The effect of relative performance feedback on academic outcomes

Rupert Sausgruber[1], Rudolf Winter-Ebmer[2], Mario Lackner[2], and Anna Schwarz[1]

[1]Wirtschaftsuniversität (WU) Wien
[2]Johannes Kepler Universität (JKU) Linz

February 18, 2022

Abstract

Despite extensive research on the effect of relative performance feedback, a consensus on the direction and mechanics of the effect, especially in the higher education context, is far. Moreover, the role of students' prior beliefs about their performance is poorly understood. Using a large-scale survey at a big European university, we can cleanly elicit these beliefs before providing relative performance feedback to a randomly selected treatment group. By linking the survey data to administrative data, we can subsequently measure the effect on academic performance in the short- and long-run and by prior beliefs. Additionally, the treatment group has to actively decide to see the information, which allows us to analyze selection into receiving relative performance feedback. The survey context further enables us to estimate the effect of relative performance feedback on several secondary outcomes, such as competitiveness, stress, academic self-concept. Our clean experimental design using survey as well as administrative data lends itself to comprehensively analyze the effect of relative performance feedback in higher education and provide evidence on the role of prior performance beliefs.

1 Background

Related Literature There has been extensive research on the effect of relative performance feedback in educational settings, but also at the work place or in the laboratory. A very good review of the literature can be found in Villeval (2020). However, a consensus on the direction of the effect, it's mechanisms and heterogeneous impacts is far. Positive effects seem to be mostly driven by improved self-esteem and competitive preferences (Villeval, 2020). Dobrescu et al. (2021) also provide evidence that relative performance feedback can benefit everyone by increasing peer interactions and encouraging social learning. On the other hand, several studies find negative effects on performance, which mostly arise at the extremes of the performance distribution. Top performers have been shown to decrease performance after relative performance feedback, when they previously underestimated their rank (Azmat et al., 2019), when they are female and averse to competition (Cabrera, Cid, et al., 2017), or when social norms lead them to adjust their performance towards the mean (Ashraf, 2019; Blader, Gartenberg, and Prat, 2020). However, Brade, Himmler, and Jackle (2020) also show, that students who underestimate themselves can react by increasing their performance, when they receive good news. Negative effects for low performers are mostly driven by learning about ability and optimal effort provision, but also discouragement or shame, and willingness to avoid discouragement (Goulas and Megalokonomou, 2021; Ashraf, Bandiera, and Lee, 2014). Finally, Hermes et al. (2021) provide evidence that communicating relative performance feedback in a dynamic way (i.e. how much the own rank has changed) can counter

such negative effects, especially for low performers and girls. Our study can contribute to the literature by providing evidence for the effects of relative performance feedback in a large-scale field experiment at a public university. By exploiting rich survey data linked to administrative records we can further examine which mechanisms are at place and which groups react in which way, especially focusing on heterogeneity in effects by prior beliefs in one's own relative performance.

Institutional context We implement our field experiment at a large public university in Europe. In comparison to Azmat et al. (2019), our sample is more general in terms of socio-economic characteristics, which adds to the external validity of our results.

2 Experimental Details

2.1 Description of the intervention

The intervention is placed within a university-survey, that is sent to all bachelor beginners at the end of their first semester. The treatment consists of providing one half of students with the opportunity to be informed about their position in the GPA (Grade Point Average) as well as the ECTS (European Credit Transfer and Accumulation System) distribution. It is thus an intention-to-treat design. We implement this by telling students the quartile of the distribution in which their performance is located. The reference group consists of all students who started their studies in the same semester and for the same field of study. The decision whether students choose to see their relative performance will be recorded. This enables us to analyze, who selects into receiving information. Allowing students to opt into receiving performance feedback may also attenuate negative effects, that are documented in the literature.

Timeline The intervention is included in the survey in February 2022. The relevant student cohort are thus all students who started their studies in October 2021. We will follow the students over several semesters to also estimate potential long-term effects of the intervention. We further have the possibility to repeat this procedure in following survey rounds and thus with upcoming student cohorts. The survey is sent every year after the winter term; i.e. at the end of January or early February.

2.2 Data & Sample

We have access to two types of data, that can be matched at the individual level.

Administrative data The university provides us with administrative data about students' baseline characteristics (gender, age, secondary school qualification, nationality) as well as academic performance (ECTS, number of courses taken and passed, and course grades). The structure of the administrative data is an individual panel, which means that we can link individual student observations over time.

Survey Data We have access to an array of survey measures taken from the standard survey conducted by the university. In addition, we include some survey items for the specified purpose of our study:

- Survey items to be answered before the intervention:
  - Competitive preferences and locus of control (Rotter, 1966)
  - Questions on how much students interact with fellow students

  - Prior beliefs about relative performance, separated by ECTS and GPA

- Survey items to be answered after the intervention:
  - Satisfaction with the studies
  - Study effort in the past and upcoming semester
  - Stress and perceived competitiveness in the studies
  - Academic self-concept (Reynolds, 1988)
  - Own and perceived prosociality in the studies

Attrition & Exclusions The administrative data covers the universe of all students; i.e. there are no concerns about attrition, by definition. Dropouts are recorded and analyzed among the outcomes of interest. In the survey, response rates typically lie between 30% and 50%, according to previous records. We can test for sample selection by means of administrative data, which is available for all students including those who do not participate in the survey. Participation in the survey is voluntary and students can stop the survey at any point. To attenuate potential attrition by survey drop-out, we place the block of our questions relatively in the beginning of the survey. Nevertheless, we will check for systematic attrition with the administrative data available.

Some students in the data are inactive and not taking any courses. The reasons for this are mainly administrative, e.g. because students do not deregister when they switch to another university. To prevent such zero entries from biasing the feedback information, we exclude such inactive students from the sample and only consider students who have passed at least one course for the calculation of performance ranks.

2.3 Treatment assignment

The assignment to treatment and control group is done via stratified randomization at the individual level. We stratify the sample by gender (binary), study program (business and economics or business law), and secondary school qualifications. The latter variable is coded in three values indicating academic secondary school, vocational secondary school, and others. The category "others" encompasses mostly second-chance educational qualifications and those obtained abroad.

3 Analysis

3.1 Hypotheses

Following the literature, it is a priori unclear how the treatment will affect students' academic outcomes. The overall effect likely depends on student characteristics and prior beliefs about relative performance. Following Azmat and Iriberri (2010), our baseline hypothesis is formulated as the Null

Hypothesis 0 *H_0:Providing students with the opportunity to receive feedback on relative past performance will have no significant effect on performance measured in GPA or ECTS.*

Self-assessment of students is found to play a critical role in this context. For example, Azmat et al. (2019) find that students who underestimate themselves experience a decrease in performance due to relative performance feedback. This sub-population is found to be responsible for an overall negative effect. This finding is contrasted by the results of Brade, Himmler, and Jackle (2020). They find that students who underestimate themselves experience a positive effect of feedback on performance. Consequently, we propose to test the following hypothesis:

Hypothesis 1a *Students experience different effects of the treatment depending on their prior beliefs of past relative performance.*

Previous results suggest that absolute and relative ability of students may play a critical role. For example, Bandiera, Larcinese, and Rasul (2015) find that relative performance feedback has a stronger positive effect on performance of more able students, compared to students with lower ability. In our experimental setting, we have the possibility to proxy the ability of students by their relative performance in the semester preceding treatment assignment. Consequently we propose to test the following alternative hypothesis:

Hypothesis 1b *The treatment affects students differently depending on their relative performance.*

Gender differences in the reaction to feedback are also documented in the literature (Wozniak, 2012; Berlin and Dargnies, 2016; Czibor et al., 2020). In an education context, gender differences in the reaction to relative performance feedback are elusive (Azmat and Iriberri, 2010; Tran and Zeckhauser, 2012). Recently, Dobrescu et al. (2021) find that short term positive effects of relative performance feedback are only significant for male students, while long-term effects in GPA are relevant for female students. To provide insights, we propose to examine whether female students respond differently to relative feedback on past performance than male students. We want to test the following hypothesis:

Hypothesis 1c *The treatment affects female and male students differently.*

The survey contains an item to measure preferences regarding competition. Drawing on previous theoretical and empirical results (Azmat and Iriberri, 2010), it is plausible that students with preferences for competition react more strongly to receiving feedback about relative performance. Thus, we plan to test the following hypothesis:

Hypothesis 1d *Students who have stronger preferences for competition will show different effects than students who are less competitive.*

Eventually, previous research has shown a relationship between internal or external locus of control and multiple work outcomes, such as performance, motivation and coping with stress (Ng, Sorensen, and Eby, 2006; Wang, Bowling, and Eschleman, 2010). Related to our study, Hermes et al. (2021) show that emphasizing a growth mindset, which is strongly related to internal locus of control, along with relative performance feedback can be beneficial particularly for low-performing students. Similarly, Azmat et al. (2019) point out that the effect of relative performance feedback may depend on people's beliefs about effort efficacy, which is again tightly linked to an internal locus of control. Therefore, we test the following hypothesis:

Hypothesis 1e *The treatment will have different effects on students characterized by internal and external locus of control.*

In addition to the causal effect of relative performance feedback on subsequent performance, our experimental setting enables us to analyze who selects into receiving relative performance feedback. Women and men are found to have different preferences concerning the willingness to compete (Niederle and Vesterlund, 2011; Lackner, 2021). Typically, men show a higher tendency to engage in competitions, sometimes detrimental to subsequent performance (Niederle and Vesterlund, 2007). We thus test the following hypothesis:

Hypothesis 2a *Female students are less likely to retrieve information on relative performance than male students.*

Eventually, personal beliefs about own relative performance will affect the willingness to retrieve feedback about past relative performance. Consequently, we plan to test the following hypothesis:

Hypothesis 2b *Students who overestimate themselves are more likely to retrieve relative performance feedback than students who underestimate themselves.*

3.2 Outcomes of interest

Primary outcomes Our primary outcomes are students' performance and whether the feedback information is retrieved.

- Information retrieved
- ECTS completed
- GPA

Secondary outcomes The secondary outcomes are used to disentangle possible mechanisms of the treatment effect. For this, we use the variables collected within the survey after the intervention. This includes study satisfaction, study effort, stress and perceived competitiveness, academic self-concept, and prosociality in the studies.

3.3 Estimation and inference

Intention-to-treat: To estimate the average effect of the treatment (referring to Hypothesis 0), we run the following regression model, which identifies the causal intention-to-treat effect under the assumption that the treatment is randomly assigned.

$$Y_{i,t} \;=\; \beta_0 \;+\; \beta_{1i} * Treated \;+\; \beta * \mathbf{X_i} \;+\; \delta_s \;+\; \epsilon_i \tag{1}$$

$Y_{i,t}$ refers to the respective outcome of interest (see above). We can observe the outcome variables at different instances of time, allowing us to estimate short-term effects (in the next semester) as well as longer-term effects (in the upcoming one to two years). β_{1i} specifies the intention-to-treat effect of being assigned to the treatment group. X_i includes several controls, such as one's rank in the performance distribution measured in quartiles, one's belief about the rank, and an array of socio-demographic characteristics. Further, δ_s refers to strata fixed effects. The university does not impose fixed class structures and students are usually mixed in different courses. Clustering standard errors at the class level, as in many education studies, is therefore not appropriate. However, we estimate robust and bootstrapped standard deviations.
As mentioned above, we expect that the treatment will have different effects on different subgroups. We conduct five pre-specified heterogeneity analyzes related to the hypotheses 1a to 1e using the following regression.

$$Y_{i,t} \;=\; \beta_0 \;+\; \beta_{1i} * Treated * (T_i) \;+\; \beta_{2i} * (T_i) \;+\; \beta * \mathbf{X_i} \;+\; \delta_s + \;\epsilon_i \tag{2}$$

In addition to regression 1, regression 2 includes interactions to test for heterogeneous treatment effects. (T_i) refers to the respective heterogeneity variable. Prior beliefs about the respondents' performance rank (separated by ECTS and GPA) are derived from questions in the survey. This variable has three values indicating whether the respective student over-, under-, or correctly estimates her performance. Another dimension of heterogeneity is the performance quartile of the student, again respectively for ECTS and GPA, which is identical to the feedback they receive in the treatment group. Locus of control and competitive preferences are derived from the corresponding survey questions and coded as dummy variables by median split. Eventually, we estimate regression 1 separately by gender to test for gender heterogeneous treatment effects.[1]

[1] We refrain from using an interaction term here, because of the perfect collinearity of gender and the strata fixed effects, as gender is one of the stratification variables used for randomization.

Average treatment effect on the treated: By our design, not everybody who is assigned to the treatment group gets treated, but only those who choose to see the information are effectively treated. Using the treatment assignment as an instrument for seeing the information, we can estimate the causal effect of seeing the relative performance feedback. This effect is equivalent to the treatment effect on the treated.[2] We, thus, estimate the following two-stage model:

$$retrieved_i \;=\; \pi_0 + \pi_{1,i} * \; Treated \;+\; \xi' X_i \;+\; \delta_s \;+\; \nu_i \tag{3}$$

$$Y_{i,t} \;=\; \beta_0 \;+\; \beta_{1,i} * \overline{retrieved_i} \;+\; \xi' X_i \;+\; \delta_s \;+\; \epsilon_i \tag{4}$$

where $retrieved_i$ is a binary variable equal to 1 if the observed treated student i has made the decision to retrieve information on past relative performance, 0 if otherwise.
Regression 3 represents the first stage effect from the treatment assignment on seeing the information, while regression 4 is the second stage, where the fitted values of the first stage are included as an independent variable. The corresponding coefficient $\beta_{1,i}$ captures the treatment effect on the treated (i.e. on those in the treatment group who actually saw the information).

Information pick-up In order to test Hypothesis 2a we estimate the following model separately by gender.

$$retrieved_i \;=\; \alpha_0 \;+\; \alpha_1 * Treated \;+\; \beta * \mathbf{X_i} \;+\; \delta_s \;+\; \nu_i, \tag{5}$$

The difference between the two estimates of α_1 then measures the gender gap in the willingness to receive feedback. To account for the randomization method, we again include strata (δ_s) fixed effects. To test hypothesis 2b, we interact the Treatment dummy in regression 5 with our measure of prior beliefs. The interaction coefficients measure the differential propensity to retrieve relative performance feedback by the students' prior beliefs.

Exploratory analysis: We consider the analysis of treatment effects on our secondary outcomes to be exploratory. Specifically, we test for treatment effects on study satisfaction, study effort, stress and perceived competitiveness, academic self-concept, and prosociality in the studies.

Treatment spill-overs: We test for treatment spill-over effects in an indicative way by comparing the treatment effect between those who have many contacts with fellow students and those with little contact. We split the sample at the median of this variable.

Adjusting for multiple outcomes and hypotheses testing: To control for a possible false discovery rate associated with testing multiple hypothesis, we take two approaches. In regard to the primary outcome variables, we additionally report the mean standardized treatment effect with its standard error adjusted for the dependency between the different outcome variables, following Duflo, Glennerster, and Kremer (2007). In our analysis, we further plan to test for heterogeneous treatment effects across different groups of subjects. Consequently, we propose to estimate step-down adjusted p-values robust to multiple hypothesis testing as outlined in Romano and Wolf (2005a) and Romano and Wolf (2005b). This approach calculates p-values testing significance of one out of multiple hypotheses, accounting for the probability of a Type I error among all tested

[2] In order for this identification to be valid, we have to make two assumptions. First, we assume that the possibility to see the true rank does not affect the outcome variable, if the information is not seen; i.e. that outcomes of *non-compliers* in the treatment group are comparable to the control group. The second assumption is that assignment to treatment has a monotonic effect on seeing the information. This means that assignment to the treatment group cannot make it less likely to see the information than for the control group, which is true by construction of the design.

hypotheses. This approach is similar, but more general than the procedure proposed by Westfall and Young (1993).

3.4 Power Analysis

We first estimate a target sample size needed to detect a reasonable effect size for our main result with 80% power and 5% significance level. For the average treatment effect, already as of the first year (most conservative scenario N=1,000) the power is sufficient to detect effect sizes of 0.13 and 0.16 of a standard deviation for ECTS and GPA, respectively (see minimum detectable effect size estimates in the appendix). After three years, the sample is sufficiently large to analyze heterogeneous treatment effects by prior beliefs (hypothesis 1a) with a power of 80%. This is calculated as follows.

As discussed in section 3.1, we expect students to react differently depending on their prior beliefs (see hypothesis 1a). General average effect sizes in the related literature are around 0.15 to 0.2 of a standard deviation. Following Brade, Himmler, and Jackle (2020), we expect that students who underestimate their performance will show the largest effects. We thus assume that this group will exert effect sizes of 0.25 of a standard deviation on ECTS and GPA. We also calculate power for 0.2 and 0.3 of a standard deviation as one more conservative and one more optimistic scenario. Brade, Himmler, and Jackle (2020) even estimate an effect size of around 0.5 of a standard deviation for the subsample of students whose performance is above average, but who underestimate themselves.[3]

To calculate the sample size needed to detect the above mentioned effect sizes in our setting, we use data from the previous student cohort (i.e. those that started their studies in October 2020). Power analysis via simulation can additionally account for our stratified randomization procedure and the inclusion of some control variables. We run the power analysis for several sample sizes, using the effect sizes specified above to check at which N we surpass the threshold of 80% power. For each N, we randomly draw a sample of size N from the whole student cohort population with replacement, i.e. a test sample. Half of the respective sample is allocated to the treatment group, using the same randomization method that we apply to the data (i.e. stratified by gender, study program, and secondary school qualifications). We then simulate treatment effect regressions following the specification in regression 2. The effect size for the subgroup of students who underestimate themselves is chosen exogenously, as described above. This procedure is repeated 10,000 times for each sample size. For each N, the share of interaction effects for the students underestimating themselves that are significant at the 5% level is equivalent to the power of the analysis.

We have to estimate the prior belief distribution, as the data on beliefs is not contained yet in the data of the previous cohort. For this, we refer to existing knowledge on performance beliefs among university students from the literature. Following Azmat et al. (2019) and Brade, Himmler, and Jackle (2020), a fraction of around 50% of students underestimate themselves, 30% of the students overestimate themselves, and 20% hold correct beliefs about their performance. We also know that beliefs about performance are correlated with actual performance: we assume a correlation of $\rho = 0.5$ with their actual rank.[4]

[3] Azmat et al. (2019) estimate a negative effect of 0.15 standard deviations on GPA for the group underestimating themselves. However, their institutional setting differs from ours and we thus rely more on the estimates of Brade, Himmler, and Jackle (2020), who are more comparable to our setting and also more in line with other effect sizes in the related literature, such as Dobrescu et al. (2021) or Elsner, Isphording, and Zölitz (2021).

[4] Power is inversely related to ρ, the correlation between beliefs and actual performance. However, the results change very little, for example, if ρ changes from 0.5 to 0.2, the baseline MDE for N=1,000 decreases from 3.161 to 3.069. Additionally, we have greater power with equally sized groups. In that case MDEs decrease by about 20% across sample sizes, which is what one would expect.

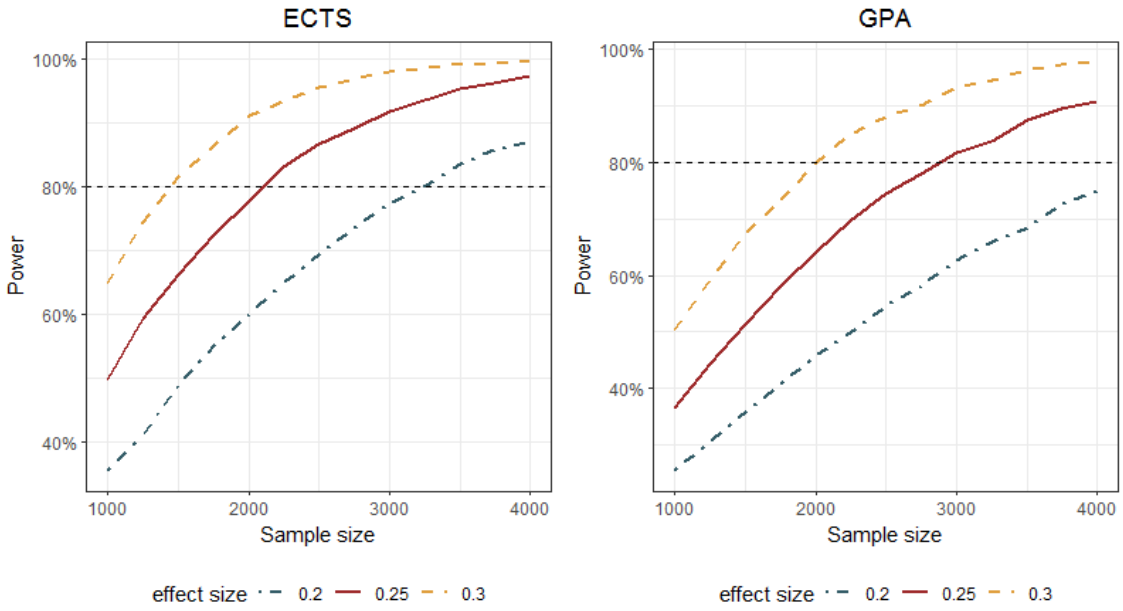

Figure 1: Statistical power by sample and effect size

In figure 1, we show the calculated power for estimating effect sizes regarding the interaction of the treatment dummy with a variable indicating whether students underestimate their position in the performance distribution. We expect effect sizes to be around 0.25 of a standard deviation for both outcomes, ECTS and GPA, as argued above. We thus need around 2.000 students for the ECTS outcome and around 3.000 students for GPA to be able to estimate heterogeneity effects of 0.25 of a standard deviation with sufficient power. 3.000 students is thus our target sample size. This gives us enough power to detect effect sizes of 0.2 of a standard deviation regarding ECTS, as in the more conservative scenario, and 0.25 regarding GPA. However, as we do not have perfect control over the sample size in our field setting, this should be seen as an approximate target. The average sample size for the survey has been 1,485 over the last five years, but last year only 1,279 students participated, which could be a Covid-effect. We therefore expect that we need to run the intervention for three years to reach sufficient power for our analysis on the heterogeneity by prior beliefs.

Additionally, we have done a power analysis for a specification where we estimate the treatment effect within the subgroup of students who underestimate themselves (as compared to estimating an interaction term in the full sample). In this specification, we reach sufficient power already at relatively lower N and smaller effect sizes, as can be seen in figure 2. If we miss the target sample size because of unexpectedly low response rates to the survey, we thus could still resort on this subgroup-specification to analyze heterogenous treatment effects with respect to beliefs.

We additionally estimate minimum detectable effect sizes for our all hypotheses, specified in section 3.1 for different sample sizes, which can be seen in the appendix. We do this analysis for 7 different sample sizes, ranging from 1,000 as a lower bound, to 3,000 as our target sample size.[5] The

[5] 1,279 and 1,485 are the number of students participating in the survey last year and on average over the last

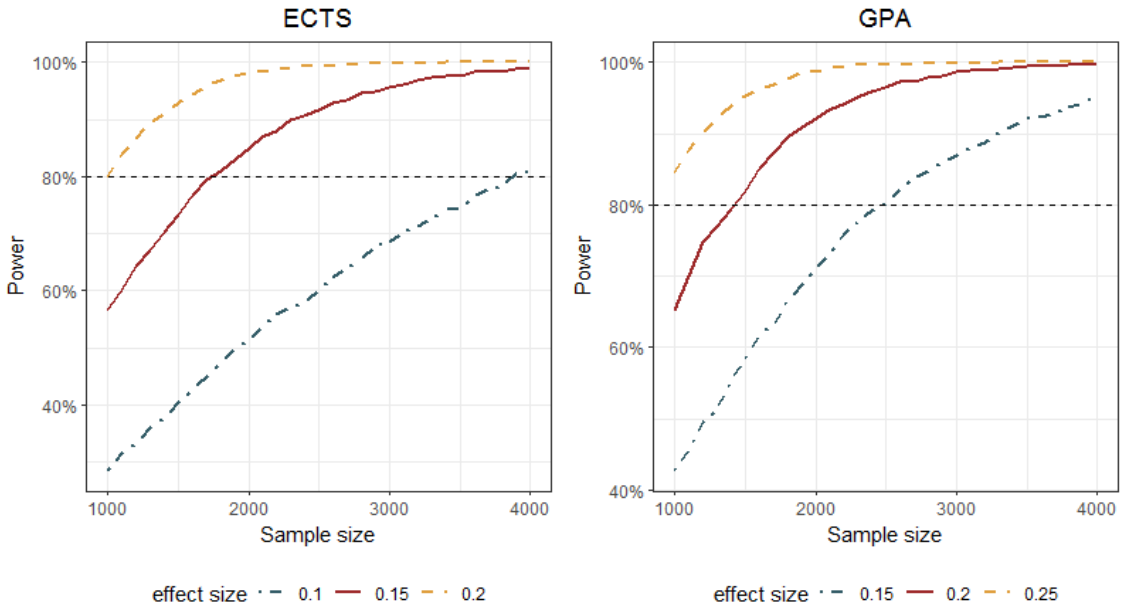

Figure 2: Statistical power by sample and effect size - subgroup analysis

procedure is the same as above. Test samples for each N are randomly drawn with replacement from the whole student population of last year's cohort and randomly assigned to the treatment and control group. Different from above, we now do not specifiy the effect size, but estimate placebo regressions for the actual analysis, i.e. average treatment effect estimations (see hypothesis 0) and the heterogeneity analyzes (testing hypotheses 1a to 1e).[6] The process is iterated 10,000 times and we thus get a distribution of 10,000 potential placebo effect sizes. Following Campos-Mercade and Wengström (2018), we then calculate by how much we would have to shift the distribution of of estimated treatment effects in order to reach a power of 80% at an α-level of 5%.[7] The resulting MDE estimates are discussed in the appendix.

4 IRB approval

The study described in this pre-analysis plan has been reviewed and approved by the Competence Center for Experimental Research at the Vienna University of Economics and Business under reference number WU-HSRP-2022-003.

five years, respectively. 1,800 is taken as an upper bound for participation in the first year, as this was the maximum of students participating in the last 10 years. The higher sample sizes then relate to samples where we repeat the intervention in the following 1-2 years and pool the data.

[6]We refrain from estimating MDEs for hypotheses 2a and 2b as we would have to simulate the dependent variable, which is a survey measure.

[7]This effect size is independent of the direction of the effect, which means it applies for positive as well as negative effects. This approach is very straightforward and flexible and lends itself perfectly to power calculation for heterogeneity analysis, as we do not need to assume any direction of the effects.

References

Ashraf, Anik (2019). "Do performance ranks increase Productivity? Evidence from a field experiment". Mimeo, University of Warwick.

Ashraf, Nava, Oriana Bandiera, and Scott S Lee (2014). "Awards unbundled: Evidence from a natural field experiment". In: *Journal of Economic Behavior & Organization* 100, pp. 44–63.

Azmat, Ghazala, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri (2019). "What You Don't Know... Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education". In: *Management Science* 65.8, pp. 3714–3736.

Azmat, Ghazala and Nagore Iriberri (2010). "The importance of relative performance feedback information: Evidence from a natural experiment using high school students". In: *Journal of Public Economics* 94.7–8, pp. 435–452.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul (2015). "Blissful ignorance? A natural experiment on the effect of feedback on students' performance". In: *Labour Economics* 34, pp. 13–25.

Berlin, Noémi and Marie-Pierre Dargnies (2016). "Gender differences in reactions to feedback and willingness to compete". In: *Journal of Economic Behavior & Organization* 130, pp. 320–336.

Blader, Steven, Claudine Gartenberg, and Andrea Prat (2020). "The contingent effect of management practices". In: *The Review of Economic Studies* 87.2, pp. 721–749.

Brade, Raphael, Oliver Himmler, and R Jackle (2020). "Relative Performance Feedback and the Effects of Being Above Average–Field Experiment and Replication". In: *MPRA Paper* 88830.

Cabrera, José María, Alejandro Cid, et al. (2017). "Gender differences to relative performance feedback: A field experiment in education". In: *Universidad de Montevideo wpaper* 1704.

Campos-Mercade, Pol and Erik Wengström (2018). "Incentives and education". In: *AEA RCT Registry*. URL: `https://doi.org/10.1257/rct.2690-1.1`.

Czibor, Eszter, Sander Onderstal, Randolph Sloof, and C Mirjam van Praag (2020). "Does relative grading help male students? Evidence from a field experiment in the classroom". In: *Economics of Education Review* 75, p. 101953.

Dobrescu, LI, Marco Faravelli, Rigissa Megalokonomou, and Alberto Motta (2021). "Relative performance feedback in education: Evidence from a randomised controlled trial". In: *The Economic Journal* 131.640, pp. 3145–3181.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007). "Using randomization in development economics research: A toolkit". In: *Handbook of development economics* 4, pp. 3895–3962.

Elsner, Benjamin, Ingo E Isphording, and Ulf Zölitz (Apr. 2021). "Achievement Rank Affects Performance and Major Choices in College". In: *The Economic Journal* 131.640, pp. 3182–3206.

Goulas, Sofoklis and Rigissa Megalokonomou (2021). "Knowing who you actually are: The effect of feedback on short-and longer-term outcomes". In: *Journal of Economic Behavior & Organization* 183, pp. 589–615.

Hermes, Henning, Martin Huschens, Franz Rothlauf, and Daniel Schunk (2021). "Motivating low-achievers—Relative performance feedback in primary schools". In: *Journal of Economic Behavior & Organization* 187, pp. 45–59.

Lackner, Mario (2021). "Gender differences in competitiveness". In: *IZA World of Labor*.

Ng, Thomas WH, Kelly L Sorensen, and Lillian T Eby (2006). "Locus of control at work: a meta-analysis". In: *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 27.8, pp. 1057–1087.

Niederle, Muriel and Lise Vesterlund (2007). "Do women shy away from competition? Do men compete too much?" In: *The quarterly journal of economics* 122.3, pp. 1067–1101.

– (2011). "Gender and competition". In: *Annu. Rev. Econ.* 3.1, pp. 601–630.

Reynolds, William M (1988). "Measurement of academic self-concept in college students". In: *Journal of personality assessment* 52.2, pp. 223–240.
Romano, Joseph P and Michael Wolf (2005a). "Exact and approximate stepdown methods for multiple hypothesis testing". In: *Journal of the American Statistical Association* 100.469, pp. 94–108.
– (2005b). "Stepwise multiple testing as formalized data snooping". In: *Econometrica* 73.4, pp. 1237–1282.
Rotter, Julian B (1966). "Generalized expectancies for internal versus external control of reinforcement." In: *Psychological monographs: General and applied* 80.1, p. 1.
Tran, Anh and Richard Zeckhauser (2012). "Rank as an inherent incentive: Evidence from a field experiment". In: *Journal of Public Economics* 96.9-10, pp. 645–650.
Villeval, Marie Claire (2020). "Performance Feedback and Peer Effects". GLO Discussion Paper.
Wang, Qiang, Nathan A Bowling, and Kevin J Eschleman (2010). "A meta-analytic examination of work and general locus of control". In: *Journal of Applied Psychology* 95.4, p. 761.
Westfall, Peter H. and S. Stanley Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, United States.
Wozniak, David (2012). "Gender differences in a market with relative performance feedback: Professional tennis players". In: *Journal of Economic Behavior & Organization* 83.1, pp. 158–171.

A Appendix

A.1 Minimum detectable effect sizes

The MDE estimates for the average treatment effects by sample size are shown in Table 1. The estimates are in terms of Cohen's d effect sizes, i.e. adjusted for the standard deviation of the dependent variable. They are all in the realm of small effect sizes and range from 0.13 (ECTS) or 0.16 (GPA) for N=1,000 to 0.08 (ECTS) or 0.10 (GPA) for N=3,000. Thus, for average effect estimation, we have enough power to detect even small effects already in the most conservative scenario for only one round of intervention.

	1,000	1,279	1,485	1,800	2,000	2,500	3,000
ECTS	0.135	0.119	0.112	0.100	0.097	0.085	0.078
GPA	0.164	0.143	0.133	0.121	0.116	0.102	0.097

Table 1: MDE estimates for average treatment effects

MDE estimates, again adjusted by the standard deviation of the outcome variable, for the main heterogeneity analysis by prior beliefs can be seen in Table 2. The minimum detectable effect sizes for the baseline category are naturally smaller than the effects than can be estimated for the interaction terms, i.e. the difference between the baseline and those who over- or underestimate themselves. The effect sizes for the baseline range from 0.31 (ECTS) or 0.35 (GPA) of a standard deviation for N=1,000, and 0.18 (ECTS) or 0.20 (GPA) for N=3,000, respectively. Between the baseline and those who overestimate themselves MDE estimates range from from 0.38 (ECTS) or 0.46 (GPA) for N=1,000 to 0.22 (ECTS) or 0.26 (GPA) for N=3,000. Finally, for those who underestimated themselves the effect sizes range from 0.37 (ECTS) or 0.43 (GPA) for N=1,000 to 0.22 (ECTS) or 0.25 (GPA) for N=3,000. The minimum detectable effect size for the target sample corresponds to the effect size we have specified above in section 3.4 to calculate the target sample size.

		1,000	1,279	1,485	1,800	2,000	2,500	3,000
ECTS	TE right	0.313	0.276	0.256	0.229	0.220	0.195	0.182
	I*(overestimate)	0.384	0.339	0.317	0.286	0.272	0.239	0.223
	I*(underestimate)	0.369	0.330	0.302	0.276	0.261	0.233	0.215
GPA	TE right	0.353	0.310	0.286	0.259	0.249	0.223	0.203
	I*(overestimate)	0.461	0.402	0.378	0.343	0.328	0.287	0.262
	I*(underestimate)	0.431	0.377	0.354	0.317	0.301	0.262	0.247

Table 2: MDE estimates for heterogeneous treatment effects by prior beliefs

We further discuss the power for testing hypothesis 1b, concerning the treatment effects by performance quartiles.[8] In Table 3 we present the MDE estimates in terms of Cohen's d effect sizes for the treatment effects by performance quartiles. Group 1 refers to the top performers and acts as the baseline. The estimates for the baseline range between 0.28 (ECTS) or 0.31 (GPA) for N=1,000 to 0.16 (ECTS) or 0.18 (GPA) for 3,000 students. The MDEs for the other three quartiles are of roughly equal size, because they all include an equal number of students. They range between 0.41 (ECTS) or 0.47 (GPA) for N=1,000 to 0.24 (ECTS) or 0.27 (GPA) for N=3,000.

		1,000	1,279	1,485	1,800	2,000	2,500	3,000
ECTS	Best (Group 1)	0.278	0.245	0.224	0.207	0.195	0.172	0.159
	I*(Group 2)	0.414	0.365	0.337	0.305	0.288	0.258	0.237
	I*(Group 3)	0.401	0.358	0.328	0.298	0.281	0.246	0.228
	I*(Group 4)	0.348	0.303	0.283	0.258	0.248	0.217	0.197
GPA	Best (Group 1)	0.311	0.275	0.255	0.234	0.220	0.196	0.178
	I*(Group 2)	0.443	0.393	0.361	0.330	0.316	0.280	0.251
	I*(Group 3)	0.443	0.392	0.366	0.335	0.315	0.280	0.258
	I*(Group 4)	0.470	0.423	0.393	0.355	0.340	0.280	0.273

Table 3: MDE estimates for heterogeneous treatment effects by performance quartiles

Finally, we can calculate the MDE estimates for the remaining hypotheses 1c to 1e. While estimating the MDEs by gender is straightforward,we have to simulate the outcomes of the survey measures as they cannot be yet observed. Therefore, we randomly allocate the students in our test sample into two equally sized groups representing the survey measures of high vs. low level of competitiveness and internal vs. external locus of control. The MDE estimates thus have to be taken with a grain of salt, as students will most likely not be randomly distributed across these two groups, but these measures will be correlated with other variables in the regression. The MDE results for this analysis, adjusted by the standard deviation of the outcome variable, can be seen in Table 4. For the subgroup regressions by gender, the minimum detectable effect sizes range between 0.2 (ECTS) or 0.25 (GPA) for N=1,000 to 0.12 (ECTS) or 0.14 (GPA) for N=3,000. With our target sample size of 3,000 students, we are thus able to estimate relatively small effects for

[8] Analyzing the treatment effects by performance quartiles reflects the mode of students' feedback in the treatment group. However, MDE estimates would be smaller by about 30% for ECTS and 25% for GPA, if we split the performance distribution at the median.

women and men separately. Regarding other sources of heterogeneity, we are able to detect effect sizes of 0.19 (ECTS) or 0.23 (GPA) for N=1,000 and 0.11 (ECTS) or 0.13 (GPA) for N=3,000 for the baseline group. To detect differences between these two groups, effect sizes should be around 0.27 (ECTS) or 0.33 (GPA) for N=1,000 and around 0.16 (ECTS) or 0.19 (GPA) for N=3,000. As already indicated above, it should be kept in mind that these variables needed to be simulated and the results are thus less precise.

		1,000	1,279	1,485	1,800	2,000	2,500	3,000
ECTS	men	0.204	0.180	0.164	0.151	0.142	0.128	0.116
	women	0.193	0.172	0.160	0.143	0.137	0.120	0.110
GPA	men	0.247	0.213	0.197	0.180	0.170	0.154	0.140
	women	0.230	0.201	0.190	0.172	0.160	0.146	0.131
ECTS	low competitiveness or external locus of control	0.191	0.167	0.155	0.142	0.134	0.120	0.111
	I*high competitiveness or internal locus of control	0.272	0.240	0.220	0.202	0.192	0.170	0.155
GPA	low competitiveness or external locus of control	0.230	0.204	0.187	0.171	0.164	0.146	0.130
	I*high competitiveness or internal locus of control	0.328	0.290	0.270	0.248	0.231	0.206	0.186

Table 4: MDE estimates for heterogeneous treatment effects by gender and survey measures