.008	-0.004	-0.037	0.003
	(0.010)	(0.006)	(0.016)	(0.047)	(0.035)	(0.034)	(0.030)	(0.039)
Panel D: 1st Decile Split ($D^{10\%}$)								
Difficult question	-0.001	0.007	-0.045*	-0.010	-0.067	-0.057	-0.056	-0.045
	(0.014)	(0.008)	(0.022)	(0.052)	(0.038)	(0.048)	(0.038)	(0.042)
Panel E: Continuous Variable (D^{cont})								
Difficult question	0.011	0.019	-0.064	-0.018	-0.144	-0.041	-0.153	-0.055
	(0.020)	(0.012)	(0.036)	(0.091)	(0.081)	(0.102)	(0.094)	(0.079)
Mean of Outcome	0.879	0.650	0.613	0.602	0.604	0.615	0.618	0.621
Number of Observations	6,887	6,051	915	193	188	166	169	141

Notes: The table presents the estimated effect of starting with a difficult question in a given test on the probability of participating in the subsequent test (column 1), on the result of the subsequent test (column 2), and on the final exam result (columns 3–8). In columns 4–8, the effect of a treatment variable being switched on (having a difficult first question) in a given test is considered while in column 3 all tests are pooled together. In columns 1–2 the data from both the University of Hamburg and Free University Berlin is utilized. In the columns 3–8 only the data from the University or Hamburg is used, as final exams are not observed for the other university. Different definitions of the treatment variable "difficult question" are presented in panels A–E. The number of observations in columns 1 and 2 is lower because we have to exclude the exams and final tests of each semester. The analyses in columns 3 to 8 include only students from the UHH sample in the winter semester 2020/21 and summer semester 2021, as these were the only semesters in which the exam was administered through mcEmpirics. Robust standard errors (in parentheses) allow for clustering at the individual level (** p<0.01, * p<0.5).
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table A.5: Heterogeneity: The Effect of a Difficult First Question
on Overall Assessment Performance
University of Hamburg

	Gender		Motivation		Assessment	
	Females	Males	High	Low	Tests	Exams
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Median Split ($D^{50\%}$)						
Difficult question	-0.004	-0.004	-0.007	-0.012	-0.007	0.018
	(0.019)	(0.011)	(0.011)	(0.015)	(0.010)	(0.021)
Panel B: 3rd Decile Split ($D^{30\%}$)						
Difficult question	0.034	0.006	0.010	0.014	0.016	0.015
	(0.019)	(0.012)	(0.012)	(0.016)	(0.011)	(0.023)
Panel C: 2nd Decile Split ($D^{20\%}$)						
Difficult question	0.019	-0.005	-0.002	0.007	0.003	0.017
	(0.024)	(0.015)	(0.015)	(0.018)	(0.014)	(0.025)
Panel D: 1st Decile Split ($D^{10\%}$)						
Difficult question	0.038	-0.008	-0.002	0.012	0.006	0.030
	(0.037)	(0.019)	(0.020)	(0.027)	(0.020)	(0.035)
Panel E: Continuous Variable (D^{cont})						
Difficult question	0.010	0.000	-0.001	0.003	-0.012	0.050
	(0.058)	(0.031)	(0.031)	(0.045)	(0.033)	(0.041)
Mean of outcome	0.637	0.649	0.692	0.575	0.661	0.559
Number of observations	595	1,162	1,072	741	1,508	305

Notes: The table presents the heterogeneous effects of a difficult first question on the total score (measured by the share of correct answers in the assessment). In columns 3–4, students are split by their motivation into highly-motivated (column 3) and low motivation (column 4). See also notes to Table A.3. In columns 5–6, the assessments are split into online tests performed during the semester (column 5) and final exams (column 6). Different definitions of the treatment variable "difficult question" are presented in panels A-E. Robust standard errors (presented in parentheses) allow for clustering at the individual level (** p<0.01, * p<0.05).
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

E-Learning at Universities: Does Starting with Difficult Questions Affect Student Performance?

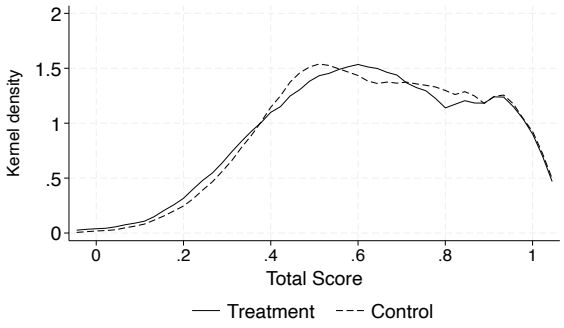
– Online Appendix –

Agata Galkiewicz, Jan Marcus, and Thomas Siedler

Figure O.1: Distribution of Overall Assessment Performance by Treatment Status
University of Hamburg

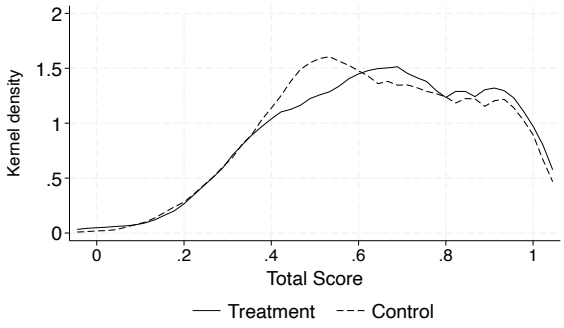
Panel A. Median Split ($D^{50\%}$)

(Kolmogorov-Smirnov test p-value: 0.855)



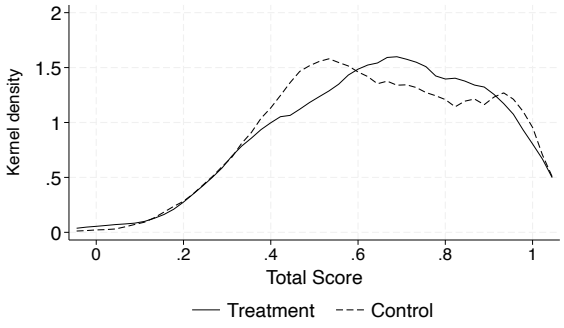
Panel B. 3rd Decile Split ($D^{30\%}$)

(Kolmogorov-Smirnov test p-value: 0.074)



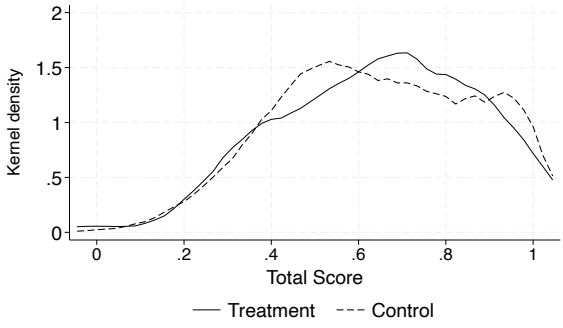
Panel C. 2nd Decile Split ($D^{20\%}$)

(Kolmogorov-Smirnov test p-value: 0.137)



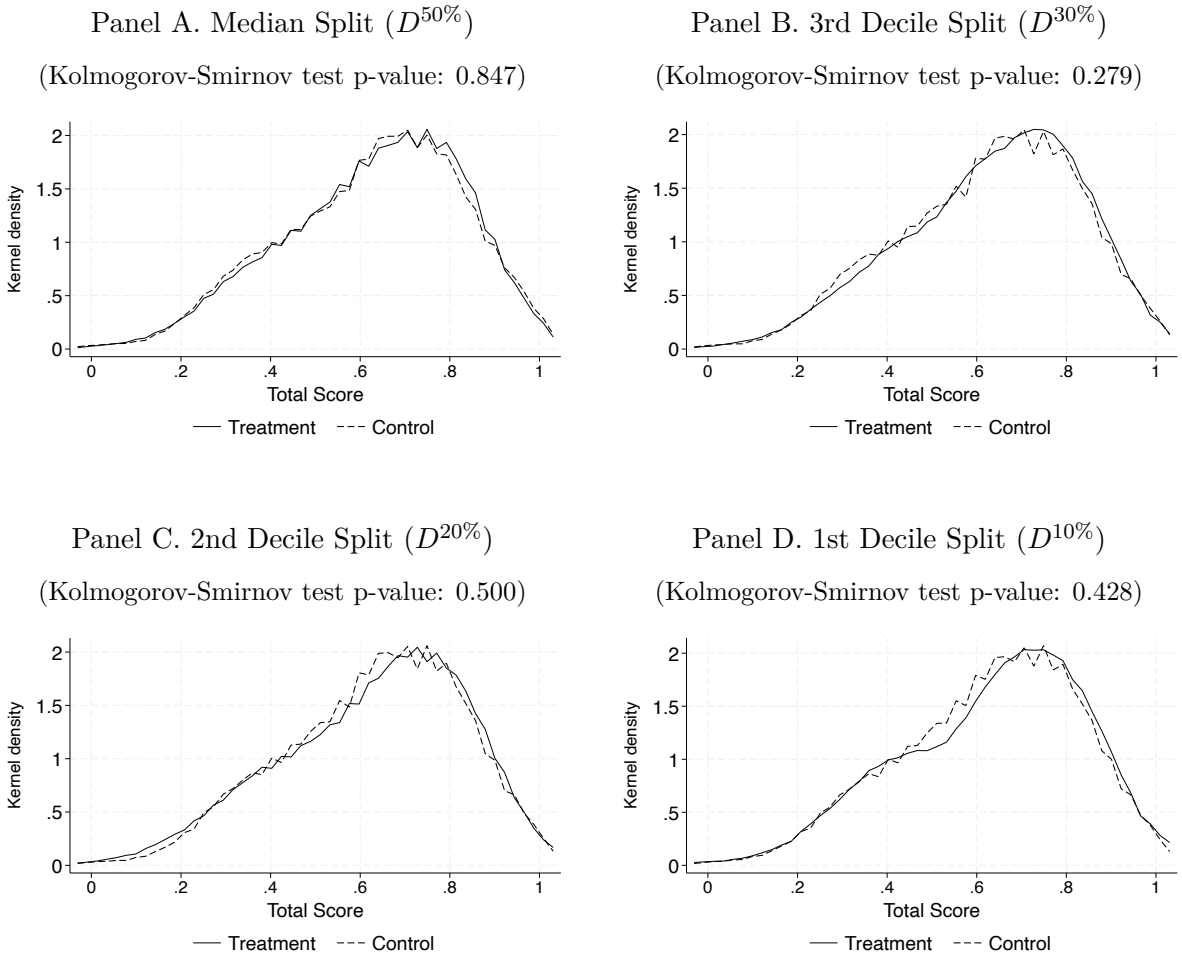
Panel D. 1st Decile Split ($D^{10\%}$)

(Kolmogorov-Smirnov test p-value: 0.608)



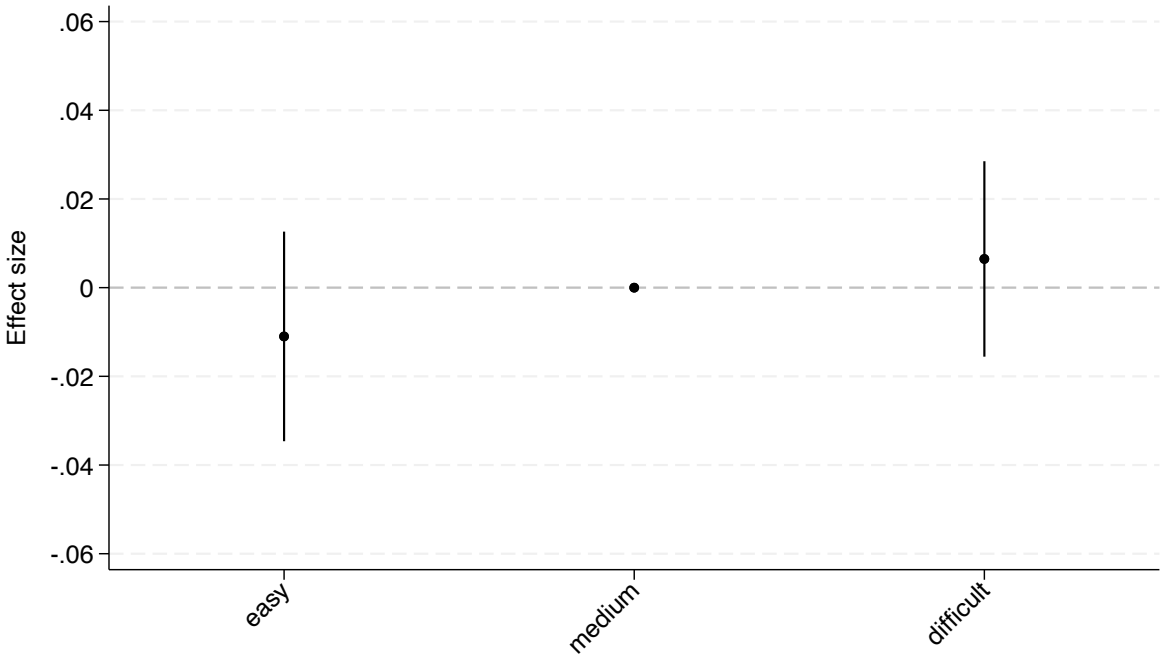
Notes: The figure displays kernel density of the total score (measured by the share of correct answers in the assessment) for treatment (solid line) and control (dashed line) groups. In panel A, the treatment indicator "difficult question" is based on the median value of the difficulty-based question distribution, in panel B on the 3rd decile value, in panel C on the 2nd decile value, and in panel D on the 1st decile value. The reported p-values are the approximate asymptotic p-values from the two-sample Kolmogorov-Smirnov tests.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure O.2: Distribution of Overall Assessment Performance by Treatment Status
Free University of Berlin



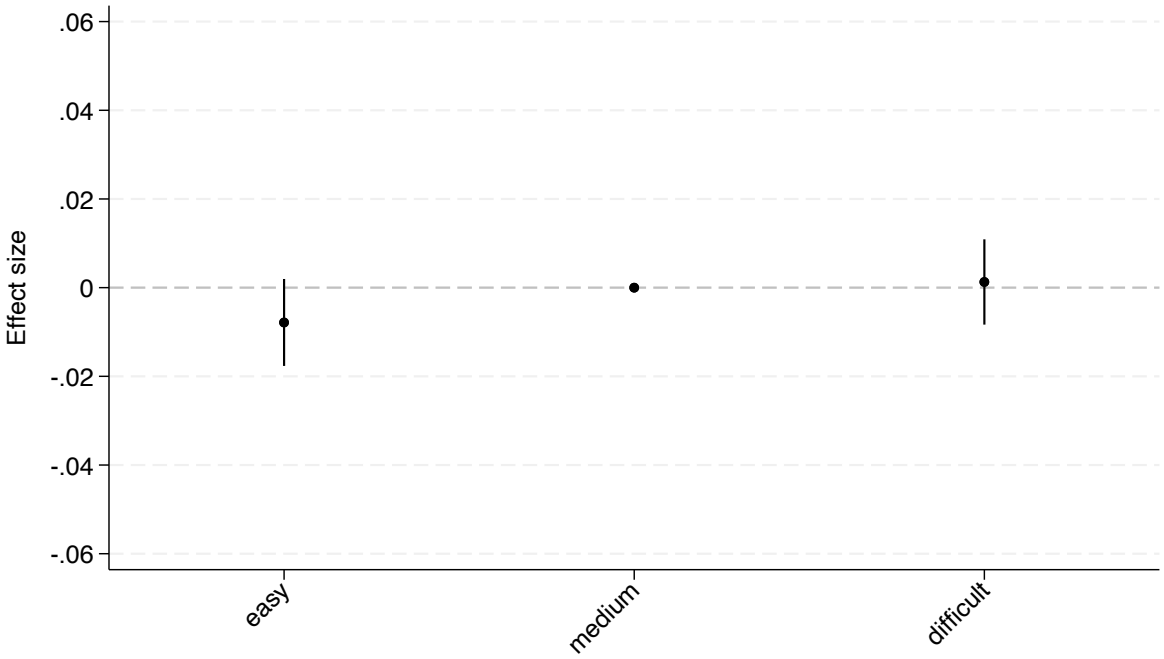
Notes: The figure displays kernel density of the total score (measured by the share of correct answers in the assessment) for treatment (solid line) and control (dashed line) groups. In panel A, the treatment indicator "difficult question" is based on the median value of the difficulty-based question distribution, in panel B on the 3rd decile value, in panel C on the 2nd decile value, and in panel D on the 1st decile value. The reported p-values are the approximate asymptotic p-values from the two-sample Kolmogorov-Smirnov tests.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure O.3: Robustness Checks with Three Question Types
University of Hamburg



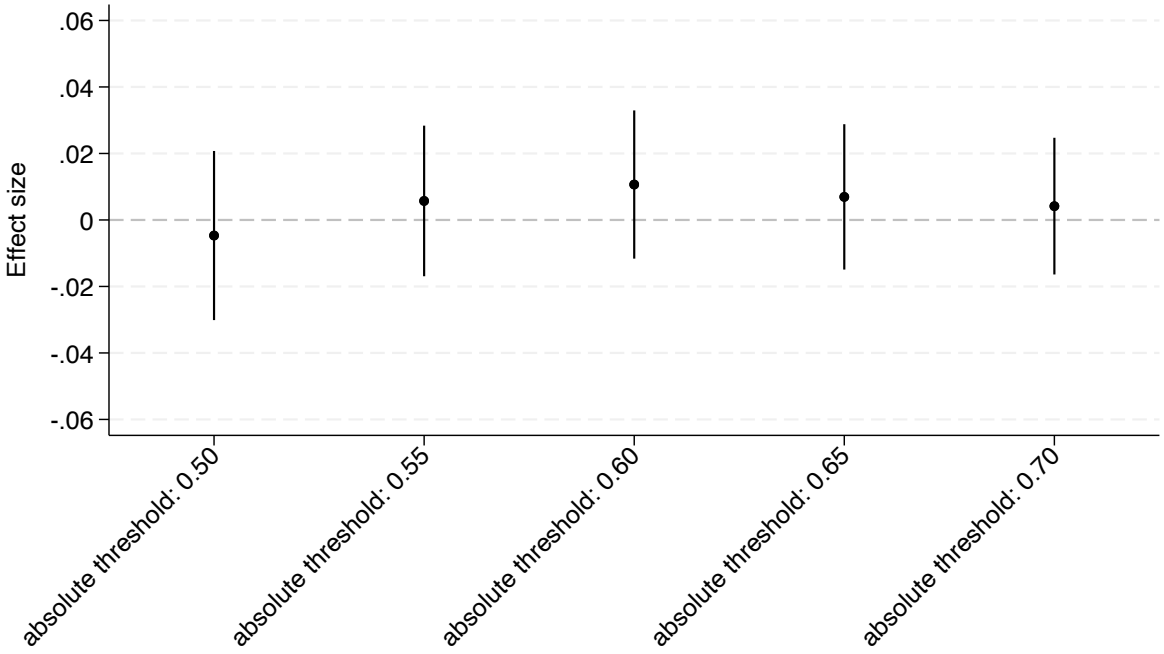
Notes: The figure presents point estimates (and 95% confidence intervals) from regressions of assessment performance (measured as the proportion of correct answers in the assessment) on an alternative treatment definition. This alternative treatment distinguishes between three question types instead of two (first question easy and first question difficult). First questions are defined as easy if the proportion of correct answers given by fellow students is higher than the 66th percentile of the distribution of these proportions for a given assessment, and are defined as difficult if this proportion is lower than the 33rd percentile (see Section **??** for more details). The remaining questions are defined as being of medium difficulty and form the baseline group. The number of observations is 1,813.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure O.4: Robustness Checks with Three Question Types
Free University of Berlin



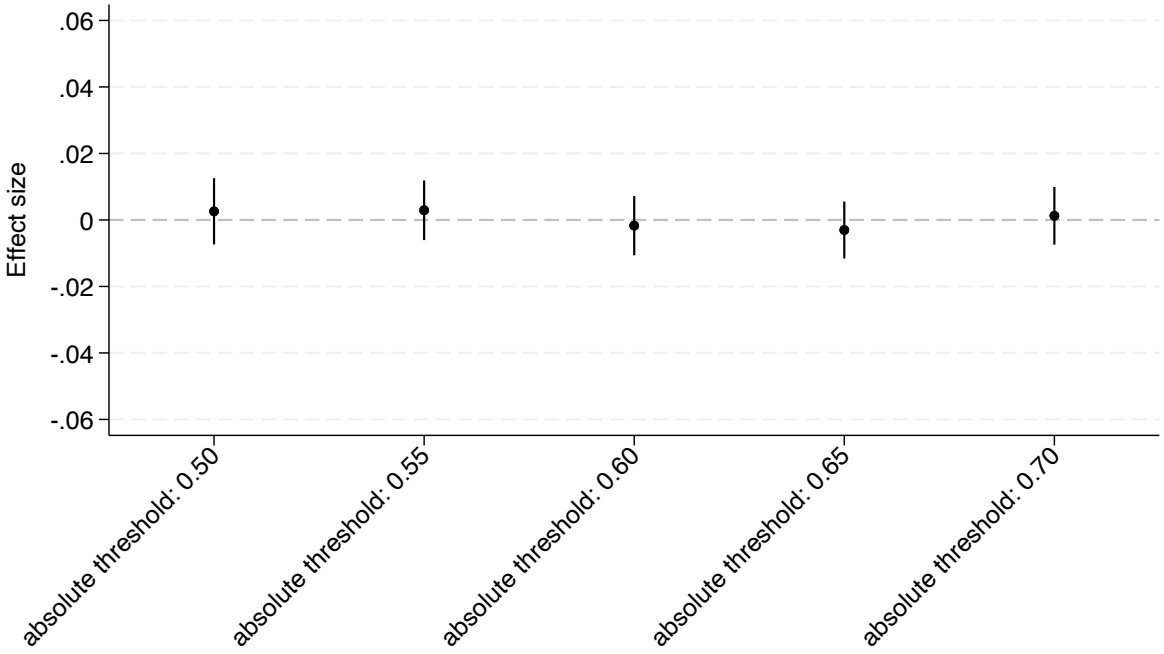
Notes: The figure presents point estimates (and 95% confidence intervals) from regressions of assessment performance (measured as the proportion of correct answers in the assessment) on an alternative treatment definition. This alternative treatment distinguishes between three question types instead of two (first question easy and first question difficult). First questions are defined as easy if the proportion of correct answers given by fellow students is higher than the 66th percentile of the distribution of these proportions for a given assessment, and are defined as difficult if this proportion is lower than the 33rd percentile (see Section **??** for more details). The remaining questions are defined as being of medium difficulty and form the baseline group. The number of observations is 6,583.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure O.5: Robustness Checks with Absolute Thresholds
University of Hamburg



Notes: The figure displays point estimates (and 95% confidence intervals) from regressions of assessment performance (measured by the share of correct answers in the assessment) on an alternative definition of the treatment variable. This alternative treatment is defined based on an absolute threshold. The first questions are defined as easy if the proportion of correct answers, given by fellow students in a given assessment, is higher than or equal to an absolute threshold, and are defined as difficult if this proportion is smaller than the threshold. The number of observations is 1,813.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Figure O.6: Robustness Checks with Absolute Thresholds
Free University of Berlin



Notes: The figure displays point estimates (and 95% confidence intervals) from regressions of assessment performance (measured by the share of correct answers in the assessment) on an alternative definition of the treatment variable. This alternative treatment is defined based on an absolute threshold. The first questions are defined as easy if the proportion of correct answers, given by fellow students in a given assessment, is higher than or equal to an absolute threshold, and are defined as difficult if this proportion is smaller than the threshold. The number of observations is 6,583.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table O.1: Robustness Checks: The Effect of Difficult First Question(s) on Overall Assessment Performance
University of Hamburg

						First Two Questions		Average Difficulty		Last Question
	All Controls	Test FE x Sem. FE	SE Clustered at Test Level	Two-way Clust.	Ind. FE	Two Easy Questions	Two Difficult Questions	First Two Questions	First Three Questions	Difficult Question
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Median Split ($D^{50\%}$)										
Treatment effect	-0.012	-0.012	-0.004	-0.004	0.002	-0.003	0.000	0.001	0.007	-0.006
	(0.009)	(0.009)	(0.010)	(0.009)	(0.008)	(0.011)	(0.014)	(0.011)	(0.011)	(0.010)
Share of observations						28%	18%	50%	50%	46%
Panel B: 3rd Decile Split ($D^{30\%}$)										
Treatment effect	0.010	0.009	0.015	0.015	0.019*	-0.009	-0.019	-0.004	0.000	0.001
	(0.010)	(0.010)	(0.011)	(0.010)	(0.008)	(0.010)	(0.020)	(0.012)	(0.013)	(0.010)
Share of observations						51%	6%	70%	70%	28%
Panel C: 2nd Decile Split ($D^{20\%}$)										
Treatment effect	0.000	-0.002	0.004	0.004	0.016	0.002	-0.018	0.007	0.008	0.017
	(0.012)	(0.012)	(0.010)	(0.010)	(0.010)	(0.011)	(0.033)	(0.014)	(0.015)	(0.012)
Share of observations						66%	2%	80%	80%	19%
Panel D: 1st Decile Split ($D^{10\%}$)										
Treatment effect	0.005	0.004	0.008	0.008	0.020			0.008	0.020	0.032
	(0.016)	(0.017)	(0.015)	(0.015)	(0.014)			(0.018)	(0.019)	(0.018)
Share of observations								90%	90%	8%
Panel E: Continuous Variable (D^{cont})										
Treatment effect	-0.001	-0.003	0.004	0.004	0.028			0.008	0.068	-0.016
	(0.027)	(0.027)	(0.023)	(0.022)	(0.025)			(0.040)	(0.048)	(0.028)

Notes: Column 1 includes all available students characteristics as control variables: gender, semester enrolled in, study program and motivation. In column 2 all control variables are included, together with the interaction terms between assessment fixed effects and semester fixed effects. In column 3, standard errors at clustered at the assessment level instead of the individual level. In column 4, standard errors are clustered at the assessment and individual level together (two-way clustering), and in column 5, we also control for individual fixed effects. In columns 6–7, the difficulty-based question order of the first two questions are examined; column 6 displays the effect of having two easy questions first, and column 7 of having two difficult questions first. The baseline group in this specification consists of students who got one easy and one difficult question among the first two (irrespective of their order). Columns 8–9 are based on the average difficulty level of the first two and three questions, respectively. Column 10 estimates the effect of the last question being difficult. Different definitions of the treatment variable "difficult question" are presented in panels A–E. In column 1 and 2, the sample size is 1,757 observations and the sample mean of the outcome variable is 0.645. In all other specifications, the sample size is 1,813 observations and the sample mean of the outcome variable is 0.644. Robust standard errors (in parentheses) allow for clustering. ** p<0.01, * p<0.5.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).

Table O.2: Robustness Checks: The Effect of a Difficult Question(s) on Overall Assessment Performance: Free University of Berlin

				First Two Questions		Average Difficulty		Last Question
	SE Clustering at Test Level	Two-way Clustering	Ind. FE	Two Easy Questions	Two Difficult Questions	First Two Questions	First Three Questions	Difficult Questions
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Median Split ($D^{50\%}$)								
Treatment effect	0.006	0.006	-0.000	-0.001	-0.001	0.005	0.003	0.007
	(0.004)	(0.004)	(0.003)	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
Share of observations				24%	23%	50%	50%	50%
Panel B: 3rd Decile Split ($D^{30\%}$)								
Treatment effect	0.005	0.005	-0.000	-0.009*	-0.006	-0.001	0.008	0.008
	(0.005)	(0.005)	(0.004)	(0.004)	(0.008)	(0.005)	(0.005)	(0.004)
Share of observations				49%	8%	70%	70%	30%
Panel C: 2nd Decile Split ($D^{20\%}$)								
Treatment effect	0.006	0.006	0.001	-0.008	-0.012	-0.002	0.007	0.006
	(0.005)	(0.005)	(0.004)	(0.004)	(0.011)	(0.005)	(0.006)	(0.005)
Share of observations				65%	3%	80%	80%	18%
Panel D: 1st Decile Split ($D^{10\%}$)								
Treatment effect	0.010	0.010	0.005			0.005	0.003	0.009
	(0.007)	(0.007)	(0.006)			(0.007)	(0.007)	(0.007)
Share of observations						90%	90%	9%
Panel E: Continuous Variable (D^{cont})								
Treatment effect	0.006	0.006	-0.005			0.001	0.019	0.015
	(0.009)	(0.009)	(0.009)			(0.015)	(0.018)	(0.010)

Notes: All estimations are based on 6,583 observations and the sample mean of the outcome variable is 0.628. In columns 1–3, results of the regression of the overall assessment performance (measured by the share of correct answers in the assessment) on the treatment variable "difficult question" are displayed. In column 1, standard errors are clustered at the assessment level instead of the individual level. In column 2, standard errors are clustered both at the assessment and individual level (two-way clustering). In column 3, individual fixed effects are added. In columns 4 and 5, the difficulty-based question order of the first two questions are considered; column 4 shows the effect of having two easy questions first and column 5 of having two difficult questions first. The baseline group in this specification consists of students who got one easy and one difficult question among the first two (irrespective of their order). Columns 6–7 are based on the average difficulty level of the first two and three questions, respectively. In column 8, the effect of the last question being difficult is estimated. Different definitions of the treatment variable "difficult question" are presented in panels A–E. Robust standard errors (in parentheses) allow for clustering. ** p<0.01, * p<0.05.
Source: Own calculations based on data from the e-learning platform mcEmpirics (`https://www.mcempirics.com`).