Gutenberg School of Management and Economics
& Research Unit “Interdisciplinary Public Policy”
Discussion Paper Series

Feedback in Homogeneous Ability Groups: A Field Experiment

Tim Klausmann

September 30, 2021

Discussion paper number 2114

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
https://wiwi.uni-mainz.de/

Contact Details:

Tim Klausmann
Department Law and Economics
Johannes Gutenberg-University Mainz
Jakob-Welder-Weg 4
55128 Mainz
Germany
tim.klausmann@uni-mainz.de

Feedback in Homogeneous Ability Groups: A Field Experiment *

Tim Klausmann†

September 30, 2021

Relative performance feedback often increases effort and performance on average. However, in the context of education, learners with low ability often do not profit from relative performance feedback. Less is known on how learners react to feedback when changing the feedback group composition. In a randomized field experiment we allocated 7352 learners into (i) *homogeneous* ability feedback groups, (ii) *heterogeneous* ability feedback groups, and (iii) a *control* group. All learners were observed in an online learning enviroment with anonymity between them. We find that on average relative performance feedback increases learning effort by 0.11 standard deviations. However, we do not observe any difference between learners in homogeneous and heterogeneous feedback groups on average. Further, we analyze the differential treatment effect for different ability levels between homogeneous and heterogenous feedback groups.

*We would like to thank Daniel Schunk, Florian Hett, Pawel Wasiak, Alexander Dzionara, Sigrid Ruland, Patrick Schneider, Isabell Zipperle, Niklas Witzig, Valentin Wagner, and the participants of the Luzern research workshop for helpful discussions and valuable feedback. We are grateful for the fruitful cooperation with the vocanto GmbH team, including Johannes Laudenberg, Johannes Schulte, and Stephan Hansen, and for receiving their unique data. We received financial support from the research priority program Interdisciplinary Public Policy (IPP) at the University of Mainz.

†Corresponding author, tim.klausmann@uni-mainz.de, Johannes Gutenberg University Mainz, Joachim-Becher-Weg 31, 55128 Mainz, Germany

Keywords: Feedback, Relative Performance, Heterogeneity, Education, Online Education, Peer Quality, Tracking, Learning Behavior, Gamification

JEL-Codes: C93, I20, I21

1. Introduction

Digital environments allow us to study aspects of human behavior in the field that were out of scope for experimentalists for decades. Both in workplaces and schools peer effects are confounds to central economic phenomena like the effects of feedback on workers or grouping students by ability. As an illustrative example from the analog world we follow and observe weak student *W*, mediocre student *M*, and strong student *S* on their first day of a new term. They walk into their school awaiting the assignment of their new classroom. Depending on classroom assignment they experience different schooling: the distinguishing feature of their school is an omnipresent relative performance feedback (RPF) score board in each classroom. Villeval (2020), Hattie and Clarke (2019) and many others informed the principal that the presence of feedback increases learning effort. However, if struggling *W* ends up in a heterogeneous classroom with *M* and *S*, *W* anticipates that the score board will be discouraging as '*W*' will be stuck at the bottom (compare Haenni, 2019; Gneezy and Fershtman, 2011). On the other hand, in a weak-learners-only classroom *W* will experience the effects of *homogeneous feedback* like vivid competition without constant discouragement. In a strong-learners-only classroom *S* will experience the positive effects of *homogeneous feedback*, as *S* will not be at the top of the score board. Staying on top will be a challenge rather than a given. In an intermediate classroom *M* expects to be somewhere in the middle of the board as in a classroom with *heterogeneous feedback*. However, in a room full of other mediocre students the score board race might be more stimulating than in a heterogeneous classroom with *W* who will always stay at the bottom and *S* who is unbeatable. In classrooms

with homogeneous feedback every action in the classroom might cause rank changes in an ever-so-close race. In addition to peer effects as a confound (Falk and Ichino, 2006), in this setting teachers could adapt their teaching style to the ability level of the classroom (Banerjee et al., 2016) and weaker classrooms might act as a negative signal (Steenbergen-Hu et al., 2016). However, with digitization in this study, we can pinpoint the pure effect of ability grouping feedback as illustrated above.

The contribution of this work, thus, is to apply insights from the ability grouping in classrooms and schools (Steenbergen-Hu et al., 2016; Betts, 2011; Kimbrough et al., 2020; Hanushek and Wößmann, 2006; Card and Giuliano, 2016; Dustmann et al., 2017) to feedback (recently reviewed in Villeval, 2020). The research on the total effect of ability grouping has policy relevance for actual classroom settings (Steenbergen-Hu et al., 2016; Betts, 2011). In our digital learning context, in contrast, where many anonymous learners meet on a learning platform, our setting gains policy relevance: group assignment for the specific reason of feedback provision is inevitable as there is no 'natural' grouping like with local schools (Thaler and Sunstein, 2008). The present work will show how *W*, *M*, and *S* learn when facing RPF with peers similar to themselves in contrast to peers drawn from the full population of learners.

We conducted a pre-registered randomized controlled trial (Klausmann, 2020)[1] with vocational learners in an anonymous digital learning platform to investigate effects of homogeneous feedback grouping on educational outcomes. Differing from ability grouping in the traditional classroom setting this helps to isolate the effects of feedback. In other words, we analyze the effect of feedback in a more homogeneous (ability grouped) group of learners compared to feedback in the heterogeneous group of randomly drawn learners. 2477 learners do not receive feedback (control group), whereas 2414 learners are sorted into heterogeneous feedback groups and 2461 learners are sorted into

[1] We received an ethics approval by the joint ethics commission of the departments of economics at Goethe University Frankfurt and University of Mainz (Gemeinsame Ethikkommission Wirtschaftswissenschaften der Goethe-Universität Frankfurt und der Johannes Gutenberg-Universität Mainz) on February 6th 2020.

homogeneous feedback groups. Both treatment groups receive instantaneous RPF in feedback groups of ten learners. The learners in the heterogeneous feedback treatment group are randomly assigned to feedback groups. Thus the group composition in expectation reflects the composition of the full sample of learners in our intervention. The learners in the homogeneous feedback treatment group are split into quintiles based on their past performance and assigned to a feedback group consisting of members of the same performance quintile.

The randomized controlled nature of this trial allows us to answer two questions: First, can homogeneous feedback grouping increase the positive average effect feedback has on learning? Second, does this effect differ with respect to performance, i.e., do weak learners profit more, similar, or less from feedback in homogeneous groups than strong learners? Additionally, the comparison between our treatment groups and the control group allows us to establish the average effect of anonymous feedback in a field setting.

In line with comparable field experiments in education (Azmat and Iriberri, 2010; Goulas and Megalokonomou, 2021; Jalava et al., 2015; Andrabi et al., 2017; Celik Katreniak, 2018; Fischer and Wagner, 2018) we find that providing RPF to learners significantly increases their learning effort. Effort, i.e., the number of tasks learners solve during the intervention, in the two treatment groups is on average 0.11 standard deviations higher than in the control group that receives no feedback. This is noteworthy for two reasons: First, participants in our trial are absolutely anonymous to one another. Thus, despite the field setting, direct social comparison concerns can be ruled out (Smith, 2000). Second, our effort measure reflects participants' choices compared to leisure as their alternative time use: our participants learn unsupervised and unmonitored in their free time and keep other forms of learning constant according to a survey measure.

We do not find mean differences between homogeneous and heterogeneous groups. Learners on average do not seem to profit from homogenizing groups in our setting.

When analyzing the differential effect by ability we find suggestive evidence for weaker learners increasing their effort because of homogenization and stronger learners reducing their effort. This hints at our intervention combining positive effects of feedback on average and decrease the inequality that arises when strong learners increase their effort due to feedback and weak learners do not (Haenni, 2019; Gneezy and Fershtman, 2011).

This paper contributes to several strands of literature. In the following we introduce related work on the total effect of relative performance feedback, other attempts to reduce inequality with feedback, rank response, ability grouping and peer effects.

Feedback is often found to have a positive effect on effort and performance overall (Villeval, 2020; Azmat and Iriberri, 2010; Kuhnen and Tymula, 2012; Goulas and Megalokonomou, 2021; Jalava et al., 2015; Andrabi et al., 2017; Celik Katreniak, 2018; Fischer and Wagner, 2018; Brade et al., 2020). At the lower end of the performance distribution, i.e., for weak learners like *W*, there is evidence for negative or null effects of feedback on effort. Gneezy and Fershtman (2011) find that weaker students in a sports competition quit when running side-by-side with stronger runners. Except for its physical nature, this setup comes close to the RPF provided in this study as we also provide feedback in real time, in contrast to many of the following studies that provide accumulated feedback after a period of data collection (e. g., grades after exams). Jalava et al. (2015) describe a more continuous relationship by interacting a feedback treatment with an ability measure: high achieving and mediocre 6th graders who receive feedback increase their effort, while low achievers do not react to the feedback. Haenni (2019) finds evidence of discouragement as losers in amateur tennis take longer breaks before returning to competitions. Brade et al. (2020) find similar effects in higher education. They find that the effect stems from the motivation of above-average feedback. Franco (2019) finds a positive average treatment effect of feedback on investment into academic inputs.

However, the point estimate for weak students is smaller than for strong students. This low impact of feedback on weak learners can even be anticipated, as learners react before the feedback is revealed. Ashraf et al. (2014) observe workers reduce their effort only with the knowledge of receiving feedback, therefore reducing the information in the signal and possible bad news.

An alternative to our approach to deal with the discouraging effects of feedback on weak learners is to use a different measure to score participants. Hermes et al. (2021) use performance improvements as the measure that students receive feedback on. The authors analyze low achievers, who are ranked better on performance improvements than they are used to in setups where they are ranked on absolute performance. These learners improve without reducing the positive effect of feedback for high achievers. Fischer and Wagner (2018) use a comparable measure which they call change feedback. They do not find average differences between performance and change feedback. However, they find that negative change feedback is especially motivating.

The literature on ordinal rank response is related to this work as it also uses variation in group composition. The variation in Elsner and Isphording (2017) and Elsner and Isphording (2018) stems from stronger and weaker cohorts in high school. They find students with higher rank but similar ability finish high school more often, are more likely to attend college, and engage less in risky behavior like smoking, drinking, unprotected sex, and fighting. Denning et al. (2020) use a similar variation and find positive long term effects of rank feedback on earnings. Murphy and Weinhardt (2020) find similar effects between primary and secondary school. Hett and Schmidt (2018) establish a rank response type in the lab and find that this type is consistent across tasks. Gill et al. (2019) estimate a rank response function without using variation in group assignment, but by resolving rank conflicts – two learners with the same performance – randomly.

In the context of the literature on ability grouping[2] our estimation can be understood

[2] Steenbergen-Hu et al. (2016) point out that the terms tracking and ability grouping are sometimes used

as the partial effect of between-class ability grouping in the context of anonymously provided feedback. Steenbergen-Hu et al. (2016) summarize meta-studies on ability grouping for within-class grouping – where teachers form smaller groups of learners based on their ability during class – and between-class grouping – where classrooms or whole schools are only open to students of a certain ability. The authors find no effect of between-class grouping, while other forms of ability grouping, like within-class grouping increase students performance by 0.19 - 0.3 standard deviations. Betts (2011) summarizes work, mostly from the field of economics, and concludes that ability grouping might aggravate inequality. A comprehensive work of this type is Hanushek and Wößmann (2006) who find increased inequality in test scores in a large international difference-in-difference study on ability grouping in schools. Betts (2011) also points out that – among several other components of ability grouping – well-designed within-school ability grouping might have dramatically positive effects. In a more recent example of this, Card and Giuliano (2016) find that stronger minority students profit from ability grouping reducing inequality along race lines. Kimbrough et al. (2020) take a comparable approach to ours by identifying a partial effect of ability grouping. They analyze how peer effects change between homogeneous and heterogeneous ability groups. The authors conclude that ability grouping offsets the benefits that weaker learners have from classroom peers.

Our intervention also relates to peer effect interventions: We can exclude effects driven by image concerns due to peer effects with the anonymous nature of our intervention. However, any effect of our intervention can be due to mutual monitoring and norm conformity as well as due to competitive preferences, the two other possible mechanisms Villeval (2020) discusses in the domain of peer effects. This literature usually finds that high ability participants have positive effects on their peers (Villeval, 2020; Epple and Romano, 2011). In the context of education, however, there are many

to differentiate the same policy between primary school (ability grouping) and high school (tracking) or flexible (ability grouping) and more long-term (tracking). We follow their approach also based on Loveless (2013) and use the term ability grouping for both phenomena.

exceptions: Austen-Smith and Fryer (2005) describe "acting white" as a phenomenon of peer effects where students are punished for abandoning a norm of under-average performance. Bursztyn and Jensen (2015) show that high-achieving students reduce their performance by 40% when public feedback is provided, while weaker students improve slightly. They conclude that all students adhere to a performance norm that is only slightly above the performance of the weakest. The anonymity in our setting might reduce the pressure of such norms.

Our intervention, thus, contributes to research on *feedback*, *ability grouping* and *peer effects*. Within the literature on feedback we contribute to work on *anonymous feedback*, the *differential effect along the lines of ability* and *rank response.* Like the laboratory studies on feedback (Kuhnen and Tymula, 2012; Gill et al., 2019) we can estimate the effect of *anonymous feedback*, however, in a field setting. This also contributes to the literature on *peer effects* as anonymity excludes one dimension of peer effects – image concerns – from feedback in a field setting. With respect to the *differential effect of feedback along the lines of ability,* assigning participants to feedback groups of equals is one possible approach for reducing discouragement. Applying our intervention, Gneezy and Fershtman (2011)'s runners might not quit if the distance to the better runner is in the milliseconds and Haenni (2019)'s tennis players might sign up for the next tournament if they loose in the tie breaker. Regarding *rank response* our work is a supplement to, e. g., Denning et al. (2020), as we analyze the effect of better peers despite identical education when we look at the strongest students. The rank response functions estimated by Elsner and Isphording (2017), Elsner and Isphording (2018), Gill et al. (2019), and Denning et al. (2020) display an ordinal performance measure – the participants' rank – on x-axis and outcomes on the y-axis. The heterogeneity analysis in Section 5.3 that disentangles the treatment effect along the lines of performance is similar, but displays lagged performance on the x-axis. Our effects are therefore orthogonal to rank

effects between participants. We enrich the literature on *ability grouping* as we look at the partial effect of ability grouped feedback on effort and performance.

This paper is structured as follows: Section 2 provides intuitions on how homogeneous groups might change the effect of feedback for learners on different ability levels. Based on the model of education by Cunha and Heckman (2007) it provides behavioral arguments from which the hypotheses for the later analysis are derived. Section 3 provides an in-depth look at the unique setting of our experiment and Section 4 describes the data. Section 5 tests the previously derived hypotheses. In Section 6 we summarize and provide an outlook that can be understood as a research agenda on group assignment in a digitized world.

2. Concepts of Homogeneous Group Feedback

In this section, we dive into possible theoretical mechanisms for the effect of homogeneous feedback groups on different types of learners. From this we can derive the hypotheses needed for the later analysis. Both the story of *W*, *M* and *S* as well as the literature above provide a motivation for homogenizing feedback groups. They also hint at the mechanisms of how past ability, performance, or skills could translate into future performance. Here, we describe possible relationships and then argue which preferences and biases could underpin these relationships.

As outcomes we analyze learning effort and performance and thus educational inputs and outputs. We derive predictions based on a simplistic version of the seminal theoretical work that models educational inputs and outputs: for the three treatment groups we will derive skill formation curves based on Cunha and Heckman (2007). For such an endeavor we need to relate the terms (1) *performance* and *effort* that we use in line with the feedback literature with (2) Cunha and Heckman (2007)'s *skill* and *investment*

and (3) *ability*[3] as described by the literature on ability grouping (Steenbergen-Hu et al., 2016; Betts, 2011). We use *performance* synonymous to Cunha and Heckman (2007)'s stock of *skills* θ and as our measure for *ability* (Steenbergen-Hu et al., 2016; Betts, 2011). We use effort as the empirical equivalent to Cunha and Heckman (2007)'s investment I_t. While *effort* might only be one element of investment into learning it is the choice variable we are interested in. A learner thus has a performance θ_t at the beginning of a period of learning and a new performance θ_{t+1} afterwards. Following Cunha and Heckman (2007), the new performance is determined by parental characteristics which we abstract from, the previous performance level θ_t, and the investment into learning I_t. For us the relationship therefore simplifies to $\theta_{t+1} = f_t(\theta_t, I_t)$. The shape of the skill formation curve that we will later derive is based on Cunha and Heckman (2007)'s concepts of self-productivity ($\partial f_t(\theta_t, I_t)/\partial\theta_t > 0$) and dynamic complementarity ($\partial^2 f_t(\theta_t, I_t)/\partial\theta_t\partial I_t > 0$) – the increasing and growing relationship between θ_t and θ_{t+1}.

Providing feedback increases the salience of performance θ_t. Additionally, Gneezy and Fershtman (2011), Haenni (2019), Jalava et al. (2015), Brade et al. (2020), and Franco (2019) describe the differential effect of heterogeneous feedback (as the authors do not implement more homogeneous settings) along the lines of past performance θ_t on current performance θ_{t+1}. Feedback in these works has a positive average effect. The effect is insignificant, negative but small, or positive but small for the weakest learners or sportsmen depending on the setup. Therefore, in our model we assume that heterogeneous feedback has no effect on the weakest learner – the intercept of the skill formation curve is similar without feedback and with heterogeneous feedback – while the effect increases with past performance – the slope of the skill formation curve is steeper with heterogeneous feedback than without feedback (compare all panels of Figure 2.1).

In contrast to the differential effect of past performance and heterogeneous feedback

[3] In the economics of education ability often refers to unobserved ability, a term that can hardly be related to performance. Ability grouping in contrast is often based on grades and test performance (between-class) or comprehension questions (within-class).

we are not aware of empirical work on the differential effect of homogenization and past performance. Thus, we argue below how homogeneous feedback groups compared to heterogeneous groups evoke different behavioral reactions. Both, performance θ and effort I at different points in time could be elements of learners' preferences. Thus, in the latter analysis, we will analyze both outcomes effort and performance, despite the fact that the only instantaneous choice our participants have is their effort I_t.

Panel A of Figure 2.1 shows how participants in homogeneous feedback groups react differently from participants in heterogeneous feedback groups assuming that *individuals optimize* θ_{t+1}. This could stem from both, preparing for any test that approximates θ, or optimizing own future ability for other reasons (e.g., as a signal for future wage negotiations). To derive this we assume a fixed time budget in which learners can choose between investing into learning I_t and leisure and look at the effect of assigning a learner to a homogeneous feedback group instead of a heterogeneous feedback group. In a heterogeneous feedback group dynamic complementarity leads to every unit of effort I_t being amplified by the higher current θ_t – high current performance leads to even higher future performance. The signal that stems from receiving feedback makes θ more salient, but does not change rational beliefs. When such individuals receive feedback in homogeneous groups – especially blind to treatment – learners receive a different signal. A weaker learner – in a weak group of homogeneous learners – on average receives a better signal than in heterogeneous groups. Weak learners, therefore, experience that their (relative) performance is higher, anticipate higher dynamic complementarity, and increase their learning at the cost of leisure. Stronger learners, in turn, experience that their (relative) performance is lower, anticipate lower dynamic complementarity, and reduce their learning while increasing leisure. The signal to mediocre learners will remain to be the mean signal and thus we expect no effect (compare Panel A of Figure 2.1). The average treatment effect of homogenizing feedback groups will be zero, as the positive

effect on weak learners and the negative effect on strong learners cancel out.

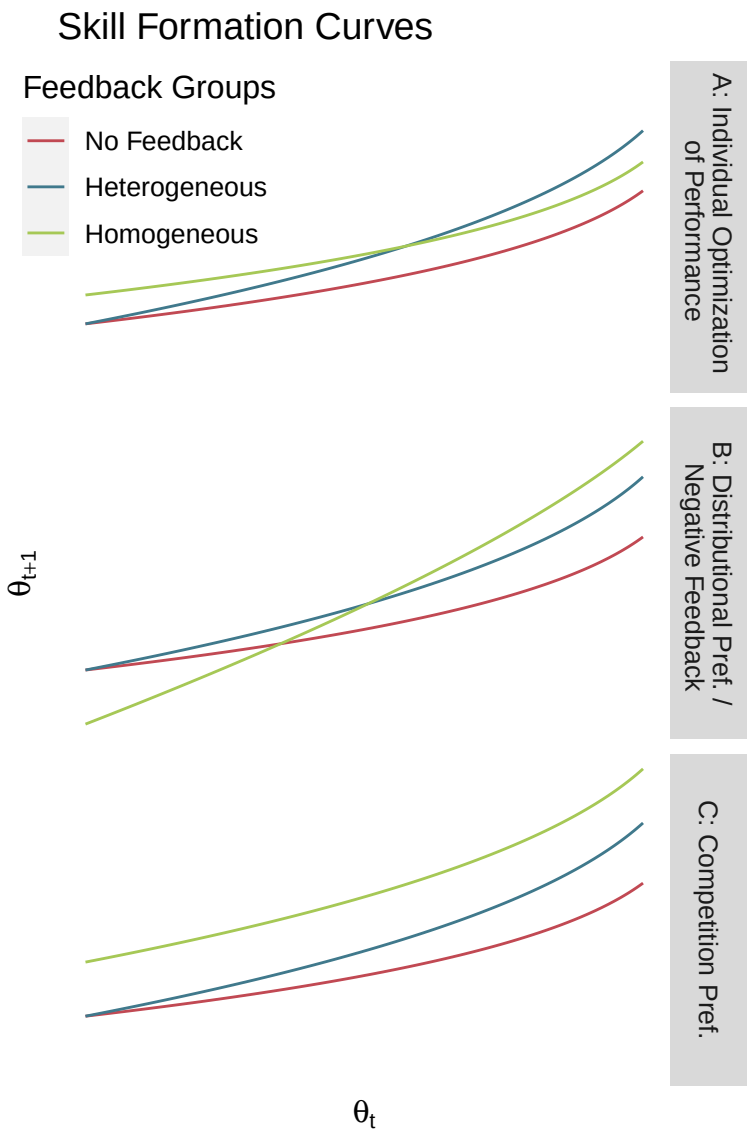

Figure 2.1: The three panels show possible skill formation curves of the three treatments based on Cunha and Heckman (2007). The x-axis shows past performance θ_t and the y-axis future performance θ_{t+1}. The shape of the skill formation curves in the control group and with heterogeneous feedback is similar in all panels, while the shape with homogeneous feedback changes between mechanisms.

Panel B of Figure 2.1 shows a skill formation curve based on behavioral economics and empirical evidence. The skill formation curve could stem from learners displaying *distributional preferences* (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). With grading on a curve (Czibor et al., 2020; Kulick and Wright, 2008) learners might view their ranking as information on their final exam grade. In our intervention learners do not know how groups are composed (homogeneously or heterogeneously). Taking their rank as a signal for their standing in the population might lead to effects that depend on learners' ability: In our homogeneous feedback groups compared to heterogeneous groups weaker learners reach a better rank with the same effort. With the goal of reaching a fixed grade and therefore fixed spot in the distribution they could

take their higher rank as a signal of their expected grade and reduce their effort I_t. Equally, stronger learners are on lower ranks on the RPF score board than in the population and might therefore increase their effort to achieve their expected grade. Mediocre learners observe the mean signal independent of treatment and thus should not change their behavior. Also based on distributional preferences we expect to observe no average treatment effect of homogenization.

Another explanation for the skill formation curve in Panel B of Figure 2.1 could stem from empirical evidence in the feedback literature. Comparing similar individuals in homogeneous and in heterogeneous feedback groups means that these similar individuals face different *feedback valence*: A good learner in a heterogeneous group always faces positive feedback; the same learner might face negative feedback in a homogeneous group. Azmat and Iriberri (2016) and Burgers et al. (2015) find no effect of valence, which implies that this difference might not matter for the effect of feedback. Fischer and Wagner (2018) find that negative RPF induces higher performance than positive RPF. This effect would produce a similar skill formation curve as the mechanism based on distributional preferences above since weaker learners receive more positive RPF and stronger learners more negative in homogeneous feedback groups compared to heterogeneous feedback groups.

Last, Panel C of Figure 2.1is based on both the literature on ability grouping (e.g., Duflo et al., 2011) as well as the literature on feedback (e.g., Haenni, 2019) discuss the effects of *preferences for competition* that could produce the skill formation curves. In homogeneous groups compared to heterogeneous groups learners on a RPF score board are more similar with respect to their absolute performance, the in-feedback-group variance of performance is smaller. A fixed amount of effort input by a given learner is, therefore, more likely to change the learners' rank on the RPF score board. A learner with strong preferences for competition can – in this case – experience more competition and thus

derive more utility from homogeneous feedback groups than from heterogeneous[4]. Panel C in Figure 2.1, therefore, shows an upward shift of the homogeneous feedback skill formation curve as learners across all skill levels experience increased competition. Strong preferences for competition would lead to more learning by all learners. While Haenni (2019) does not experimentally identify or discuss this effect, their evidence suggest that tennis players competing with players that are ranked more similar to themselves engage in the next competition sooner and thus more often in total than if they compete with either much stronger or much weaker players. If this correlation stems from a causal relationship it supports that all learners in homogeneous groups increase their effort due to tighter competition.

As using homogeneous feedback groups centrally changes the relationship between ability and what feedback participants receive, the discussion above provides a baseline for the analysis in Section 5. With respect to average treatment effects we expect:

Hypothesis 1: Learners in the feedback treatments invest more effort and consequentially perform better than those in the control group without feedback.

Depending on the model from above the average difference between the homogeneous feedback arm and the heterogeneous feedback arm is predicted to be different. Following *Individual Optimization* (Panel A) and *Distributional Preferences* (Panel B) we do not expect an average difference. Following *Competition Preferences* (Panel C) we expect effort and performance to increase on average:

Hypothesis 2: Learners in the homogeneous feedback treatment invest the same / more effort and consequentially perform similar / better than in the heterogeneous feedback treatment.

Also with respect to the intercept and the slope of the skill formation curve we expect different reactions. The *Distributional Preferences* (Panel B) model suggests that weak

[4]Here, we ignore extreme cases at the ends of the distribution. E.g., a learner with much higher performance than all other learners in the group does not necessarily experience this effect.

learners in the homogeneous feedback arm perform worse. Therefore, we expect a lower intercept than in the heterogeneous treatment arm. The the two other approaches based on *Individual Optimization* (Panel A) and *Competition Preferences* (Panel C) suggest that weak learners profit from homogeneous groups and we would thus expect a higher intercept.

Hypothesis 3: Learners in the homogeneous feedback treatment have a lower / higher intercept of the skill formation curve than in the heterogeneous feedback treatment.

The slope in homogeneous groups compared to heterogeneous groups should be flatter based on *Individual Optimization* (Panel A) and steeper based on *Distributional Preferences* (Panel B). Based on *Competition Preferences* (Panel C) the slopes should not differ between homogeneous and heterogeneous treatment groups.

Hypothesis 4: Learners in the homogeneous feedback treatment have a flatter / equal / steeper slope of the skill formation curve than in the heterogeneous feedback treatment.

3. Intervention Design

Our randomized controlled intervention ran on a learning platform for college-level students that train to become bank clerks from March to November 2020.[5] The majority of them are around 20 years old. Usage of the platform is voluntary and cannot be monitored by schools or employers. Usually, these students have access to the platform over their full 2-3 year training, which consists of classroom and practice spells. The learning platform enables students to learn and prepare for exams also in times when teachers are not available. All participants in our sample had their account pre-paid by their employer, i.e., usage is costless for them. We have access to pseudonymized data on all interactions between students and the learning platform; thus, we can monitor individual learning behavior of each participant in real time.

[5]The platform can be reached at https://www.pruefungs.tv.

Primarily, the learning platform featured a large number of learning videos. Additionally, we introduced a training tool to the platform that involves quizzes. Each quiz includes 10 items – henceforth called 'tasks'. Upon answering, students receive direct information on correctness as well as arguments as to why their answers were correct or wrong. This tool can be used for acquiring new knowledge as well as for repeating already known learning content. We relied on material previously produced by experts for each subject. Both on the landing page of the platform as well as on the page initiating each quiz we introduced an additional visual element to provide RPF and the number of correctly solved tasks as a performance metric (similar to the P-index in Dobrescu et al., 2021) in the two treatment groups (compare Figure 3.1).

We designed our randomized experiment such that we can compare learners receiving feedback and learners that do not receive feedback ceteris paribus. More specifically, we compare a control group that receives no feedback to two treatment groups that receive feedback. Participants in one treatment group (henceforth denoted by 'heterogeneous feedback') receive feedback with ten learners that are drawn from the full sample of learners in our intervention. Participants in the second treatment group (denoted by 'homogeneous feedback') receive feedback with ten learners who performed in the same quintile before. Homogeneous feedback groups are filled by a share of randomly assigned participants that cannot be assigned to a quintile as we do not have data on previous performance. Every participant stays in the same treatment arm, but not in the same feedback group for the complete intervention. In our data we might have many more influences on effort and performance than our treatments – a learner might increase or decrease effort and performance due to shocks and unobservables. To allow that homogeneous groups stay homogeneous throughout the time of our intervention, it is important to restart them based on a current measure of performance regularly. To keep this constant between treatments feedback groups are reset, newly randomized,

and restarted on the first day of every month. Importantly, participants are blind to the mechanisms of assignment into feedback groups.

Participants in the homogeneous treatment arm are sorted by their performance in the previous month, i.e., how many tasks they solved correctly. This sorted list is divided into five equi-sized groups or quintiles: last month's top 20% performers, the 20% just below them, the middle 20%, and two more such groups for the following and the last 20% of learners. In the homogeneous feedback arm, feedback groups will consist primarily out of participants from the same quintile as discussed below.

A central challenge of providing feedback in an online platform with voluntary usage is the irregular and hard to predict participation: In the summer break month of July only 1020 of the 7352 participants solved any tasks at all (for all months compare Figure A.2 in the appendix). Therefore, including all participants in feedback groups of ten would have led to an average of 8.6 participants with a zero score at the end of the month and even higher shares at the beginning. Such groups do not seem informative to us. Therefore we chose a group assignment mechanism that only adds participants to a group after they completed their first task in a month.

In the homogeneous treatment arm data on the current performance of participants

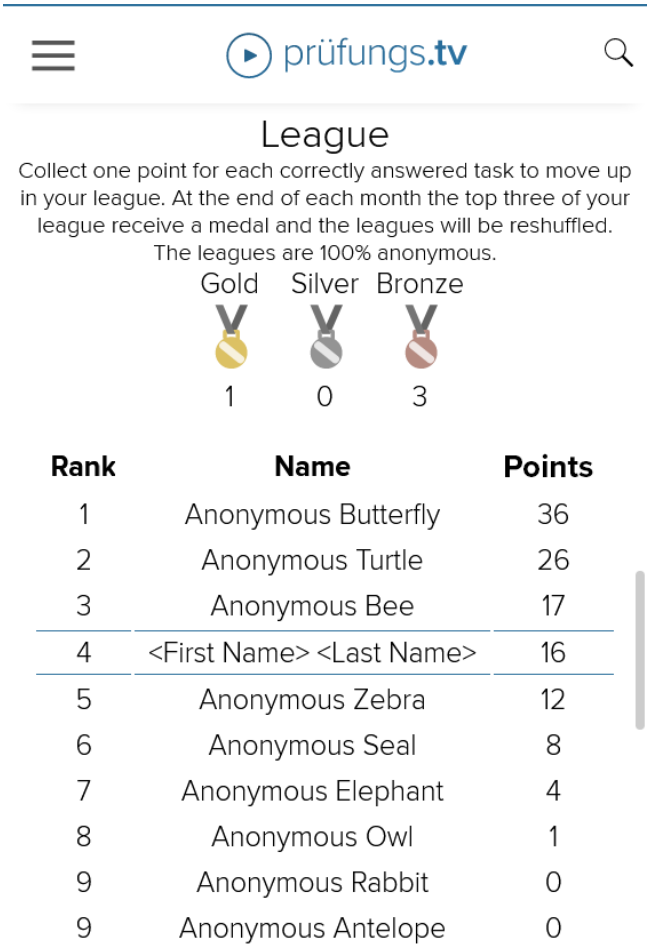

Figure 3.1: Smartphone screen shot of the real time feedback translated to English. Find the original German version in appendix Figure A.1.

is needed for monthly reassignment as described above. However, as usage of the platform is voluntary not all participants used the platform in the preceding month. Therefore, participants are assigned to homogeneous feedback groups in two ways depending on their engagement in the previous month. With the first assignment mechanism, all participants who were active in the last month, i.e., completed at least one task, are assigned to feedback groups. The first participant from a given quintile generates a new feedback group. Every following participant – as long as there are open slots in a feedback group of their quintile – is assigned to this group. If there is no open slot a new group is initiated. With the second assignment mechanism we add all participants that are in the homogeneous treatment arm, but did not complete a task in the previous month, randomly to one of the five currently open quintile groups.

In the heterogeneous treatment arm participants are assigned similarly. We also open five different groups in parallel – corresponding to the five quintile groups in the homogeneous treatment arm – to keep the above-mentioned time structure of assignment fixed. However, in this heterogeneous treatment arm all participants are assigned randomly to one of the open groups.

Naturally, this assignment mechanisms leads to groups that are not yet filled with ten participants when they are initiated. We choose to always display nine anonymous other participants in addition to a participant's own name on the RPF score board.[6] The slots that no participant is assigned to are displayed with zero solved tasks. This anticipates that participants will fill these blank slots on the RPF score board as soon as they complete their first task (compare the last two rows of Figure 3.1).

[6] In the case that a group is not filled until the end of a moth this displays more learners than are present in the group. We chose this strategy instead of a flexible list length to keep the visual components of the intervention constant over all participants.

4. Collected Data

Every interaction with the learning platform is recorded and stored. We obtained information on every attempted task (correctness, timing) as well as the initiation of each quiz. Furthermore, as a proxy for whether participants substitute quizzes presented on the platform with other possible learning, we observe each video consumed by the participants in the learning platform.

After the intervention, we collected data from an incentivized voluntary questionnaire. All participants that answered the questionnaire received 5€ for answering 13 items. In this questionnaire we recorded test performance in the standardized high-stake final test with which participants complete their degree. This test determines the full final grade of the degree and it is taken by all participants at the same time. Additionally, we collected gender, academic track, and grades of the participants' high school education. Furthermore, we collected beliefs on the participants' ranking after the intervention, the number of hours that students learned offline per week, and GRIT (Growth Resilience Integrity And Tenacity), which is based on four questions from Schmidt et al. (2019)'s German validation of Lee Duckworth et al. (2009). Descriptive summary statistics can be found in Table A.1.

In total the questionnaire was completed by 396 out of 7352 participants (i.e., 5.4%, 133 from the control group, 148 from the heterogeneous feedback treatment, and 115 from the homogeneous feedback treatment). We pre-registered the performance in the final exam at the end of the intervention period as secondary outcome (Klausmann, 2020). We based this pre-registration on a previous study on the same learning platform, but with different learners (Klausmann and Schunk, 2021). The response rate to the questionnaire was 14%. The reasons for this difference between Klausmann and Schunk (2021) and this study are unclear to us. Due to the much lower than expected response rate we do not discuss these variables with respect to causal effects in the following

analysis to circumvent the threats of under-powered analysis and potential self-selection. The result is attached in the appendix (compare Figure A.5).

5. Results

7352 participants were randomly assigned to one of three treatment arms (compare Table 5.1). Participants in the control group did not receive any feedback, but faced the identical platform otherwise. Participants were assigned into monthly feedback groups in the two feedback treatment arms. In the heterogeneous feedback group, participants received feedback in feedback groups with learners from the full distribution of same age learners. In the homogeneous feedback arm, participants received feedback with learners form the same performance quintile as they are in.

		Treatment arm	Participants
Feedback	*No*	Control	2477
	Yes	Heterogeneous Feedback	2414
		Homogeneous Feedback	2461

Table 5.1: Our treatments can be understood as splitting the sample, first, into receiving feedback and no feedback and second, spliting those that recieve feedback into a homogeneous feedback treatment arm and a heterogeneous feedback treatment arm. The control group recieves no feedback. Participants in the homogenous treatment arm face RPF with participants from the same quintile and the participants in the heterogeneous treatment arm face RPF with randomly drawn participants.

The subsequent analysis is structured as follows: first, in Section 5.1, we establish that anonymous feedback in this specific online learning platform affects learning effort and performance in the platform (Hypothesis 1). Learning effort in the platform is measured as the average number of completed tasks. The number of correctly answered tasks is the measure for the absolute performance corresponding to θ_{t+1} from Section 2. Second, in Section 5.2, we discuss the average treatment effect (ATE) of homogeneous feedback compared to heterogeneous feedback on the same outcome variables within

the learning platform (Hypothesis 2). A positive average treatment effect of homogenization would entail that, compared to the heterogeneous feedback treatment arm, participants completed more tasks and thus invested more effort into learning and performed better. However, we cannot establish such a difference. Third, in Section 5.3, we estimate the skill formation curve for the three treatment arms to answer if learners react to homogeneous feedback differently depending on their initial ability (Hypothesis 3 and 4).

It is important to point out that we have assigned participants randomly to the control or either of the two treatment groups in the moment of their first interaction with the quiz tool. Table A.2 in the appendix shows that randomization with respect to the obtained control variables worked well. Not only are the variables that were measured in the platform before participants experienced our treatment balanced, but also variables that we collected in the ex-post questionnaire (gender, High School GPA and GRIT) are balanced between treatments.

5.1. Feedback Increases Learning Effort

Result 1: Anonymous instantaneous relative performance feedback increases learning effort and performance.

The left panels of Figure 5.1 summarize how participants in all three treatments use the platform on average and aggregated over the whole intervention time span. Focusing on the top left panel, participants in the heterogeneous feedback arm complete 0.12 standard deviations more tasks compared to the control group on average (47.97 additional tasks or 125.59% of the control group). This difference is highly significant (Wilcoxon rank sum test $p < 0.001$). Participants in the homogeneous feedback arm complete 0.11 standard deviations more tasks compared to the control group (43.84 additional tasks or 123.38%, Wilcoxon rank sum test $p < 0.001$). The same holds true

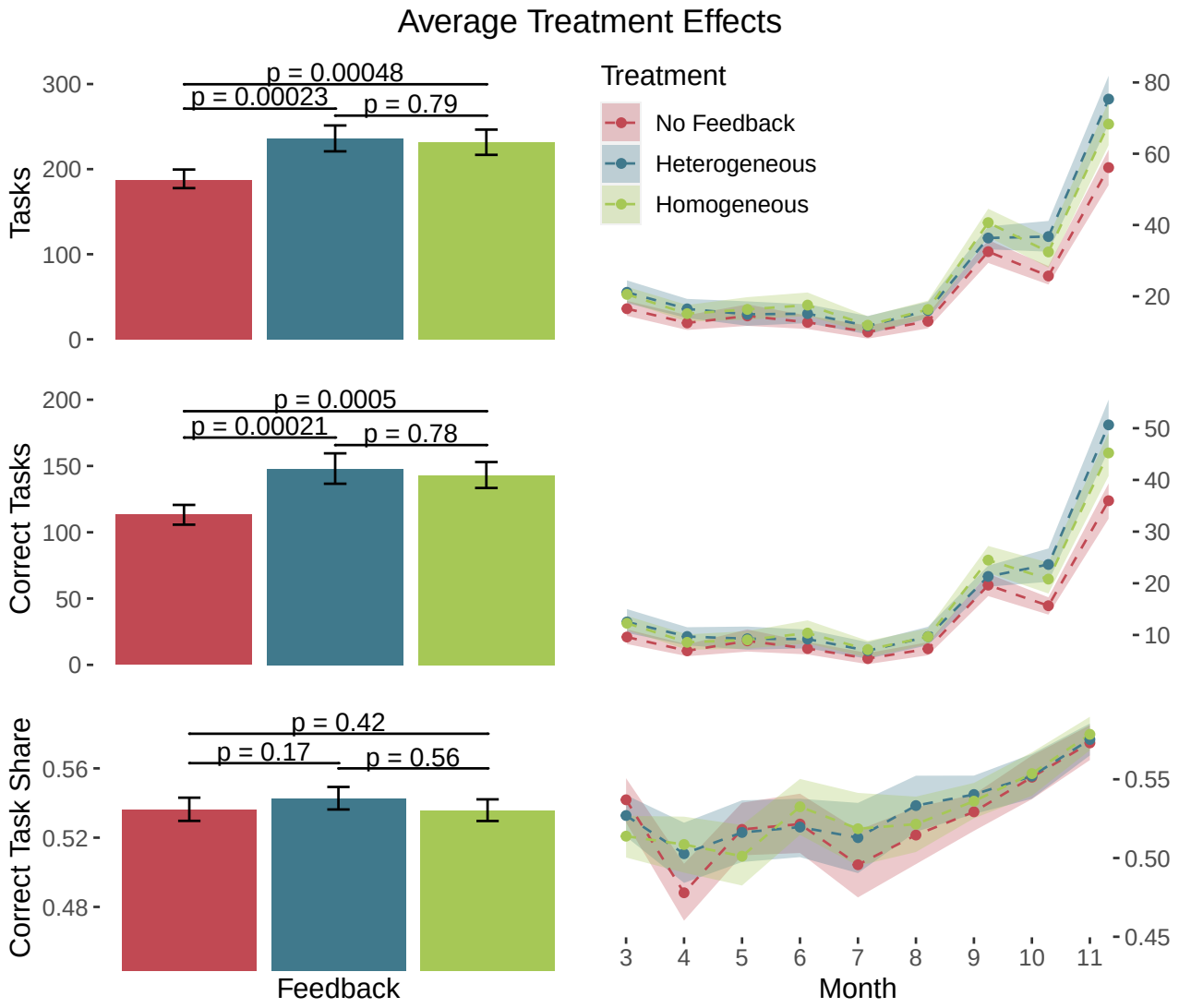

Figure 5.1: We observe a positive treatment effect for both feedback treatments on effort, absolute performance, and constant relative performance between all treatments. In the first row of panels the outcome variable (y-axis) that we interpret as effort is the average number of completed tasks (effort). In the second row of panels the outcome variable (y-axis) that we interpret as absolute performance is the average number of correctly solved tasks. In the third row of panels the outcome variable (y-axis) that we interpret as relative performance is the share of correctly solved tasks out of the completed tasks. All panels on the left aggregate over the whole intervention in each of the three treatment groups. All panels on the right aggregate monthly for each of the three treatment groups. The error bars show nonparametric bootstrapped 95% confidence intervals. The brackets comparing treatment groups are non-parametric Wilcoxson rank sum tests.

for our absolute performance measure, the number of correctly solved tasks in the center left panel. Participants in the heterogeneous feedback arm complete 0.12 standard deviations more tasks correctly and participants in the homogeneous feedback arm complete 0.12 standard deviations more tasks correctly. Both these differences are highly significant (Wilcoxon rank sum tests $p < 0.001$). This additional absolute performance corresponds to the additional effort: Participants in the treatment groups do not only attempt to solve more tasks, but maintain their share of correctly solved tasks. The bottom left panel of Figure 5.1 shows that all treatment groups solve a similar share of tasks correctly (0.54% of all attempted tasks). 42% (60%) of the additional tasks that participants in the homogeneous (heterogeneous) treatment arm did more than participants in the control group where completed in the final two months of the intervention. This additional effort can only carry it's fruit in a following period possibly after the intervention that we cannot measure. In the right panels of Figure 5.1 we can observe how in later months, closer to the final high-stake test, not only the effort increased, but also the performance in absolute and relative terms.[7]

All in all this is evidence that the feedback we use in this intervention influences learning behavior. Participants engage and spend more time which leads to higher effort provision and consequentially higher absolute performance.

5.2. No ATE of Homogenizing Feedback Groups

Result 2: Assigning learners to homogeneous feedback groups compared to heterogeneous feedback groups does not increase learning effort or performance on average.

Comparing the heterogeneous and the homogeneous treatment arm in all panels of Figure 5.1 suggests that learners in these two groups exert the same effort and show the same performance on average. Learners in the homogeneous treatment arm complete

[7] In Figure A.5 in the appendix we show the insignificant treatment effect on exam outcomes in a regression context based on the failed and therefore under-powered final questionnaire.

-0.01 standard deviations less tasks than learners in the heterogeneous treatment arm. This difference is insignificant (Wilcoxon rank sum test $p = 0.8$) and far below the minimal detectable effect size (MDES) of 0.06 standard deviations. The same holds for the number of correctly solved tasks: The difference between the feedback treatment arms is -0.02 standard deviations and insignificant. Further, the share of correctly solved tasks does not differ between the heterogeneous and the homogeneous treatment group with an insignificant difference of -0.01%. Also, over time we find no difference between the homogeneous and the heterogeneous treatment groups. During the first 6 months with generally low activity the difference between both treatments is low and insignificant. In the final three months the differences are larger (September: 0.04 sd, October: -0.04 sd, and November: -0.07 sd), but still insignificant and below the MDES.

One reason for this null finding could be the dilution of homogeneous feedback groups. As described in Section 3 we assign participants that were not active in the previous month randomly to quintile groups. On average that meant randomly assigning 43% of participants to the homogeneous groups (compare Figure A.3 for the monthly share). It is hard to evaluate whether this is a high or low share compared to full random assignment in the heterogeneous treatment arm. The intervention, however, is based on the number of correctly solved tasks which we can observe. Figure 3.1 shows that participants can observe the number of correctly solved tasks of every other participant in their feedback group as points. The ranking is based on the same number. Empirically we can, therefore, evaluate if participants that were assigned to higher quintiles on average saw more tasks solved correctly in their feedback group than participants in lower quintile groups. That would mean that participants in a higher quintile faced others that were better than if they would have been assigned to a feedback group in a lower quintile on average. Figure 5.2 suggests that this is the case. The top quintile completes on average 134 tasks while the lowest quintile completes 78 tasks. Except for the two bottom

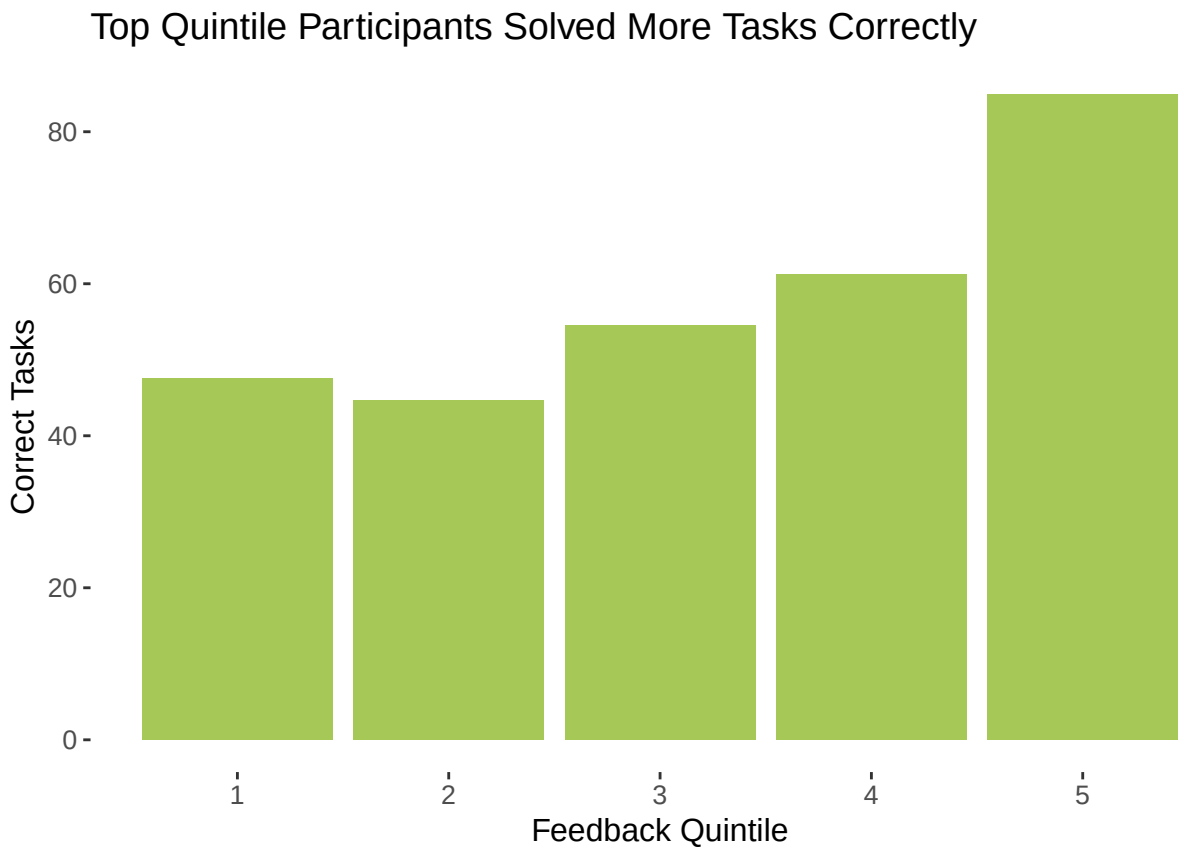

Figure 5.2: In the homogeneous treatment arm, participants in the higher quintile groups solve more tasks correctly than in the lower quintile groups. In direct comparison only participants in the bottom quintile solved slightly more tasks correctly than in the second lowest group. The y-axis shows the number of completed tasks. Each bar shows the mean solved tasks in one quintile groups over the whole intervention period. The number of correctly solved tasks determines the feedback ranking and is salient to all participants in the treatment groups.

quintiles participants in each quintile faced others that were better than if they would have been assigned to a feedback group in a lower quintile on average. This shows that the homogeneous feedback groups were different from each other. The null findings between the the heterogeneous and the homogeneous treatment arms can therefore not be explained entirely by the dilution of feedback groups.

In total, we find no average difference between the heterogeneous and the homogeneous treatment arms. In groups that learn with learners similar to themselves and in groups with all kinds of learners the average effort and performance in the platform is similar. In the following we will dive into these effects and split learners along the lines of their own performance to find if these null effects on average are also present for weaker and stronger learners.

5.3. Skill Formation in Homogeneous Feedback Groups

The effect of homogenizing groups may vary along performance. In Section 2 we discuss how learners could react to receiving feedback in groups with other, similar learners. From this we derive theoretical skill formation curves for learners in the three treatment arms. Here we will estimate the empirical equivalents of skill formation curve for all treatment arms including the control group to answer if – in our our empirical setup – learners react to homogeneous feedback differently depending on their initial ability.

To do so, we need to impose additional structure on the problem: we need to specify (1) a period length to split the data along the time dimension into periods t and $t+1$, (2) an empirical measure for performance in line with θ in Section 2, and (3) a functional form that is in line with Cunha and Heckman (2007)'s concepts of self-productivity and dynamic complementarity. Addressing (1), our setup entails feedback resets and reassignment to new feedback groups every month. Therefore, we choose monthly intervals for t. Regarding the empirical performance measure (2) we follow the intervention design

as closely as possible. Participants in the homogeneous feedback groups were assigned to quintiles based on the participants' last month's rank with regard to number of correctly solved tasks. Consequentially, we choose this rank as measure for the current level of performance θ_t. We scale this rank from 0 to 5 to represent the intervals in which participants in the homogeneous feedback treatment arm were assigned to the respective feedback groups. A participant with $\theta_t \in (0, 1)$ was assigned to the lowest feedback quintile group in the homogeneous treatment arm, participant with $\theta_t \in (1, 2)$ was assigned to the second to lowest feedback quintile group, and so on. θ_{t+1} as described in Chapter 2 is the performance in the period after θ_t realized. As θ_{t+1} is an outcome variable in the following the scaling described above for θ_t is not necessary and we can use the number of correctly solved tasks parallel to the analysis above. With respect to (3) we are looking for a function that can produce positive second order derivatives and at the same time adheres to Ockham's razor with as little parameters to estimate as possible. Furthermore, in Section 2 we predicted different levels (y-axis cutoffs) and slopes of the skill formation function. Given these requirements an exponential function seems to be a reasonable choice with flexible second order derivatives and only two parameters to estimate that correspond to the predictions from above:

$$\theta_{t+1} = \alpha + \beta * e^{\theta_t}$$

In the following we present results from a pooled OLS framework, collecting data from all months, and an OLS framework with data form the final month of November only (compare Table 5.2). The econometric challenge in this context is the serial dependence of θ_t and θ_{t+1} that makes both the pooled OLS setup (Columns (1) and (2) in Table 5.2) and any classical panel model (within, random effects, and first differences estimators) inconsistent (Cameron and Trivedi, 2005). A possible approach to estimate a consistent panel model would be the Arellano-Bond estimator (Arellano and Bond, 1991). However,

Table 5.2: Skill Formation Curve Estimates

	Dependent variable:			
	Correct tasks			
	Pooled	Pooled (Clu.)	Nov.	Nov. (Clu.)
	(1)	(2)	(3)	(4)
Constant	10.943***	10.943***	37.296***	37.296***
	p = 0.000001	p = 0.000	p = 0.000001	p = 0.000
Heterogeneous Feedback	−0.497	−0.497	−19.088*	−19.088**
	p = 0.875	p = 0.844	p = 0.064	p = 0.024
Homogeneous Feedback	3.303	3.303	−1.940	−1.940
	p = 0.293	p = 0.234	p = 0.850	p = 0.837
e^{θ_t}	0.776***	0.776***	1.193***	1.193***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Hetero. Feedback * e^{θ_t}	0.333***	0.333**	1.336***	1.336***
	p = 0.000001	p = 0.011	p = 0.000	p = 0.001
Homo Feedback * e^{θ_t}	0.209***	0.209	0.717***	0.717*
	p = 0.002	p = 0.139	p = 0.002	p = 0.072
p-value Equal Intercept	0.187	0.651	0.058	0.526
p-value Equal Slope	0.353	0.876	0.128	0.808
Observations	13,080	13,080	2,212	2,212

Note: We find that the intercept skill formation curve for our heterogeneous feedback treatment group is – in some models marginally significantly – lower than the intercept for the homogeneous feedback treatment group. At the same time the slope of the homogeneous feedback treatment group is insignificantly flatter. The slope of both feedback treatment groups is steeper than the slope of the skill formation curve in the control group. The first two columns pool data from all months of the intervention, while the latter two use only data from the final moth of the intervention. In columns (2) and (4) we cluster the standard errors on the level of randomization. The first three variables show the intercept of the skill formation curve with the x-axis for the control group, the homogeneous feedback treatment group, and the heterogeneous feedback treatment group. The latter three show estimates for the slope of the skill formation curve for the control group, the homogeneous feedback treatment group, and the heterogeneous feedback treatment group. The p-values at the bottom compare the estimates for intercept and slope of the heterogeneous and homogeneous feedback treatment groups using a pairwise Wald test. Legend: * p<0.1; *** p<0.05; *** p<0.01

this would require three consecutive months of activity by any participant that we could include in the regression. In our unbalanced panel with voluntary and, therefore, endogenous engagement in each month this might lead to a sample selection bias that might be especially hurtful in our setup where weaker participants might choose to engage less often. The approach we prefer is to look at one cross-section only to circumvent these estimation challenges. The final month of the intervention (November) with the highest engagement is the trivial choice for this cross-section (Columns (3) and (4) in Table 5.2). One downside of this approach is that in contrast to a within model, we cannot control for individual characteristics. Thus the estimate for the intercept and the slope of the skill formation curve is not purely determined by the θ_t, but also by all other individual characteristics. As we are only interested in comparing treatment effects on this intercept and slope, we can abstract from the specific intercept and slope and focus on the treatment driven differences. To correct our outcomes for joint variation that stems from the ten participants that are in the same feedback groups, in Column (2) and (4) in Table 5.2 we cluster standard errors at the level of randomization. The level of randomization is the super-group of learners that arrive at the same time and are either randomized (heterogeneous treatment arm) or assigned into feedback groups. This correction, however, might overestimate the effect of clustering as users that arrive at a similar time in a month might share other characteristics than our assignment only, e.g. a similar learning routine.

Result 3: Homogenizing feedback groups mitigates the negative effect of feedback on weak learners.

Figure 5.3 visualizes the estimates discussed above. The intercept represents the performance θ_{t+1} – number of correctly solved tasks – of the weakest learners. Comparing these intercepts we find that learners in the heterogeneous treatment arm perform worse than learners in the control group (19.09 correct tasks less for the weakest learn-

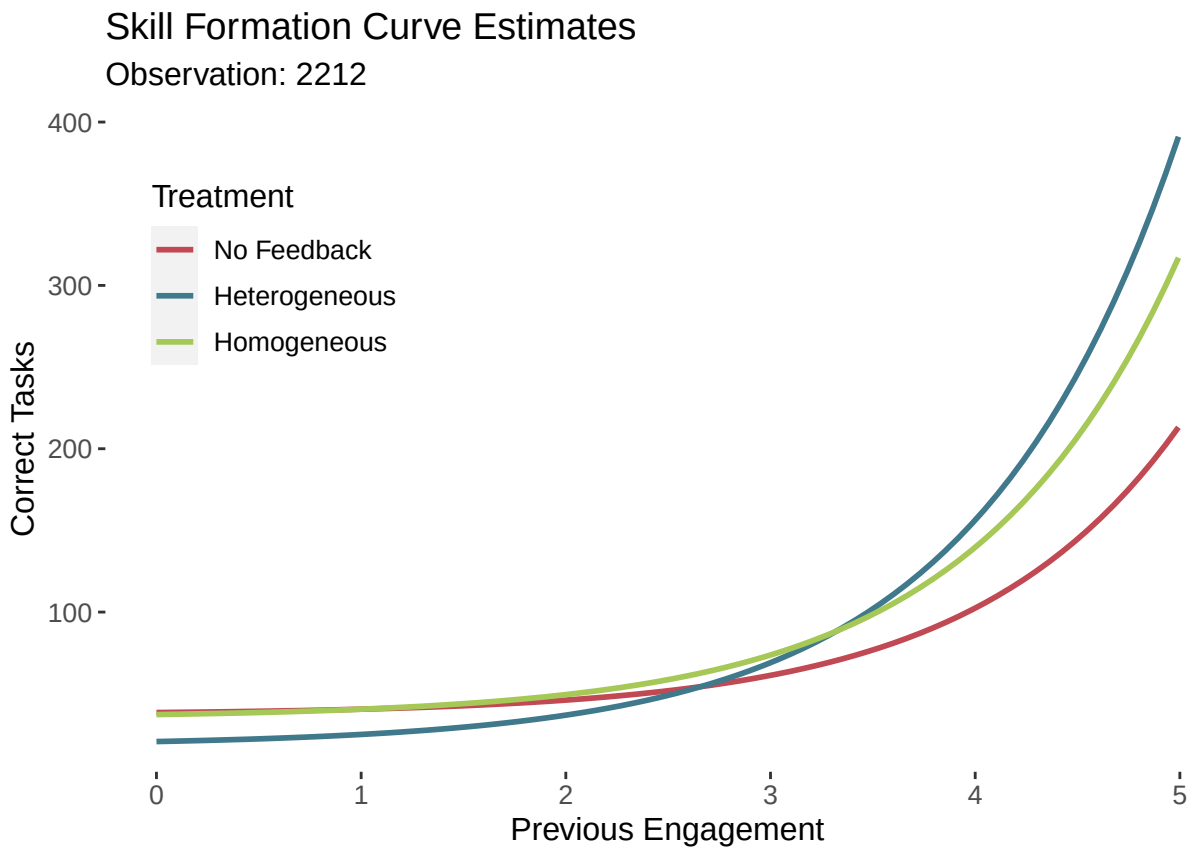

Figure 5.3: We find that mainly high ability learners (quintiles 4 and 5) in our two treatment groups perform better than in the control group. The slope of the skill formation curve is slightly flatter in the homogeneous feedback group than in the heterogenous feedback group. The data in this figure is based on data from the final month (November) parallel to column (4) of Table 5.2. The x-axis plots the position of participants, based on correctly answered tasks that a learner obtained in the previous month and is scaled such that an increase by one corresponds to moving up one feedback quintile group in the homogeneous treatment. The y-axis plots performance measured by the correctly solved tasks in the current month. The three curves are estimates for the skill formation curves for each treatment group. Figure A.6 in the appendix also contains confidence intervals.

ers; compare column (4) of Table 5.2). In the model with clustered standard errors this difference is significant (p-value = 0.024). Graphically, a similar effect holds in the weakest two quintiles. The weakest learners in the homogeneous treatment group solved an insignificant -1.94 tasks less than the control group. Summarizing, this indicates, though insignificantly, that feedback in heterogeneous groups has a negative impact on weak learners in line with Gneezy and Fershtman (2011) and Haenni (2019) and furthermore that homogenizing feedback groups can mitigate this effect.

Learners in the top two quintiles of the treatment groups outperform the control group. Figure A.4 shows that this difference is significant in the strongest feedback quintile in the subset of data from the final month of the intervention.

Result 4: The skill formation curve in homogeneous feedback groups is insignificantly flatter than in heterogeneous feedback groups.

In Figure 5.3 we observe that the point estimate for the slope of the skill formation curve in the homogeneous feedback treatment arm is smaller than in the heterogeneous treatment arm. However, with and without clustered standard errors this difference is insignificant. The positive point estimate supports the hypothesis that under individual optimization and dynamic complementarities homogeneous feedback groups lead to a flatter skill formation curve. Also improbable with the significantly lower performance of the participants in the top quintile we cannot exclude the model based on competition preferences or possibly no effect.

6. Conclusion

Grewenig et al. (2020) found that learners reduced their learning time during the recent Corona virus pandemic with home schooling by about half. One reason for this reduction could be the reduced feedback as learners cannot compare their achievements in class and in the school yard. Both, the direct effect of more digital education and the

additional need for feedback will increase the presence of feedback in digital learning tools. A feature of digital learning platforms is that learners are often not grouped like in a local school: thousands of learners might face the same platform at the same time. This large amount of learners might not be the optimal group size (Abuseileek, 2012, find that five is the optimal group size in their digital learning setup). Assigning learners to smaller groups opens the opportunity and necessity to design the group composition. One such strategy could be to apply the long-standing literature on ability grouping (Steenbergen-Hu et al., 2016; Betts, 2011; Kimbrough et al., 2020; Hanushek and Wößmann, 2006; Card and Giuliano, 2016; Dustmann et al., 2017) and form homogeneous ability groups. In this work we randomly assigned learners in an online learning platform to a control group that receives no feedback and two treatment groups: one in which learners received RPF in heterogeneous groups with other randomly drawn learners and one with learners with similar performance. Learners in the feedback treatment groups invested more effort and performed better than learners that did not receive feedback. We find that feedback in such homogeneous ability grouping does not increase the average learning effort, but does effect learning depending on ability: weaker learners learn marginally more in homogeneous ability groups, while stronger learners reduce their effort.

With these results we contribute to the feedback and the ability grouping literature. We show that feedback increases learning effort in anonymous groups in a field setting. By identifying the effect of ability grouping feedback only we identify one channel of ability grouping.

There are two main challenges in our study that might reduce possibly larger effects and could be tackled in future research. First, the voluntary nature of the platform we utilize leads to lower engagement than could be expected in a schooling environment. This has two effects: participants in one feedback group seldom meet and learn at the

same time. Thus, oftentimes the only score board changes participants observe while being logged on is their own rank (moving up with completed tasks), while the changes in other participants scores happen while the participant is offline. Also the difference in ability that we observe between the middle and lower groups is small in our setup. Ability grouping makes only a small difference between these groups. Second, past month performance is an imperfect measure for ability and dilutes possibly larger effects of our intervention. A closer monitoring of the learning behavior of participants might make it possible to assign groups more precisely. Alternatively, regular assessments might make such classification more precise, but at the same time raise costs. Lastly, machine learning techniques could help with forming group and use past data to classify also learners without recently recorded data into homogeneous groups. This way, feedback could – in the future – have positive impacts on learners of all abilities.

A. Appendix

A.1. Tables

Table A.1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Number of Tasks	7,352	217.919	409.210	1	7,008
Correct Tasks	7,352	134.539	283.857	0	5,928
Correct Tasks Share	7,352	0.538	0.200	0.000	1.000
Number of videos	7,352	35.930	95.483	0	1,187
Beliefs before interv. (in %)	1,018	47.281	18.812	0.000	100.000
Beliefs after interv. (in %)	396	66.129	19.215	0.000	100.000
Questionnaire participation	7,352	0.054	0.226	0	1
Score in final test (in %)	373	77.150	9.911	25.000	99.500
Score in final test (written only, in %)	396	74.744	11.398	22.333	100.000
Score in final test (oral only, in %)	373	83.649	10.523	33.000	100.000
GRIT	396	15.025	2.293	8.000	20.000
Female	396	0.604	0.490	0.000	1.000
Academic Track	396	0.790	0.408	0.000	1.000
HS GPA	374	−2.182	0.638	−4.000	−1.000
Offline learning (Hours/Week)	396	3.725	1.525	1.000	6.000

We observe individuals in continuous time. This table sheds light on the individual dimension. 7352 learners participated in our main intervention completing at least one task. Of these 296 participated in the final questionnaire that includes the grade in the final standardized offline tests.

Table A.2: Randomization Check

	No Feedback	Heterogeneous	Homogeneous	Hetero. vs Control	Homo. vs Control	Homo. vs Hetero.
First quiz mean	2020-07-10	2020-07-09	2020-07-09	0.680	0.659	0.979
Subscription end	2021-06-17	2021-11-22	2021-06-14	0.305	0.723	0.298
Female	0.6240602	0.5608108	0.6347826	0.283	0.862	0.226
Academic track	0.8045113	0.7837838	0.7826087	0.669	0.673	0.982
HS GPA	-2.165840	-2.196549	-2.182617	0.684	0.847	0.870
GRIT	15.23308	15.02027	14.79130	0.433	0.126	0.432

This table summarizes exogeneous variables to check if our data is balanced between treatment arms. The first three columns show the means of each variable in a treatment group. The last three columns show t-tests comparing the groups. We do not find significant differences between the treatment groups in any of these variables.

A.2. Figures

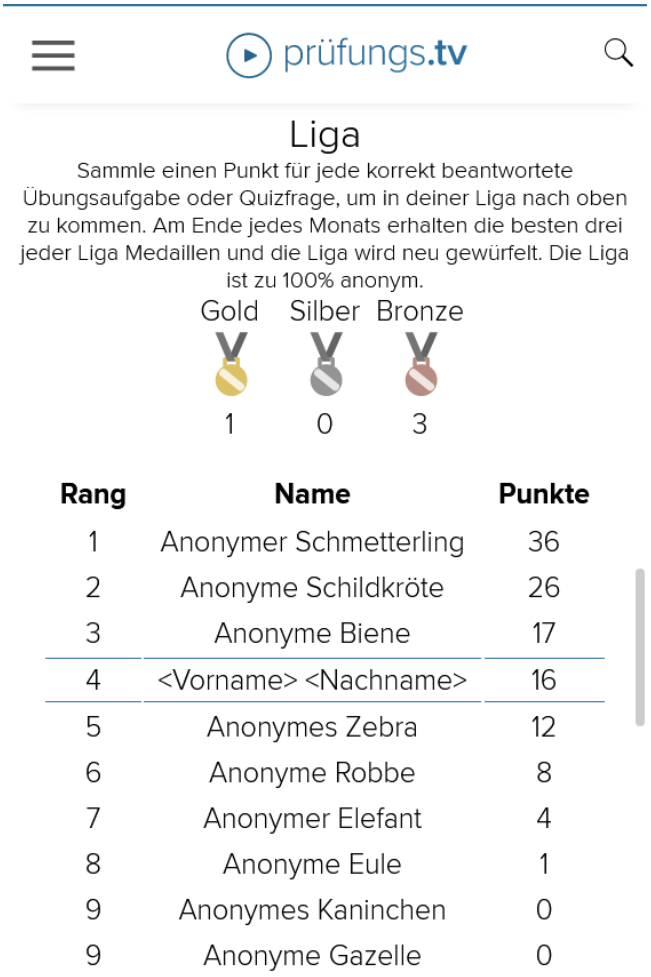

Figure A.1: Real time feedback in the learning platform.

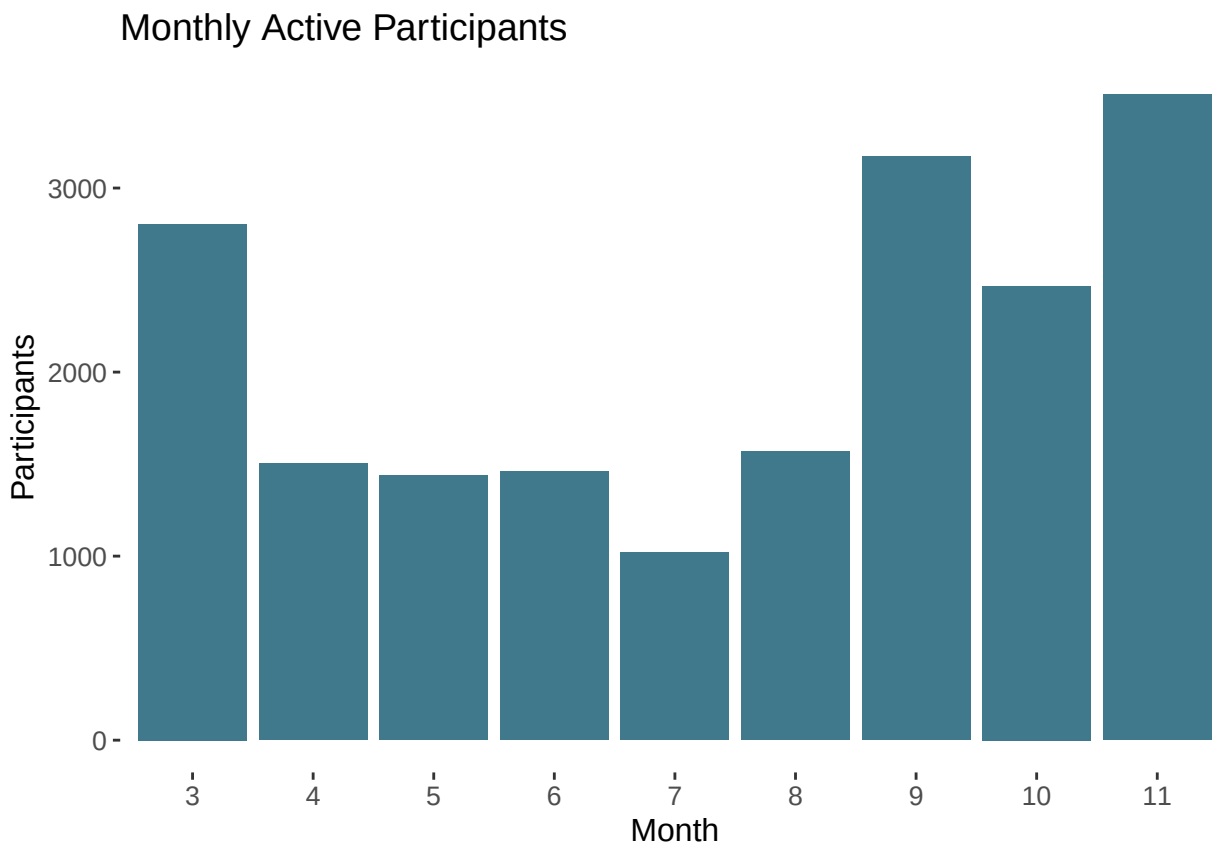

Figure A.2: Participants were most active in the first months we analyzed and in the final month before the exam. The months of the year 2020 are shown on the x-axis. The number of participants active are shown on the y-axis.

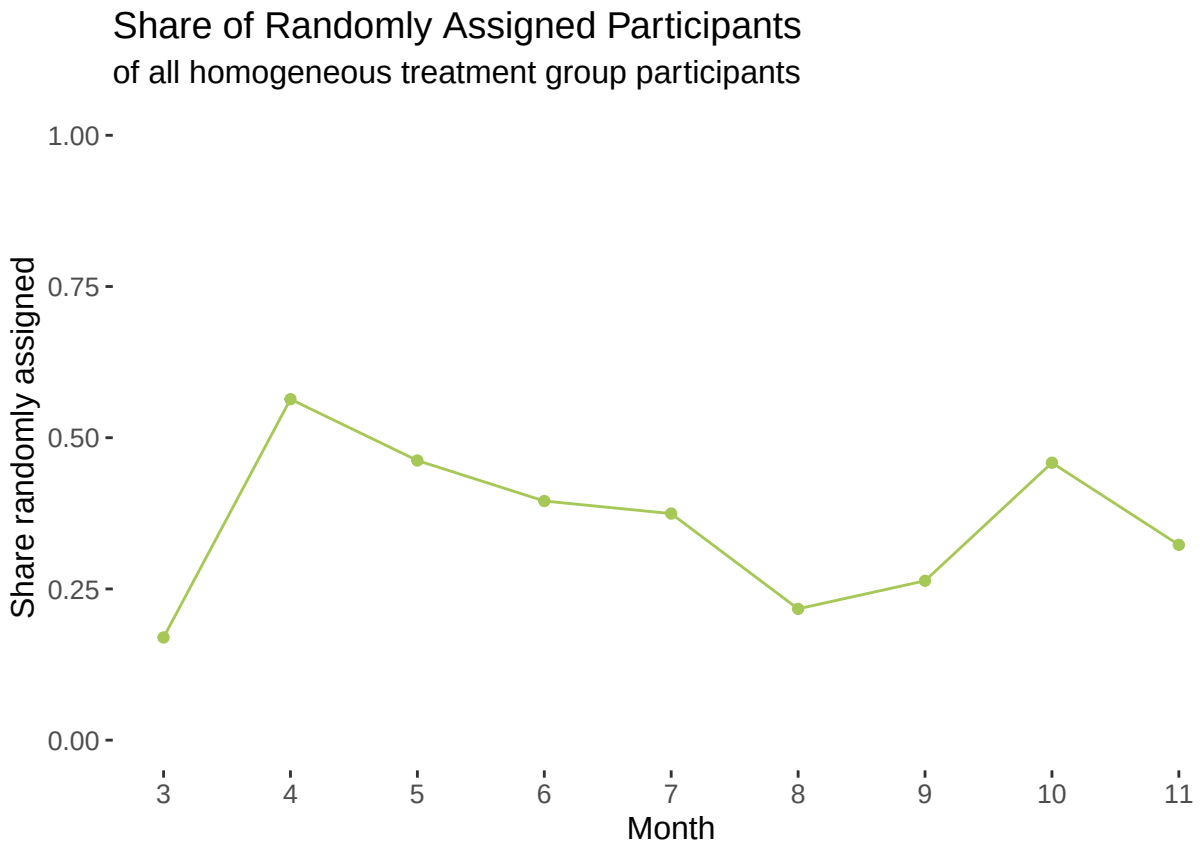

Figure A.3: In the homogeneous treatment groups not all individuals were assigned according to their ability quintile. All participants that were inactive in the previous month were assigned randomly to quintile groups. This figure shows the share of such randomly assigned participants by month.

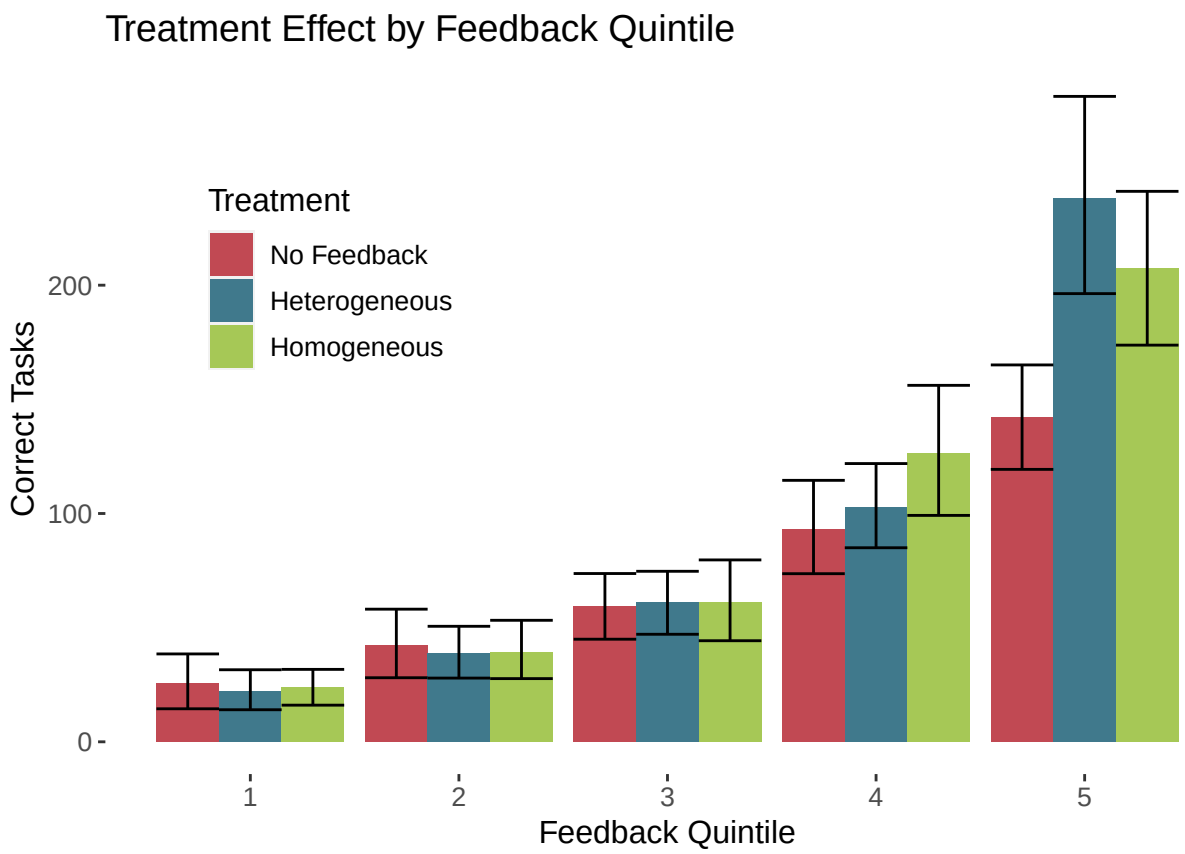

Figure A.4: While only participants in the homogeneous treatment arm receive feedback in homgeneous feedback groups, we can separate the treatment effect along the lines of correct tasks solved in the last month into the feedback quintiles participants would be assigned to. This figure shows data for the last month of the intervention. The x-axis shows these feedback quintiles from the weakest (1) to the strongest (5). On the y-axis the performance in the current month – the number of correct tasks – is displayed. The bars represent group means. The error bars show non-parametric bootstrapped 95% confidence intervals.

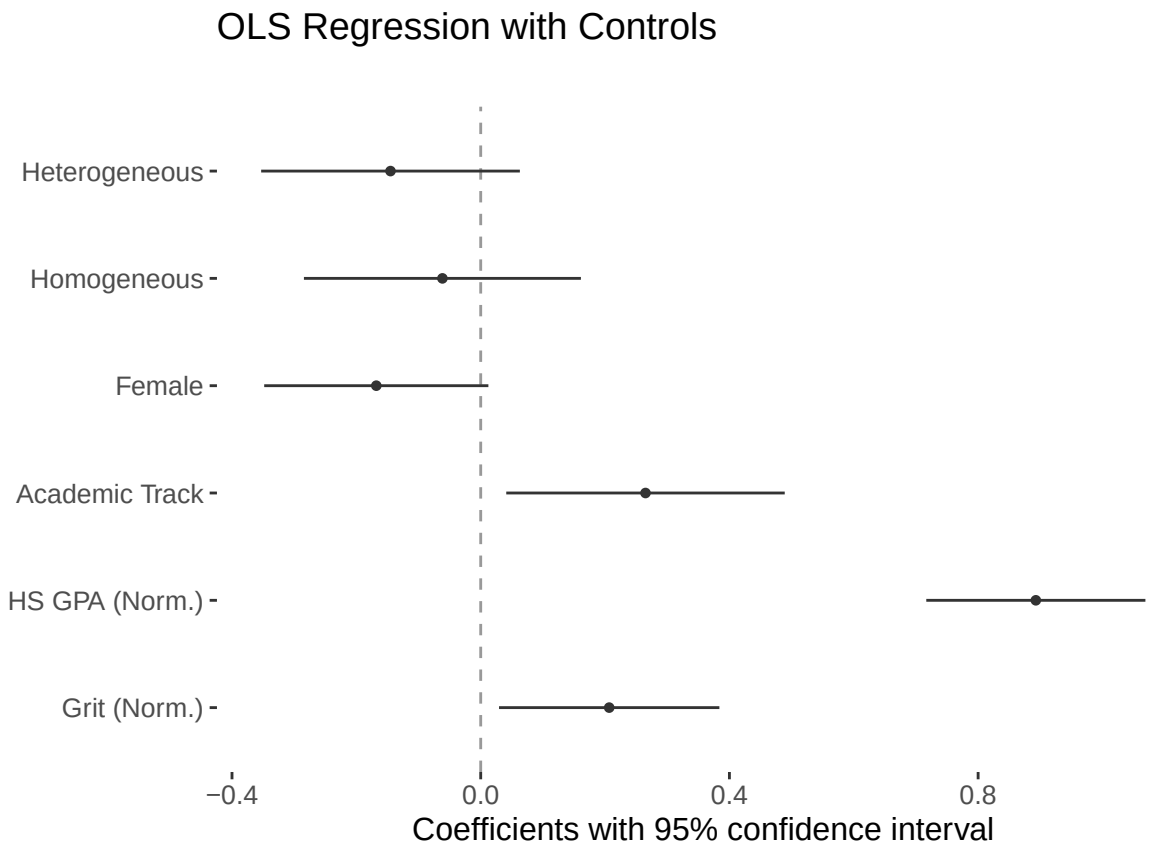

Figure A.5: We find no link between our treatments and grades in the final standardized high-stakes exam. This figure shows the estimates in standard deviations from the mean grade. Heterogeneous and Homogeneous are indicator variables that are one if a participant is in the respective treatment arm. To reduce standard errors we include further control variables. Female is an indicator for gender. Academic track is an indicator that is one if a participant attended the academic track of high school and zero otherwise. GPA is the normalized final grade in high school. Grit is a questionaire measure based on Schmidt et al. (2019)'s German validation of Lee Duckworth et al. (2009).

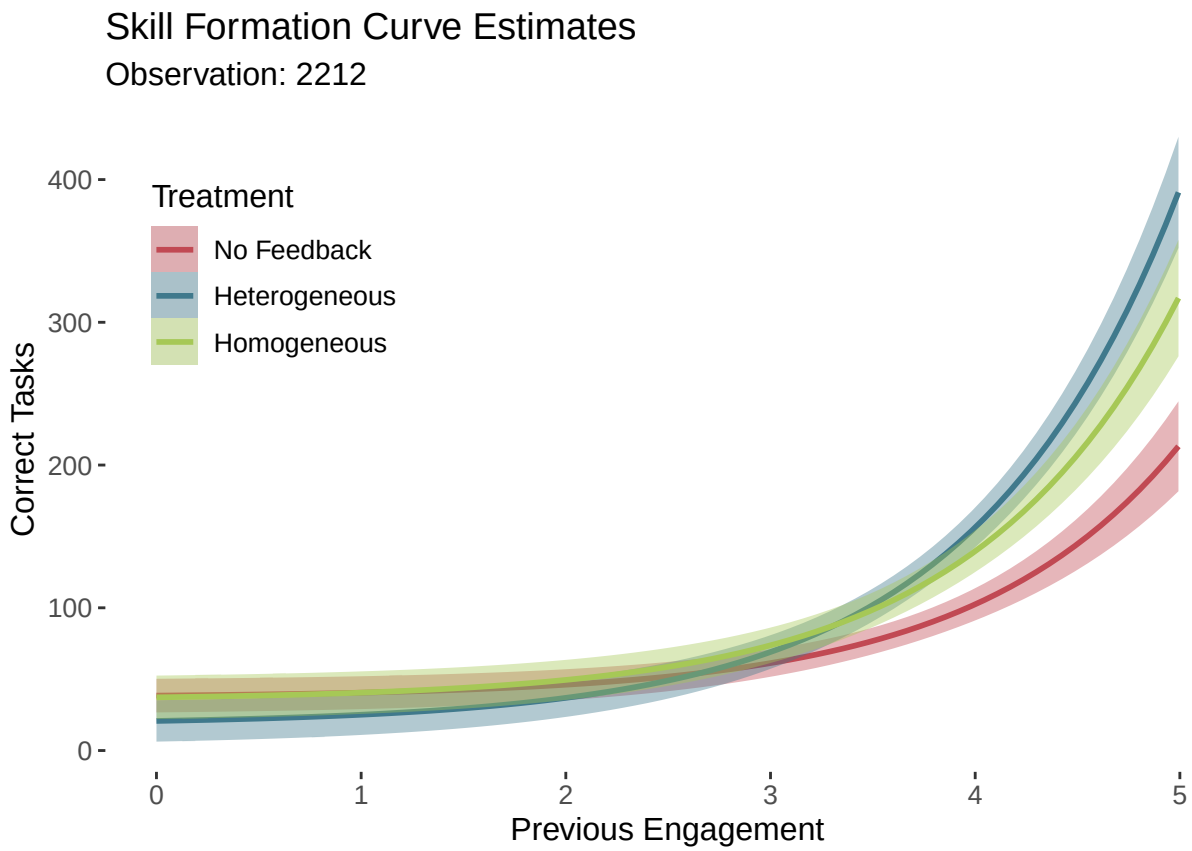

Figure A.6: We find that mainly high ability learners (quintiles 4 and 5) in our two treatment groups perform better than in the control group. The slope of the skill formation curve is slightly flatter in the homogeneous feedback group than in the heterogenous feedback group. The data in this figure is based on data from the final month (November) parallel to column (4) of Table 5.2. The x-axis plots the position of participants, based on correctly answered tasks that a learner obtained in the previous month and is scaled such that an increase by one corresponds to moving up one feedback quintile group in the homogeneous treatment. The y-axis plots performance measured by the correctly solved tasks in the current month. The three curves are estimates for the skill formation curves for each treatment group.

References

Abuseileek, A. F. (2012). The effect of computer-assisted cooperative learning methods and group size on the EFL learners' achievement in communication skills. *Computers and Education*, 58(1):231–239.

Andrabi, T., Das, J., and Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review*, 107(6):1535–1563.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data:monte carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2):277–297.

Ashraf, N., Bandiera, O., and Lee, S. S. (2014). Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior and Organization*, 100:44–63.

Austen-Smith, D. and Fryer, R. G. (2005). An Economic Analysis of "Acting White". *The Quarterly Journal of Economics*, 120(2):551–583.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452.

Azmat, G. and Iriberri, N. (2016). The Provision of Relative Performance Feedback Information: An Experimental Analysis of Performance and Happiness. *Journal of Economics & Management Strategy*, 25(1):77–110.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., Shotland, M., and Walton, M. (2016). Mainstreaming An Effective Intervention: Evidence From Randomized Evaluations Of "Teaching At The Right Level" In India.

Betts, J. R. (2011). *The Economics of Tracking in Education*, volume 3. North Holland.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Brade, R., Himmler, O., and Jäckle, R. (2020). Relative Performance Feedback and the Effects of Being Above Average – Field Experiment and Replication.

Burgers, C., Eden, A., van Engelenburg, M. D., and Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, 48:94–103.

Bursztyn, L. and Jensen, R. (2015). How Does Peer Pressure Affect Educational Investments? *The Quarterly Journal of Economics*, 130(3):1329–1367.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics Methods and Applications*. Cambridge University Press, Cambridge, 1 edition.

Card, D. and Giuliano, L. (2016). Can tracking raise the test scores of high-ability minority students? *American Economic Review*, 106(10):2783–2816.

Celik Katreniak, D. (2018). Persistent Overconfidence: Evidence from a Randomized Control Trial in Uganda.

Cunha, F. and Heckman, J. (2007). The Technology of Skill Formation. *AEA Papers and Proceedings*, 97(2):31–47.

Czibor, E., Onderstal, S., Sloof, R., and van Praag, C. M. (2020). Does relative grading help male students? Evidence from a field experiment in the classroom. *Economics of Education Review*, 75:101953.

Denning, J., Murphy, R., and Weinhardt, F. (2020). Class Rank and Long-Run Outcomes.

Dobrescu, L. I., Faravelli, M., Megalokonomou, R., and Motta, A. (2021). Relative Performance Feedback in Education: Evidence from a Randomised Controlled Trial*. *The Economic Journal (forthcoming)*.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Dustmann, C., Puhani, P. A., and Schönberg, U. (2017). The Long-term Effects of Early Track Choice. *Economic Journal*, 127(603):1348–1380.

Elsner, B. and Isphording, I. (2017). A Big Fish in a Small Pond: Ability Rank and Human Capital Investment. *Journal of Labor Economics*, 35(3):787–828.

Elsner, B. and Isphording, I. E. (2018). Rank, Sex, Drugs, and Crime. *Journal of Human Resources*, 53(2):356–381.

Epple, D. and Romano, R. E. (2011). Peer Effects in Education. In *Handbook of Social Economics*, volume 1, pages 1053–1163. Elsevier B.V.

Falk, A. and Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics*, 24(1):39–57.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Fischer, M. and Wagner, V. (2018). Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment.

Franco, C. (2019). How does relative performance feedback affect beliefs and academic decisions? Evidence from a field experiment.

Gill, D., Kissová, Z., Lee, J., and Prowse, V. (2019). First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision. *Management Science*, 65(2):494–507.

Gneezy, U. and Fershtman, C. (2011). The tradeoff between performance and quitting in high power tournaments. *Journal of the European Economic Association*, 9(2):318–336.

Goulas, S. and Megalokonomou, R. (2021). Knowing who you are: The Effect of Feedback Information on Short and Long Term Outcomes. *Journal of Behavior & Organization*, 183:589–615.

Grewenig, E., Schmidt, K. M., Lergetporer, P., Werner, K., Woessmann, L., and Zierow, L. (2020). COVID-19 and Educational Inequality: How School Closures Affect Low-and High-Achieving Students COVID-19 and Educational Inequality: How School Closures Affect Low-and High-Achieving Students.

Haenni, S. (2019). Ever tried. Ever failed. No matter? On the demotivational effect of losing in repeated competitions. *Games and Economic Behavior*, 115:346–362.

Hanushek, E. A. and Wößmann, L. (2006). Does educational tracking affect performance and inequality? differences-in-differences evidence across countries. In *Economic Journal*, volume 116, pages C63–C76. Oxford Academic.

Hattie, J. and Clarke, S. (2019). *Visible Learning: Feedback*. Routledge, London; New York, 1 edition.

Hermes, H., Huschens, M., Rothlauf, F., and Schunk, D. (2021). Motivating low-achievers—Relative performance feedback in primary schools. *Journal of Economic Behavior & Organization*, 187:45–59.

Hett, F. and Schmidt, F. (2018). Pushing Through or Slacking Off? Heterogeneity in the Reaction to Rank Feedback.

Jalava, N., Joensen, J. S., and Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196.

Kimbrough, E. O., McGee, A. D., and Shigeoka, H. (2020). How Do Peers Impact Learning? An Experimental Investigation of Peer-to-Peer Teaching and Ability Tracking. *Journal of Human Resources*, pages 0918–9770R2.

Klausmann, T. and Schunk, D. (2021). Understanding Adaptive Learning with a Field Experiment.

Klausmann, T. A. (2020). Feedback in Homogenous Groups Pre-Registration.

Kuhnen, C. M. and Tymula, A. (2012). Feedback, Self-Esteem, and Performance in Organizations. *Management Science*, 58(1):94–113.

Kulick, G. and Wright, R. (2008). The Impact of Grading on the Curve: A Simulation Analysis. *International Journal for the Scholarship of Teaching and Learning*, 2(2):1–17.

Lee Duckworth, A., Quinn, P. D., Duckworth, A. L., and Quinn, P. D. (2009). Development and validation of the short Grit Scale (Grit-S). *Journal of Personality Assessment*, 91(2):166–174.

Loveless, T. (2013). The 2013 Brown Center Report on American Education: How well are Students American Learning? Technical report, The Bookings Institution, Washington, D.C.

Murphy, R. and Weinhardt, F. (2020). Top of the Class: The Importance of Ordinal Rank. *The Review of Economic Studies*, (4815):1–50.

Schmidt, F. T., Fleckenstein, J., Retelsdorf, J., Eskreis-Winkler, L., and Möller, J. (2019). Measuring grit: A German validation and a domain-specific approach to grit. *European Journal of Psychological Assessment*, 35(3):436–447.

Smith, R. H. (2000). Assimilative and Contrastive Emotional Reactions to Upward and Downward Social Comparisons. In Suls, J. and Wheeler, L., editors, *Handbook of Social Comparison*, chapter 10, pages 173–200. Springer, Boston, 1 edition.

Steenbergen-Hu, S., Makel, M. C., and Olszewski-Kubilius, P. (2016). What One Hundred Years of Research Says About the Effects of Ability Grouping and Acceleration on K–12 Students' Academic Achievement. *Review of Educational Research*, 86(4):849–899.

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge*. Yale University Press, New Haven, 1 edition.

Villeval, M. C. (2020). Performance Feedback and Peer Effects. In Zimmermann, K. F., editor, *Handbook of Labor, Human Resources and Population Economics*, pages 1–38. Springer, Cham.